A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Michele Loi
University of Zürich
michele.loi@uzh.ch

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Andreas Krause
ETH Zürich
krausea@ethz.ch

Abstract

Equality of opportunity (EOP) is an extensively studied conception of fairness in political philosophy. In this work, we map recently proposed notions of algorithmic fairness to economic models of EOP. We formally show that through our proposed mapping, many existing definition of algorithmic fairness, such as predictive value parity and equality of odds, can be interpreted as special cases of EOP. In this respect, our work serves as a unifying moral framework for understanding existing notions of algorithmic fairness. Most importantly, this framework allows us to explicitly spell out the moral assumptions underlying each notion of fairness, and also interpret recent fairness impossibility results in a new light. Last but not least and inspired by luck egalitarian models of EOP, we propose a new, more general family of measures for algorithmic fairness. We empirically show that employing a measure of algorithmic (un)fairness when its underlying moral assumptions are not satisfied, can have devastating consequences on the subjects' welfare.

1 Introduction

Equality of opportunity (EOP) is a widely supported ideal of fairness, and it has been extensively studied in political philosophy over the past 50 years [Rawls, 2009; Sen, 1979; Dworkin, 1981a,b; Arneson, 1989; Cohen, 1989]. The concept assumes the existence of a broad range of *positions*, some of which are more desirable than others. In contrast to *equality of outcomes* (or positions), an equal opportunity policy seeks to create a *level playing field* among individuals, after which they are free to compete for different positions. The positions that individuals earn under the condition of equality of opportunity reflect their *merit* or *deservingness*, and for that reason, inequality in outcomes is considered ethically acceptable [Roemer, 2002].

Equality of opportunity emphasizes the importance of personal (or native) qualifications, and seeks to minimize the impact of circumstances and arbitrary factors on individual outcomes [Cohen, 1989; Dworkin, 1981a,b; Rawls, 2009]. For instance within the classic context of employment, one (narrow) interpretation of EOP requires that desirable jobs are given to those persons most likely to perform well in them—e.g. those with the necessary education and experience—and not based on arbitrary reasons such as their race or family background. According to Rawls's (broader) interpretation of EOP, native talent and ambition can justify inequality in social positions, whereas circumstances of birth and upbringing such as sex, race, and social background can not. Many consider the distinction between morally acceptable and unacceptable inequality the most significant contribution of the egalitarian doctrine [Roemer and Trannoy, 2015].

Prior work in economics has sought to formally characterize conditions of equality of opportunity to allow for its precise measurement in practice (see e.g. [Fleurbaey, 2008; Roemer, 2009]). At a high level, in these models an individual's outcome/position is assumed to be affected by two main factors: their *circumstance* c and their desert[1]/effort e. Circumstance c is meant to capture all factors that are

[1] "Desert in philosophy is the condition of being deserving of something, whether good or bad." Wikipedia, entry on *Desert (philosophy)*.

deemed irrelevant or out of the scope of responsibility for the individual; for instance c could specify the socio-economic status he/she is born into. e captures all desert factors, that is, factors that can justify inequality. For the sake of concreteness, prior work in economics refers to e as effort. For any circumstance c and any effort level e, a policy ϕ induces a distribution of *utility* among people of circumstance c and effort e. Formally, an EOP policy will ensure that an individual's final utility will be, to the extent possible, only a function of their effort and not their circumstances.

While EOP has been traditionally discussed in the context of employment practices, its scope has been continually expanded to cover other areas, including lending, housing, college admissions, and beyond [Wikipedia, 2018]. Decisions made in these domains are increasingly automated and made through Algorithmic Data Driven Decision Making systems (A3DMs). We argue, therefore, that it is natural to study fairness for A3DMs through the lens of EOP. In this work, we draw a formal connection between recently proposed notions of fairness for supervised learning and economic models of EOP. Our work is based on the premise that in practice, predictive models inevitably make errors (e.g. the model may mistakenly predict the applicant won't pay back a loan). Sometimes these errors end up being beneficial to the subject, and sometimes they cause harm. We posit that in this context, EOP would require that individuals must have the same prospect of receiving this benefit/harm irrespective of their irrelevant characteristics.

More precisely, we assume that a person's features can be partitioned into two sets: those for which we consider it morally acceptable to hold him/her accountable, and those for which it is not so. We will broadly refer to the former set of attributes as the individual's *meritocratic* features, and the latter, as their *arbitrary* or *irrelevant* features. Note that there is considerable disagreement on the criteria to determine what factors should belong to each category. Roemer [1993] for instance proposes that societies decide this democratically. We take a neutral stance on this issue and leave it to domain experts and stake-holders to reach a resolution. Throughout, we assume this partition has been identified and is given.

We distinguish between an individual's *actual* and *deserved* utility when subjected to algorithmic decision making. We assume an individual's final condition or total (undeserved) utility as the result of being subject to A3DMs, is the difference between their actual and deserved utility (Section 2). Our main conceptual contribution is to map the supervised learning setting to economic models of EOP by treating predictive models as policies, irrelevant features as individual circumstance, and deserved utilities as desert/effort (Figure 1). We show that using this mapping many existing notions of fairness, such as predictive value parity [Kleinberg *et al.*, 2016] and equality of odds [Hardt *et al.*, 2016], can be interpreted as special cases of EOP. In particular, equality of odds is equivalent to fair EOP if we assume all individuals with the same true label are equally deserving (Section 3.1). Similarly, predictive value parity is equivalent to luck egalitarian EOP if the predicted label is assumed to reflect an individual's deserved utility (Section 4). In this respect, our work serves as a unifying framework for understanding existing notions of algorithmic fairness as special cases of EOP. Importantly, this framework allows us to explicitly spell out the moral assumptions underlying each notion of fairness, and interpret recent fairness impossibility results [Kleinberg *et al.*, 2016] in a new light. Last but not least, inspired by Roemer's model of egalitarian EOP we present a new, more general family of measures for algorithmic (un)fairness. We empirically show that employing the wrong measure of algorithmic fairness—that is, when the underlying assumptions of the measure are not satisfied—can have devastating consequences on the subjects' welfare.

1.1 Equality of Opportunity: An Overview

Equality of opportunity has been extensively debated among political philosophers. Philosophers such as Rawls [2009], Dworkin [1981a], Arneson [1989], and Cohen [1989] contributed to the egalitarian school of thought by proposing different criteria for making the cut between arbitrary and desert factors. The detailed discussion of these influential ideas is outside the scope of this work, and the interested reader is referred to excellent surveys by Arneson [2015] and Roemer and Trannoy [2015].

