Normative Principles for Evaluating Fairness in Machine Learning

Derek Leben
leben@pitt.edu
University of Pittsburgh, Johnstown

ABSTRACT

There are many incompatible ways to measure fair outcomes for machine learning algorithms. The goal of this paper is to characterize rates of success and error across protected groups (race, gender, sexual orientation) as a distribution problem, and describe the possible solutions to this problem according to different normative principles from moral and political philosophy. These normative principles are based on various competing attributes within a distribution problem: intentions, compensation, desert, consent, and consequences. Each principle will be applied to a sample risk-assessment classifier to demonstrate the philosophical arguments underlying different sets of fairness metrics.

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; *Machine learning*.

KEYWORDS

fairness, machine learning, political philosophy, discrimination, algorithmic decision making

ACM Reference Format:
Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Societyety (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375808

1 INTRODUCTION

Machine learning algorithms are increasingly being used in both the public and private sectors to make decisions about hiring, financial applications, college admissions, medical diagnoses, and prison sentences. The algorithms we are most concerned with are classifiers that produce a binary output like "approve/reject" for loan applications, "dangerous/safe" for prisoners, "hire/pass" for job applicants, "malignant/benign" for medical diagnoses, and so on. The appeal of these algorithms is clear; they can vastly increase the efficiency, accuracy, and consistency of decisions. Because they are playing a role in the distribution of important resources and opportunities, we are obligated to ensure that algorithms are also free of discriminatory bias towards historically underrepresented groups. However, there are many ways of measuring whether a binary classifier is indeed free of bias towards these protected groups, and as Kleinberg et al. [11] and Chouldechova [3] have demonstrated, it is mathematically impossible to achieve parity according to every fairness metric. Under these circumstances, Binns [2] notes that normative principles from moral and political philosophy may be required to justify which design choices were made. This paper is an attempt to take a more detailed step in this direction. If we characterize rates of success and error as a distribution problem, then there are very specific prescriptions from moral and political philosophy that can be used to justify using one set of fairness metrics over another.

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375808

2 PARITY METRICS IN ML

If we have a binary classifier, $r(X)$, that is trained on some data set, (x, y), where x is a vector of input values and y is the classification (either 0 or 1), then $r(X)$ will give us the predicted category, $\hat{y}_i$, for some new set of data, x_i. For example, an algorithm trained on images of skin marks that are labeled as '1= cancer' or '0= no cancer' will produce a judgment about whether or not new skin marks are cancerous. Here, the x-values are patterns of pixels, the y-values are whether the image is labeled as cancerous or not, and $r(X)$ is a score that will produce a classification of $\hat{y} = 1$ for cancer and $\hat{y} = 0$ for no cancer. The outputs of the model can then be evaluated by the match between predicted categories and actual categories of new data. Because there are two possible values for the predicted and actual values, the comparisons fall into the categories of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

The ideal model will have all its outputs in the categories of TP and TN, but that is unrealistic. All classifiers will have some amounts of error, and the question is how to evaluate the success of the model (the TP and TN results) relative to this error. Considering the rates of one outcome within another set of outcomes that contains it will generate a conditional probability. For instance, considering the rate of True Positives within all positive predictions will give the *positive predictive rate*, which is also the probability: $p(y = 1|\hat{y} = 1)$. This metric is telling us how many positive predictions actually have the targeted trait. On the other hand, we could also ask how many of the negative predictions actually lack the trait, which is the *negative predictive rate*: $p(y = 0|\hat{y} = 0)$.