In this section, we briefly mention several prominent interpretations of EOP and discuss their relevance to A3DMs. Following Arneson [2018], we recount three main conceptions of equality of opportunity:

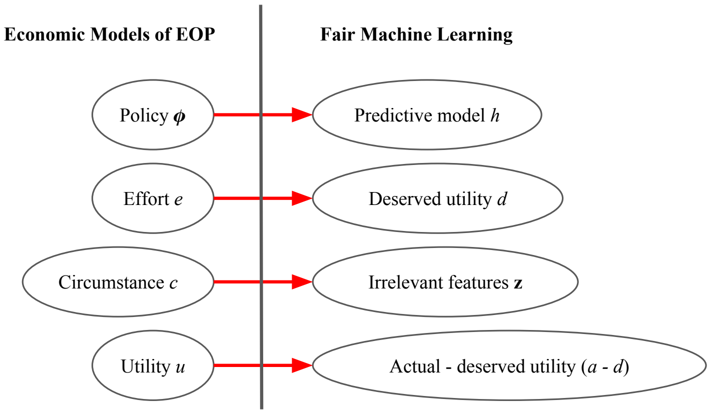


Figure 1: Our proposed conceptual mapping of the Fair ML literature to EOP.

- **Libertarian EOP:** A person is morally at liberty to do what she pleases with what she legitimately owns (e.g. self, business, etc.) as long as it does not infringe upon other people's moral rights—that is, the use of force, fraud, theft, or damage on persons or property of another individual is prohibited. Other than these restrictions, any outcome that occurs as the result of people's free choices on their legitimate possessions is considered just. In the context of A3DMs and assuming no gross violations of individuals' data privacy rights, this interpretation of EOP leaves the enterprise at total liberty to implement any algorithm it wishes for decision making. The algorithm can utilize all information available, including individuals' sensitive features such as race or gender, to make (statistically) accurate predictions.

- **Formal EOP:** Also known as "careers open to talents", formal EOP require desirable social positions to be open to all who possess the attributes relevant for the performance of the duties of the position (e.g. anyone who meets the formal requirements of the job) and wish to apply for them [Roemer, 2009]. The applications must be assessed only based on relevant attributes/qualifications that advances the morally innocent goals of the enterprise. Direct discrimination based on factors deemed arbitrary (e.g. race or gender) is therefore prohibited under this interpretation of EOP. Formal EOP would permit differences in people's circumstances—e.g. their gender—to have indirect, but nonetheless deep impact on their prospects. For instance, if women are less likely to receive higher education due to past social injustices, as long as a hiring algorithm applies the same educational requirement to male and female applicants and is blind to gender, formal equality of opportunity is maintained.

  In context of A3DMs, Formal EOP is equivalent to the removal of the sensitive feature information from the learning pipeline. In the fair ML community, this is sometimes referred to as "fairness through blindness".

- **Substantive EOP:** Substantive EOP moves the starting point of the competition for desirable positions further back in time, and requires not only open competition for desirable positions, but also fair access to the necessary qualifications for the position. This implies access to qualifications (e.g. formal requirements for a position) should not to be affected by arbitrary factors, such as race gender or social class. The concept is closely related to indirect discrimination: If the A3DM indirectly discriminates against people with a certain irrelevant feature (e.g. women or African-Americans) this may be an indication that the irrelevant/arbitrary feature has played a role in the acquisition of the requirements. When there are no better explanations of the indirect discrimination, it is considered in violation of substantive EOP.

Our focus in this work is on substantive EOP, and in particular, on two of its refinements, called Rawlsian EOP and Luck Egalitarian EOP.

Rawlsian EOP According to Rawls, those who have the same level of talent or ability and are equally willing to use them must have the same *prospect* of obtaining desirable social positions, regardless of arbitrary factors such as socio-economic background [Rawls, 2009]. This Rawlsian conception of EOP has been translated into precise mathematical terms as follows [Lefranc *et al.*, 2009]:[2] Let c denote circumstance, capturing factors that are not considered legitimate sources of inequality among individuals. Let scalar e summarize factors that are viewed as legitimate sources of inequality. (For the sake of brevity, the economic literature refer to e as "effort", but e is meant to capture an individual's overall desert). Let u specify individual utility, which is a consequence of effort, circumstance, and policy. Formally, let $F^{\phi}(.|c,e)$ specify the cumulative distribution of utility under policy ϕ at a fixed effort level e and circumstance c. Rawlsian/Fair EOP requires that for individuals with similar effort e, the distribution of utility should be the same—regardless of their circumstances:

Definition 1 (Rawlsian Equality of Opportunity (R-EOP)) *A policy ϕ satisfies fair EOP if for all circumstances c, c' and all effort levels e,*

$$F^{\phi}(.|c,e) = F^{\phi}(.|c',e).$$

Note that this conception of EOP takes an *absolutist* view of effort: it assumes e is a scalar whose absolute value is meaningful and can be compared across individuals. This view requires effort e to be inherent to individuals and not itself impacted by the circumstance c or the choice of the policy ϕ.

Luck Egalitarian EOP Unlike fair EOP, luck egalitarian EOP offers a *relative* view of effort, and allows for the possibility of circumstance c and implemented policy ϕ impacting the distribution of e. In this setting, Roemer [2002] argues that "in comparing efforts of individuals in different types, we should somehow adjust for the fact that those efforts are drawn from distributions which are different". As the solution he goes on to propose "measuring a person's effort by his rank in the effort distribution of his type/circumstance, rather than by the absolute level of effort he expends".

Formally, let $F_E^{c,\phi}$ be the effort distribution of type c under policy ϕ. Roemer argues that "this distribution is a characteristic of the type c, not of any individual belonging to the type. Therefore, an inter-type comparable measure of effort must factor out the goodness or badness of this distribution". Roemer declares two individuals as having exercised the same level of effort if they sit at the same quantile or rank of the effort distribution for their corresponding types/circumstances. More precisely, let the indirect utility distribution function $F^{\phi}(.|c,\pi)$ specify is the distribution of utility for individuals of type c at the πth quantile ($0 \leq \pi \leq 1$) of $F_E^{c,\phi}$. Equalizing opportunities means choosing the policy ϕ to equalize utility distributions, $F^{\phi}(.|c,\pi)$, across types, at fixed levels of π:[3]

Definition 2 (Luck Egalitarian Equality of Opportunity (e-EOP)) *A policy ϕ satisfies Luck Egalitarian EOP if for all $\pi \in [0,1]$ and any two circumstances c, c':*

$$F^{\phi}(.|c,\pi) = F^{\phi}(.|c',\pi).$$

To better understand the subtle different between fair EOP and luck egalitarian EOP, consider the following example: suppose in the context of employment decisions, we consider years of education as a merit feature, and gender as an arbitrary feature. Suppose Alice and Bob both have has 5 years of education, whereas Anna and Ben have 3 and 7 years of education, respectively. Rawlsian EOP would require Alice and Bob to have the same employment prospects, so it would ensure that factors such as sexism wouldn't affect Alice's employment chances, negatively (compared to Bob). Luck egalitarian EOP goes a step further and calculates everyone's rank (in terms of years of education) among all applicants of their gender. In our example, Alice is ranked 1st and Anna is ranked 2nd. Similarly, Bob is ranked 2nd and Ben is ranked 1st. A luck egalitarian EOP policy would ensure that Alice and Ben

[2] Note that in Rawls's formulation of FEO talent and ambition are treated as a legitimate source of inequality, even when they are independent of a person's effort and responsibility. The mathematical formulation proposed here includes talent, ability and ambition all in the scalar e. Whether natural talent should be treated as a legitimate source of inequality is a subject of controversy. As stated earlier, throughout this work we assume such questions have been already answered through a democratic process and/or deliberation among stakeholders and domain experts.

[3] Note that in Roemer's original work, utility is assumed to be a deterministic function of c, e, ϕ. Here we changed the definition slightly to allow for the possibility of non-deterministic dependence.

have the same employment prospects, and may indeed assign Bob to a less desirable position than Alice—even though they have similar years of education.

Next, we will discuss the above two refinements of substantive EOP in the supervised learning context.

2 Setting

As running example in this section, we consider a business owner who uses A3DM to make salary decisions so as to improve business productivity/revenue. We assume a higher salary is considered to be more desirable by all employees. An A3DM is designed to predict the salary that would improve the employee's performance at the job, using historical data. This target variable, as we will shortly formalize, does not always coincide with the salary the employee morally deserves.

We consider the standard supervised learning setting: A learning algorithm receives the training data set $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consisting of n instances, where $\mathbf{x}_i \in \mathcal{X}$ specifies the feature vector for individual i and $y_i \in \mathcal{Y}$, the true label for him/her (the salary that would improve his/her performance). Unless otherwise specified, we assume $\mathcal{Y} = [0, 1]$. Individuals are assumed to be sampled i.i.d. from a distribution F. The goal of a learning algorithm is to use the training data T to fit a *model* (or pick a hypothesis) $h : \mathcal{X} \rightarrow \mathcal{Y}$ that accurately predicts the label for new instances. Let $\mathcal{H}$ be the hypothesis class consisting of all the models the learning algorithm can choose from. A learning algorithm receives T as the input; then utilizes the data to select a model $h \in \mathcal{H}$ that minimizes some empirical loss, $\mathcal{L}(T, h)$. We denote the individual's predicted label by $\hat{y}$ (i.e. $\hat{y} = h(\mathbf{x})$).

Consider an individual who is subject to algorithmic decision making in this context. To discuss EOP, we begin by assuming his/her observable information, $\mathbf{x}$, can be partitioned into two disjoint sets, $\mathbf{x} = \langle \mathbf{z}, \mathbf{w} \rangle$, where $\mathbf{z} \in \mathcal{Z}$ denotes the individual's observable characteristics for which he/she is considered morally *not* responsible—this could include sensitive attributes such as race or gender, as well as less obvious attributes, such as zip code. We refer to $\mathbf{z}$ as *non-meritocratic*, *arbitrary*, or *irrelevant* features. $\mathbf{w} \in \mathcal{W}$ denotes observable characteristics that are deemed morally acceptable to hold the individual responsible for; in the running example, this could include the level of job-related education and experience. We refer to $\mathbf{w}$ as *meritocratic* or *relevant* features. We emphasize once again that determining what factors should belong to each category is entirely outside the scope of this work. We assume throughout that a resolution has been previously reached in this regard—through the appropriate process—and is given.

Let $d \in [0, 1]$ specify the individual's *deserved utility* (e.g. the utility they get if they receive their deserved salary). Deserved utility d is not directly observable, but we assume there exists an unknown function f, such that

$$d = f(\mathbf{x}, y, h).$$

That is, f links the observable information, $\mathbf{x}, y$, and h, to the deserved utility, d. Let $a \in [0, 1]$ be the *actual utility* the individual receives as the result of prediction $\hat{y}$ (e.g. the utility they get as the result of predicted salary). Throughout, for simplicity we assume higher values of a and d correspond to more desirable conditions.

Let u be the *advantage* or (undeserved) *utility* the individual earns as the result of being subject to predictive model h. We define u so that it captures the discrepancy between an individual's *actual* utility (a) and their *deserved/merited* utility d. In particular, we assume u has the following form:

$$u = (a - d).$$

With this formulation, an individual's utility is 0 when their actual and deserved utilities coincide (i.e. $u = 0$ if $a = d$).

We consider u to be the currency of equality of opportunity (i.e. it is what we hope to equalize across people with similar desert). Our moral argument for this choice is as follows: The predictive model h inevitably makes errors in assigning individuals to their deserved outcomes—this could be due to the target variable not properly reflecting desert, or simply a consequence of generalization. Sometimes these errors are beneficial to the subject, and sometimes they cause harm. u precisely

Notion of fairness	Deserved utility D	Actual utility A	Notion of EOP
Accuracy Parity	constant (e.g. 0)	$(\hat{Y} - Y)^2$	Rawlsian
Statistical Parity	constant (e.g. 1)	$\hat{Y}$	Rawlsian
Equality of Odds	Y	$\hat{Y}$	Rawlsian
Predictive Value Parity	$\hat{Y}$	Y	Egalitarian

Table 1: Interpretation of existing notions of algorithmic fairness for binary classification as special instances of EOP.

captures this benefit/harm. EOP in this setting would require that all individuals have the same prospect for earning this extra (undeserved) utility—regardless of their irrelevant attributes. As an example, let's assume the true labels in the training data reflects individuals' deserved utilities (as we will shortly argue, this assumption is not always morally acceptable, but for now let's ignore this). In this case, a perfect predictor—one that predicts the true label for every individual—will distribute no undeserved utility, but in real world applications, we almost never train such a model, so there will be some undeserved utility distributed among individuals. A fair model (with EOP rationale) would give everyone the same prospect for earning this utility—regardless of their irrelevant attributes.