To determine whether a classifier is unfair in its treatment of Group A compared with Group B, we might compare these success or error rates for both groups. If α_X is a success or error rate α for Group X, then the most important parity metrics are as follows:

Demographic Parity:

$$p(\hat{y}_a = 1) = p(\hat{y}_b = 1)$$
$$\frac{TP_A + FP_A}{FN_A + TN_A} = \frac{TP_B + FP_B}{FN_B + TN_B}$$

Positive Predictive Parity:

$$p(y_a = 1|\hat{y}_a = 1) = p(y_b = 1|\hat{y}_b = 1)$$
$$\frac{TP_A}{TP_A + FP_A} = \frac{TP_B}{TP_B + FP_B}$$

Negative Predictive Parity:

$$p(y_a = 0|\hat{y}_a = 0) = p(y_b = 0|\hat{y}_b = 0)$$
$$\frac{TN_A}{TN_A + FN_A} = \frac{TN_B}{TN_B + FN_B}$$

False Positive Parity:

$$p(\hat{y}_a = 1|y_a = 0) = p(\hat{y}_b = 1|y_b = 0)$$
$$\frac{FP_A}{TN_A + FP_A} = \frac{FP_B}{TN_B + FP_B}$$

Equality of Opportunity:

$$p(\hat{y}_a = 1|y_a = 1) = p(\hat{y}_b = 1|y_b = 1)$$
$$\frac{TP_A}{TP_A + FN_A} = \frac{TP_B}{TP_B + FN_B}$$

Let's apply these parity rates to a sample data set. Imagine we have trained a risk-assessment algorithm for assisting parole judgments, and we want to evaluate whether it is unfair in its treatment of white and black prisoners. Say that there are a total of 600 white prisoners (Group A) and 200 black prisoners (Group B). The outcomes for each group are represented in Figure 1.

	Reoffends (y=1)	No Reoffense (y=0)
High Risk ($\hat{y} = 1$)	$TP_b =$ 90 / $TP_a =$ 270	$FP_b =$ 10 / $FP_a =$ 30
Low Risk ($\hat{y} = 0$)	$FN_b =$ 50 / $FN_a =$ 20	$TN_b =$ 50 / $TN_a =$ 280

Figure 1: A sample risk-assessment classifier

Is this algorithm *unfair* to white or black prisoners? The answer depends on which parity metrics we are using to evaluate fairness. In terms of positive predictive value, the model treats both demographic groups equally: 90% for white prisoners (270/300) and also 90% for black prisoners (90/100). This means that it correctly predicts reoffense at the same rate for both groups, which was how Northpointe defended its COMPAS algorithm against allegations of bias. More generally, this sample classifier also satisfies *demographic parity*, since 50% of both black and white prisoners are labeled "high risk." Feldman et al. [6] argue that a modified version of this metric satisfies the legal standards for non-discrimination in disparate impact law. By these metrics, one could argue that the algorithm is both ethically and legally fair.

By other metrics, one could claim that the algorithm is unfair. One of these metrics is inequality in the peaceful prisoners who were incorrectly identified (the *false positive rate*), which is 16% (10/60) for black prisoners, but only 9.6% (30/310) for white prisoners. This was the metric that Pro Publica used in their allegation that the COMPAS assessment is biased. Another way that the algorithm can be called unfair is by appealing to inequality in the rates of dangerous prisoners who were *correctly* identified from both groups, which is 93% for whites (270/290) but only 64% for blacks (90/140). Because this enables those who actually have the trait to be correctly identified, we might call this *equality of opportunity* (that label obviously works better in settings where the positive condition is something beneficial, such as those who are actually qualified for a job being hired). A combination of both equal false positive rates and equality of opportunity has been called "equalized odds" by Hardt et al. [7].

Given that we cannot possibly achieve parity in all of the metrics, which ones should we care more about? To develop arguments for why we should prefer one rate to another, we must turn to normative principles for distributing goods.

3 NORMATIVE DISTRIBUTION PRINCIPLES

Following the recommendations of Binns (2018), this paper will characterize rates of success and error between groups as a distribution problem. If this characterization is apt, then the parity metrics described above can be evaluated by how closely they approximate normative principles of fair distribution. In moral and political philosophy, normative principles can be categorized into two families, Consequentialist and Deontological (Figure 3). This section will summarize these principles; further detail can be found from a philosophical perspective in [16] and a welfare economics perspective in [13].