Since d is in practice not directly observable, we can only make use of other available information (i.e. $\mathbf{x}, y, h$) to approximate it. Suppose we do this through a *deserved utility approximator* function $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}^+$ and assume $d \simeq g(\mathbf{x}, y, h)$. The approximate utility of the individual is, therefore,

$$u \simeq (a - g(\mathbf{x}, y, h)).$$

Our main conceptual contribution is to map the setting we just described to that of economic models of EOP (Section 1.1). We treat the predictive model h as a policy, non-merit features $\mathbf{z}$ as individual circumstances, and the deserved utilities d as effort/desert (Figure 1). In the next Section, we show that with this mapping, most existing statistical notions of fairness can be interpreted as special cases of EOP.

3 EOP for Supervised Learning

In this Section, we show that many existing notions of algorithmic fairness, such as statistical parity, equality of odds, equality of accuracy, and predictive value parity, can be cast as special cases of EOP. The summary of our results in this Section can be found in Table 1. To avoid any confusion with the notation, we define random variables $\mathbf{X}, Y, \hat{Y}$ to specify the feature vector, true label, and predicted label for an individual drawn i.i.d. from F. Similarly given the predictive model h, random variables A^h, D^h, U^h specify the actual utility, the deserved utility, and advantage, respectively, of an individual drawn i.i.d. from distribution F. When the predictive model in reference is clear from the context, we drop the superscript h for brevity.

3.1 Statistical Parity, Equality of Odds and Accuracy as Rawlsian EOP

We begin by translating Rawlsian EOP into the supervised learning setting using the mapping proposed in Figure 1. Recall that we proposed replacing e with deserved utility d, and circumstance c with vector of irrelevant features $\mathbf{z}$. In order for the definition of Rawlsian EOP to be morally acceptable, we need d to not be affected $\mathbf{z}$ and the model h. In other words, it can only a function of $\mathbf{w}$ and y. We define Rawlsian EOP for supervised learning as follows:

Definition 3 (R-EOP for supervised learning) *Suppose $d = f(\mathbf{w}, y)$. Predictive model h satisfies Rawlsian EOP if for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$ and all $d \in [0, 1]$,*

$$F^h(.|\mathbf{Z} = \mathbf{z}, D = d) = F^h(.|\mathbf{Z} = \mathbf{z}', D = d).$$

In the binary classification setting, if we assume the true label Y reflects an individual's deserved utility D, Rawlsian EOP translates into equality of odds across protected groups:[4]

Proposition 1 (Equality of Odds as R-EOP) *Consider the binary classification task where $\mathcal{Y} = \{0, 1\}$. Suppose $U = A - D$, $A = h(\mathbf{X}) = \hat{Y}$ (i.e., the actual utility is equal to the predicted label) and $D = g(\mathbf{X}, Y, h)$ where $g(\mathbf{X}, Y, h) = Y$ (i.e., deserved utility of an individual is assumed to be the same as their true label). Then the conditions of R-EOP are equivalent to those of equality of odds [Hardt* et al.*, 2016].*

Proof Recall that R-EOP requires that $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall d \in \mathcal{D}, \forall u \in \mathbb{R}:$

$$\mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}, D = d) = \mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}', D = d).$$

Replacing U with $(A - D)$, D with Y, A with $\hat{Y}$, the above is equivalent to

$$\begin{aligned}
& \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y \in \{0, 1\}, \forall u \in \{0, \pm 1\} : \mathbb{P}[\hat{Y} - Y \leq u | \mathbf{Z} = \mathbf{z}, Y = y] = \mathbb{P}[\hat{Y} - Y \leq u | \mathbf{Z} = \mathbf{z}', Y = y] \\
\Leftrightarrow \quad & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y \in \{0, 1\}, \forall u \in \{0, \pm 1\} : \mathbb{P}[\hat{Y} \leq u + y | \mathbf{Z} = \mathbf{z}, Y = y] = \mathbb{P}[\hat{Y} \leq u + y | \mathbf{Z} = \mathbf{z}', Y = y] \\
\Leftrightarrow \quad & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y \in \{0, 1\}, \forall \hat{y} \in \{0, 1\} : \mathbb{P}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}, Y = y] = \mathbb{P}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}', Y = y]
\end{aligned}$$

where the last line is identical to the conditions of equality of odds. ∎

The important role of the above proposition is to explicitly spell out the moral assumption underlying equality of odds as a measure of fairness: By measuring fairness through equality of odds, we implicitly assert *moral equivalence* among all individuals with the same true label in the training data set. This can clearly be problematic in practice: true labels rarely reflect desert. At best, they are only a reflection of the current state of affairs—which itself might be tainted by past injustices. For these reasons, we argue that equality of odds can only be used as a valid measure of algorithmic fairness (with an EOP rationale) once the validity of the above moral equivalency assumption has been carefully investigated and its implications are well understood in the context.

Other statistical definitions of algorithmic fairness—namely statistical parity and equality of accuracy—can similarly be thought of as special instances of R-EOP. See Table 1. For example statistical parity can be interpreted as R-EOP if we assume all individuals deserve the same utility (e.g. the highest utility).[5]

Proposition 2 (Statistical Parity as R-EOP) *Consider the binary classification task where $\mathcal{Y} = \{0, 1\}$. Suppose $U = A - D$, $A = \hat{Y}$ and $D = g(\mathbf{X}, Y, h)$ where $g(\mathbf{X}, Y, h)$ is a constant function (i.e., deserved utility of all individuals is assumed to be the same). Then the conditions of R-EOP is equivalent to statistical parity [Dwork* et al.*, 2012].*

Proof Without loss of generality, suppose $g(\mathbf{X}, Y, h) \equiv 1$, i.e. all individuals deserve utility 1. Recall that R-EOP requires that $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall d \in \mathcal{D}, \forall u \in \mathbb{R}:$

$$\mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}, D = d) = \mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}', D = d).$$

Replacing U with $(A - D)$, D with 1, and A with $\hat{Y}$, the above is equivalent to

$$\begin{aligned}
& \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall d \in \{1\}, \forall u \in \{0, -1\} : \mathbb{P}[\hat{Y} - D \leq u | \mathbf{Z} = \mathbf{z}, D = 1] = \mathbb{P}[\hat{Y} - D \leq u | \mathbf{Z} = \mathbf{z}', D = 1] \\
\Leftrightarrow \quad & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall u \in \{0, -1\} : \mathbb{P}[\hat{Y} \leq u + 1 | \mathbf{Z} = \mathbf{z}] = \mathbb{P}[\hat{Y} \leq u + 1 | \mathbf{Z} = \mathbf{z}'] \\
\Leftrightarrow \quad & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \{0, 1\} : \mathbb{P}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}] = \mathbb{P}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}']
\end{aligned}$$

where the last line is identical to statistical parity. ∎

Note that the moral assumption underlying statistical parity is that all individuals are deserving of the same utility. This assumption is justified when we believe the process by which different people earn different true labels, is driven mainly by arbitrary factors. For example, let's say we consider all

[4] Note that Hardt *et al.* [2016] referred to a weaker measure of algorithmic fairness (i.e. equality of true positive rates) as equality of opportunity.

[5] Statistical parity can be understood as equality of outcomes as well, if we assume $\hat{Y}$ reflects the outcome.

patients equally deserving of access to adequate clinical examinations. Now suppose that undergoing an invasive clinical examination has utility 1 if one has the suspected diseases and -1 otherwise, whereas avoiding the same clinical investigation has utility 1 if one does not have the suspected disease, and -1 otherwise. For all subjects, the deserved utility is the same (the maximum utility, let us suppose). In other words, all people with a disease deserve the invasive clinical investigation and all people without the disease deserve to avoid it. Consider a policy of giving clinical investigation to all the people without the disease and to no people without the disease. This would achieve an equal distribution of deserved utility (D) and distribute no (undeserved) utility U. Such policy, however, could only be achieved with a perfect accuracy predictor. For an imperfect accuracy predictor, R-EOP would require the distribution of (negative, in this case) utility (U) to give the same chance to African-Americans and whites with (without) the disease to receive (avoid) an invasive clinical exam.

Proposition 3 (Equality of Accuracy as R-EOP) *Consider the binary classification task where $\mathcal{Y} = \{0, 1\}$. Suppose $U = A - D$, $A = (\hat{Y} - Y)^2$ and $D = g(\mathbf{X}, Y, h)$ where $g(\mathbf{X}, Y, h) \equiv 0$ (i.e., deserved utility of all individuals are assumed to be the same and equal to 0). Then the conditions of R-EOP is equivalent to equality of accuracy.*

Proof Recall that R-EOP requires that $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall d \in \mathcal{D}, \forall u \in \mathbb{R}$:

$$\mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}, D = d) = \mathbb{P}(U \leq u | \mathbf{Z} = \mathbf{z}', D = d).$$

Replacing U with $(A - D)$, D with 0, and A with $(\hat{Y} - Y)^2$, the above is equivalent to $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall d \in \{0\}, \forall u \in \{0, 1\}$:

$$\mathbb{P}[(\hat{Y} - Y)^2 - D \leq u | \mathbf{Z} = \mathbf{z}, D = d] = \mathbb{P}[(\hat{Y} - Y)^2 - D \leq u | \mathbf{Z} = \mathbf{z}', D = d]$$

We can then write:

$$\begin{aligned} &\Leftrightarrow \quad \forall \mathbf{z}, \mathbf{z}', \forall u : \mathbb{P}[(\hat{Y} - Y)^2 = u | \mathbf{Z} = \mathbf{z}] = \mathbb{P}[(\hat{Y} - Y)^2 = u | \mathbf{Z} = \mathbf{z}'] \\ &\Leftrightarrow \quad \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z} : \mathbb{E}[(\hat{Y} - Y)^2 | \mathbf{Z} = \mathbf{z}] = \mathbb{E}[(\hat{Y} - Y)^2 | \mathbf{Z} = \mathbf{z}'] \end{aligned}$$

where the last line is identical to equality of accuracy. ■

Equality of accuracy is equivalent to EOP if we assume error reflects the undeserved utility distributed by the predictive model. This exposes the fundamental ethical problem with adopting equality of accuracy as a measure of algorithmic fairness: it fails to distinguish between errors that are beneficial to the subject and those that are harmful. For example, in the salary prediction example, accuracy parity would make no distinction between an individual who earns a salary higher than what they deserve, and someone who earns lower than their deserved salary.

We remark that predictive value parity (equality of false discovery and omission rates across sensitive groups) can not be thought of as an instance of R-EOP, as it requires the deserved utility to be a function of the predictive model h (more precisely, it assumes $D = h(\mathbf{X})$). This is in violation of the absolutist view of Rawlsian EOP. Next, we will show that predictive value parity can be cast as an instance of luck egalitarian EOP.

3.2 Predictive Value Parity as Egalitarian EOP

Next, we will specialize Roemer's model of Egalitarian EOP to the supervised learning setting. Recall that egalitarian EOP allows the deserved utility to be a function of the predictive model h, that is $D = f(\mathbf{X}, Y, h)$. When this is the case, following the argument put forward by Roemer we posit that the distribution of deserved utility for a given type $\mathbf{z}$ (denoted by $F_D^{\mathbf{z},h}$) is a characteristic of the type $\mathbf{z}$, not of any individual belonging to the type. Therefore, an inter-type comparable measure of deserved utility must factor out the goodness or badness of this distribution. We consider two individuals as being equally deserving if they sit at the same quantile or rank of the distribution of D for their corresponding type.

More formally, we can compute the *indirect utility distribution* function $F^h(.|\mathbf{z}, \pi)$, which is the distribution of utility for individuals of type $\mathbf{z}$ at the πth quantile $(0 \leq \pi \leq 1)$ of $F_D^{\mathbf{z},h}$. Equalizing opportunities means choosing the predictive model h to equalize this indirect utility distribution across types, at fixed levels of π.

Definition 4 (e-EOP for supervised learning) *Suppose $d = f(\mathbf{x}, y, h)$. Predictive model h satisfies egalitarian EOP if for all $\pi \in [0,1]$ and $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,*

$$F^h(.|\mathbf{z}, \pi) = F^h(.|\mathbf{z}', \pi). \tag{1}$$

Next, we show that predictive value parity can be thought of as a special case of e-EOP, where the predicted label $h(\mathbf{x})$ is assumed to reflect the individual's deserved utility.

Proposition 4 (predictive value parity as e-EOP) *Consider the binary classification task where $\mathcal{Y} = \{0,1\}$. Suppose $U = A - D$, $A = Y$ and $D = g(\mathbf{X}, Y, h)$ where $g(\mathbf{X}, Y, h) = h(\mathbf{X}) = \hat{Y}$ (i.e., deserved utility of an individual under h is assumed to be the same as their predicted label). Then the conditions of e-EOP are equivalent to those of predictive value parity [Berk* et al.*, 2017].*

Proof Recall that e-EOP requires that $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \pi \in [0,1], \forall u \in \mathbb{R}:$

$$\mathbb{P}[U \leq u | \mathbf{Z} = \mathbf{z}, \Pi = \pi] = \mathbb{P}[U \leq u | \mathbf{Z} = \mathbf{z}', \Pi = \pi].$$

Note that since $D = \hat{Y}$ and in the binary classification, $\hat{Y}$ can only take on two values, there are only two ranks/quantiles possible in terms of the deserved utility—corresponding to $\hat{Y} = 0$ and $\hat{Y} = 1$. So the above condition is equivalent to $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \{0,1\}, \forall u \in \{0, \pm 1\}:$

$$\mathbb{P}[U \leq u | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}[U \leq u | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}].$$

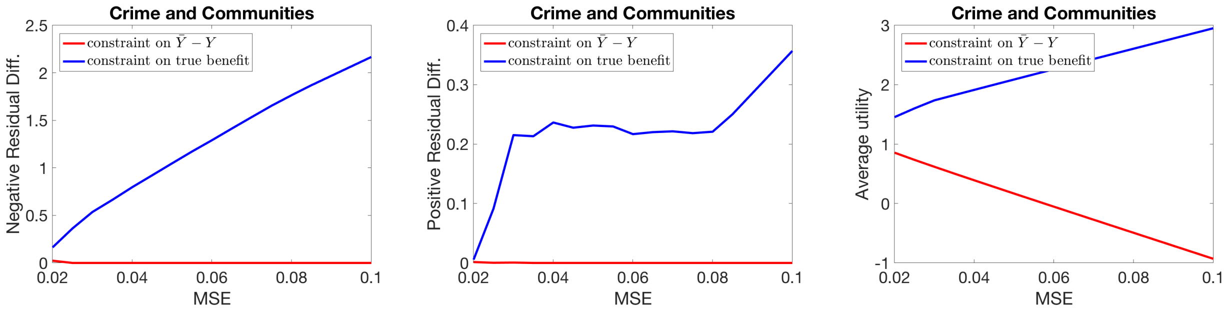
Replacing U with $(A - D)$, D with $\hat{Y}$, A with Y, the above is equivalent to

$$\begin{aligned}
& \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \{0,1\}, \forall u \in \{0, \pm 1\} : \mathbb{P}[Y - \hat{Y} \leq u | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}[Y - \hat{Y} \leq u | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}] \\
\Leftrightarrow \; & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \{0,1\}, \forall u \in \{0, \pm 1\} : \mathbb{P}[Y \leq u + \hat{y} | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}[Y \leq u + \hat{y} | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}] \\
\Leftrightarrow \; & \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \{0,1\}, \forall y \in \{0,1\} : \mathbb{P}[Y = y | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}[Y = y | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}]
\end{aligned}$$

where the last line is identical to predictive value parity. ∎

The choice of prediction $h(\mathbf{X})$ as the indicator of desert, may sound odd at first. But we argue that in certain settings this is indeed an appropriate assumption. For instance, consider the case of criminal risk assessment; predictive value parity assumes that those who are equally likely to reoffend (according to the best available predictive model h) are deserving of the same utility—even though some of them will end up actually reoffending, and the rest won't. This is an acceptable assumption if we believe that differences in actual outcome (re-offence in this example) among equally risky individuals is mainly driven by arbitrary factors such as brute luck—of course such factors should never specify desert. In cases like this, the potential/risk to be a criminal—as opposed to the actual outcome—can justify unequal treatment. We emphasize again that the plausibility of such moral assumptions must be critically evaluated in a given context before predictive parity is employed to ensure fairness.

On Recent Fairness Impossibility Results Several papers have recently shown that statistical measures of fairness, such as predictive value parity and equality of odds, are generally incompatible with one another and cannot hold simultaneously [Kleinberg *et al.*, 2016; Friedler *et al.*, 2016]. Our approach confers a moral meaning to the impossibility results: they can be interpreted as contradictions between fairness desiderata reflecting different and irreconcilable moral assumptions. For example predictive value parity and equality of odds make very different assumptions about the deserved utility d: Equality of odds assumes all persons with similar true labels are equally deserving, whereas predictive value parity assumes all persons with the same predicted label/risk are equally deserving. Note that depending on the context, usually only one (if any) of these assumptions is morally acceptable. We argue, therefore, that unless we are in the highly special case where $Y = h(\mathbf{X})$, it is often unnecessary—from a moral standpoint—to ask for both of these fairness criteria to be satisfied simultaneously.



(a)

(b)

(c)

Figure 2: NRD, PRD, and Social Welfare as a function of ϵ (the upperbound on mean squared error). The choice of the utility levels greatly impacts the welfare of subjects.

4 Experiments with Egalitarian Measures of Fairness

In this section, inspired by Roemer's model of egalitarian EOP we present a new family of measures for algorithmic (un)fairness that are applicable to settings beyond binary classification. We empirically show that employing a measure of algorithmic fairness when its underlying assumptions are not met, can have devastating consequences on the total welfare of the population.

4.1 A New Family of Measures

In settings more complicated than binary classification (e.g. regression), the requirement of equation 1 becomes too stringent (there will be infinitely many quantiles to equalize utilities over). The problem persists even if we relax the requirement from equal utility distributions to maximizing the minimum expected utility at each rank[6]. More precisely, let $v^{\mathbf{z}}(\pi, h)$ specify the expected utility of individuals of type $\mathbf{z}$ at the πth quantile of the deserved utility distribution. Let's say a predictive model h^{π} satisfies Egalitarian EOP at the π-slice for $\pi \in [0, 1]$, if:

$$h^{\pi} \in \arg\max_{h \in \mathcal{H}} \min_{\mathbf{z} \in \mathcal{Z}} v^{\mathbf{z}}(\pi, h).$$

If we are concerned only with the π-slice of individuals, then h^{π} would be the equal-opportunity predictive model. Unfortunately, beyond binary classification, we generally won't be able to find a model that is optimal for all ranks $\pi \in [0, 1]$. Therefore, we need to find a compromise. Following Roemer, we propose the following remedy:[7]

$$h^* \in \arg\max_{h \in \mathcal{H}} \min_{\mathbf{z} \in \mathcal{Z}} \int_0^1 v^{\mathbf{z}}(\pi, h) d\pi \tag{2}$$

That is, we consider h^* to be an e-EOP predictive model if it maximizes the average utility of the worst off group. Replacing the integral with its in-sample analogue, our proposed family of e-EOP measures can be evaluated as follows:

$$\mathcal{F}(h, T) = \min_{\mathbf{z} \in \mathcal{Z}} \frac{1}{n_{\mathbf{z}}} \sum_{i \in T : \mathbf{z}_i = \mathbf{z}} u(\mathbf{x}_i, y_i, h) \tag{3}$$

where $u(\mathbf{x}_i, y_i, h)$ is the utility an individual with feature vector $\mathbf{x}_i$ and true label y_i receives when we employ predictive model h.

[6] The motivation for max-min (as opposed to equality) is to address the "leveling down" objection: the argument is that disadvantaged individuals are more interested in maximizing their absolute level of utility, as opposed their relative utility compared to the advantaged.

[7] He in fact proposes 2 further alternatives: In the first solution, the objective function for each π-slice of the population is assumed to be $\min_{\mathbf{z} \in \mathcal{Z}} v^{\mathbf{z}}(\pi, h)$—which is then weighted by the size of the slice. In the second solution, he declares the equal opportunity policy to be the average of the policies h^{π}. Roemer expresses no strong preference for any of these alternatives, other than the fact that computational simplicity sometimes suggests one over the others [Roemer, 2002].

For simplicity, in the rest of this work, we assume utility has the following functional dependence on $\mathbf{x}$ and h: $u(\mathbf{z}, y, h(\mathbf{x}))$. That is, u's dependence on $\mathbf{x}$ and h are through $\mathbf{z}$ and $h(\mathbf{x})$, respectively. Suppose we have m (intersectional) groups $G_1, \cdots, G_m$ based on the values $\mathbf{z}$ can take on. We propose solving the following optimization problem: maximize the average utility of the worst off group (σ), subject to error being less than a certain upper bound (ϵ).

$$\begin{aligned} \max_{\sigma, h \in \mathcal{H}} \quad & \sigma \\ \text{s.t.} \quad & \frac{1}{n_g} \sum_{i \in G_g} u(\mathbf{x}_i, y_i, h(\mathbf{x}_i)) \geq \sigma \qquad \forall g = 1, \cdots, m \\ & \mathcal{L}(T, h) \leq \epsilon \end{aligned} \tag{4}$$

Note that if the loss function $\mathcal{L}$ is convex and u is concave in model parameters, this is a convex optimization problem that can be solved efficiently.

Next, as concrete example we illustrate a linear (in $h(\mathbf{x})$) instance of our proposed measure in a regression setting with $\mathcal{Y} = [0, 1]$. We start by specifying the utilities associated with the following $(y, \hat{y})$-pairs for each type $\mathbf{z}$: $(1, 1), (1, 0), (0, 1), (0, 0)$. We then derive the linear utility function corresponding to these values using the following lemma:

Lemma 1 *For $y, \hat{y} \in \{0, 1\}$, let $b^{\mathbf{z}}_{y,\hat{y}} \in \mathbb{R}$ be arbitrary constants specifying the utility an individual with irrelevant features $\mathbf{z}$ and ground truth label y receives when they are assigned label $\hat{y}$. Then there exists a linear utility function of form $c^{\mathbf{z}}_y \hat{y} + d^{\mathbf{z}}_y$ such that for all $y, \hat{y} \in \{0, 1\}$, $b^{\mathbf{z}}(y, \hat{y}) = b^{\mathbf{z}}_{y,\hat{y}}$.*

Proof Solving the following system of equations,

$$\forall y, \hat{y} \in \{0, 1\} : c^{\mathbf{z}}_y \hat{y} + d^{\mathbf{z}}_y = b^{\mathbf{z}}_{y,\hat{y}}$$

we obtain: $c^{\mathbf{z}}_0 = b^{\mathbf{z}}_{0,1} - b^{\mathbf{z}}_{0,0}$, $c^{\mathbf{z}}_1 = b^{\mathbf{z}}_{1,1} - b^{\mathbf{z}}_{1,0}$, $d^{\mathbf{z}}_0 = b^{\mathbf{z}}_{0,0}$, and $d^{\mathbf{z}}_1 = b^{\mathbf{z}}_{1,0}$. ∎

For example, suppose $b^0_{00} = 1, b^0_{10} = 2, b^0_{01} = 0.5, b^0_{11} = 1$ and $b^1_{00} = 1, b^1_{10} = 0, b^1_{01} = 5, b^1_{11} = 2$. Then using the above lemma, we obtain the following utility functions $u^0(y, \hat{y}) = -0.5\hat{y} - 0.5\hat{y}y + y + 1$ and $u^1(y, \hat{y}) = 4\hat{y} - 2\hat{y}y - y + 1$.

4.2 Experiments

Next, we solve optimization (4) on a real-world regression data set, and show that employing the wrong utility function can have devastating consequences in terms of the total welfare of the population. We ran our experiments on the *Crime and Communities data set* [Lichman, 2013]. The data consists of 1994 observations each made up of 101 features, and it contains socio-economic, law enforcement, and crime data from the 1995 FBI UCR. Community type (e.g. urban vs. rural), average family income, and the per capita number of police officers in the community are a few examples of the explanatory variables included in the dataset. The target variable is the "per capita violent crimes".

We preprocess the original dataset as follows: we remove the instances for which target value was unknown. Also, we remove features whose values are missing for more than 80% of instances. We standardize the data so that each feature has mean 0 and variance 1. We divide all target values by a constant so that labels range from 0 to 1. Furthermore, we flip all labels to make sure higher y values correspond to more desirable outcomes. We assume a neighborhood belongs to the protected group if and only if the majority of its residents are non-Caucasian, that is, the percentage of African-American, Hispanic, and Asian residents of the neighborhood combined, is above 50%. This divides the training instances into two groups G_0, G_1. We include this group membership information as the sensitive feature z in the dataset ($z_i = 1[i \in G_1]$).

We assume the true utility has the following form:[8]

$$u^0(y, \hat{y}) = -0.5\hat{y} - 0.5\hat{y}y + y + 1,$$

$$u^1(y, \hat{y}) = 4\hat{y} - 2\hat{y}y - y + 1.$$

[8] We tried several different utility functions, and the results where qualitatively similar.

(Recall that the above utility functions are derived from $b^0_{00} = 1, b^0_{10} = 2, b^0_{01} = 0.5, b^0_{11} = 1$ and $b^1_{00} = 1, b^1_{10} = 0, b^1_{01} = 5, b^1_{11} = 2$.) We solve the following two optimization problems:

$$\begin{aligned} \max_{\sigma,\boldsymbol{\theta}} \quad & \sigma \\ \text{s.t.} \quad & \frac{1}{n_0}\sum_{i\in G_0} -0.5\boldsymbol{\theta}.\mathbf{x}_i - 0.5(\boldsymbol{\theta}.\mathbf{x}_i)y_i + y_i + 1 \geq \sigma \\ & \frac{1}{n_1}\sum_{i\in G_1} 4\boldsymbol{\theta}.\mathbf{x}_i - 2(\boldsymbol{\theta}.\mathbf{x}_i)y_i - y_i + 1 \geq \sigma \\ & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}.\mathbf{x}_i - y_i)^2 + \lambda\|\boldsymbol{\theta}\|_1 \leq \epsilon \end{aligned} \tag{5}$$

In Optimization 5, we put a lower bound on the true utility functions.

$$\begin{aligned} \max_{\sigma,\boldsymbol{\theta}} \quad & \sigma \\ \text{s.t.} \quad & \frac{1}{n_0}\sum_{i\in G_0} \boldsymbol{\theta}.\mathbf{x}_i - y_i \geq \sigma \\ & \frac{1}{n_1}\sum_{i\in G_1} \boldsymbol{\theta}.\mathbf{x}_i - y_i \geq \sigma \\ & \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}.\mathbf{x}_i - y_i)^2 + \lambda\|\boldsymbol{\theta}\|_1 \leq \epsilon \end{aligned} \tag{6}$$

In Optimization 6, we put a lower bound on $\sum_i(\hat{y}_i - y_i)$.

We choose the value of λ by running a 10-fold cross validation on the data set. We then measure the following quantities via 5-fold cross validation:

- **Positive residual difference** [Calders *et al.*, 2013] is the equivalent of false positive rate in regression, and is computed by taking the absolute difference of mean negative residuals across groups.
- **Negative residual difference** [Calders *et al.*, 2013] is the equivalent of false negative rate in regression, and is computed by taking the absolute difference of mean negative residuals across groups.
- **Average utility** or social welfare is computed by taking the average utility of all individuals in the test data set.

Figure 2 shows the results of our experiments. As evident in Figures 2a and 2b, by enforcing a lower bound on $\sum_i(\hat{y}_i - y_i)$, positive and negative residual difference indeed go to 0 very quickly. This does not hold when we Iower-bound the true average utilities—in fact, in this case, positive and negative residual difference increase with ϵ. These trends are reversed in Figure 2c where we measure social welfare: by enforcing a lower bound on $\sum_i(\hat{y}_i - y_i)$, the trained model performs very poorly in terms of average utility of the population. Once we replace $\sum_i(\hat{y}_i - y_i)$ with the true utility function, social welfare improves significantly with ϵ.

5 Conclusion

Our work makes an important contribution to the quickly growing line of work on algorithmic fairness—by providing a unifying moral framework for understanding existing notions through philosophical interpretations and economic models of EOP. We showed that the choice between statistical parity, equality of odds, and predictive value parity can be mapped systematically to specific moral assumptions about what individuals morally deserve. We argue that determining meritocratic features and moral desert is outside the expertise of computer scientists, and has to be resolved through the appropriate process with input from stakeholders and domain experts. Indeed, in any given context reasonable people may disagree on what constitutes merit/desert, and there will rarely be a consensus on the most suitable notion of fairness. This, however, does not imply that all existing notions are equally acceptable from a moral standpoint.

References

Richard J. Arneson. Equality and equal opportunity for welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 56(1):77–93, 1989.

Richard J. Arneson. Equality of opportunity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2015 edition, 2015.

Richard J. Arneson. Four conceptions of equal opportunity. 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80. IEEE, 2013.

Gerald A. Cohen. On the currency of egalitarian justice. *Ethics*, 99(4):906–944, 1989.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

Ronald Dworkin. What is equality? part 1: Equality of welfare. *Philosophy & Public Affairs*, 10(3):185–246, 1981.

Ronald Dworkin. What is equality? part 2: Equality of resources. *Philosophy & Public Affairs*, 10(4):283–345, 1981.

Marc Fleurbaey. *Fairness, responsibility, and welfare.* Oxford University Press, 2008.

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

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

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

Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. Equality of opportunity and luck: Definitions and testable conditions, with an application to income in france. *Journal of Public Economics*, 93(11-12):1189–1207, 2009.

M. Lichman. UCI machine learning repository: Communities and crime data set. `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`, 2013.

John Rawls. *A theory of justice.* Harvard university press, 2009.

John E. Roemer and Alain Trannoy. Equality of opportunity. In *Handbook of income distribution*, volume 2, pages 217–300. Elsevier, 2015.

John E. Roemer. A pragmatic theory of responsibility for the egalitarian planner. *Philosophy & Public Affairs*, pages 146–166, 1993.

John E. Roemer. Equality of opportunity: A progress report. *Social Choice and Welfare*, 19(2):455–471, 2002.

John E. Roemer. *Equality of opportunity.* Harvard University Press, 2009.

Amartya Sen. Equality of what? *The Tanner Lecture on Human Values*, 1979.

Wikipedia. Equal opportunity. `https://en.wikipedia.org/wiki/Equal_opportunity`, 2018.