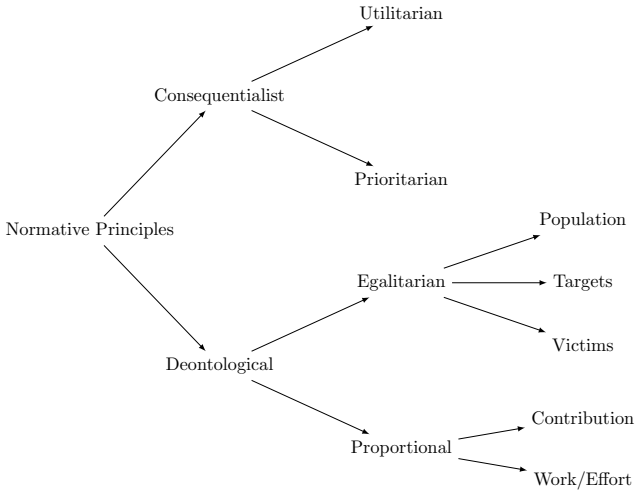

Figure 2: Families of Normative Distribution Principles

Assume we are a central planner with some amount of total goods, S, that we wish to distribute among a population of two people, Alice and Bob. Consequentialist approaches will construct some function that models each potential distribution and its effects on both Alice and Bob, in other words, a utility function: $u(d_i) \rightarrow (v_i)$, where $(d_1 \dots d_i)$ are vectors that represent possible allocations such that each vector sums to S, and $(v_1 \dots v_i)$ are measurements of well-being. The utility function is interpreted as a measure of happiness outcomes for Alice and Bob. This utility function can be discounted by a coefficient, w_i, which appropriately modifies the utility of these patients based on need or relative value. Equipped with a properly discounted utility function, a Consequentialist will then run some selection-procedure over aggregate utilities; the two most popular candidates being either maximizing the sum for Utilitarians [12], or maximizing the minimum for Prioritarians [1]:

Utilitarian Principle (Maxisum):

$$\text{UT}(d_i) = \arg\max_{x_i} \sum_{i=1}^{n} w_i u(d_i, v_i)$$

Prioritarian Principle (Maximin):

$$\text{PR}(d_i) = \arg\max_{x_i} \min_{x_i} w_i u(d_i, v_i)$$

In distributing integer divisions of \$10 to Alice and Bob, a Prioritarian would clearly select an allocation of (5,5). If there are no weights attached to outcomes, then Utilitarians will be indifferent between all allocations, since all potential allocations sum to \$10. Yet if we weight the outcomes to account for a diminishing marginal utility, or include an envy-weight for both players when the other receives a greater share, then Utilitarians will also select (5,5).

Deontological approaches, by contrast, will evaluate distributions based on rights. An *Egalitarian* approach confers equal rights, and thus equal shares, to every member of the population. Equal rights can also be distributed only to a subset of the population who qualify, perhaps as reward to the *intended* beneficiaries [10], or as *compensation* to those who make sacrifices for the process [14]. Other deontological approaches do not grant equal rights across a segmented population, but rather, adjust the rights (and allocations) of each person proportionally, based on their contributions to production.

Formally, Deontological principles can be characterized by a loss function, $\ell(d_i, F)$, that measures the distance between distribution outcomes and some fairness standard. The simplest fairness standard is based only on the size of the population, where F is simply each individual's expected value of the total goods, $\frac{1}{n}S$. If we are distributing \$10 in integer amounts to Alice and Bob, the Egalitarian distribution will be five dollars to each. More generally, when we can't achieve a perfectly equal distribution, the Egalitarian will prefer a distribution which minimizes the distance to this fairness standard:

Egalitarian Principle:

$$\text{EG}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{1}{n}S)$$

Proportional Deontological principles will distribute based on factors that went into the process of production, such as: the resources from each member of the population that went into production, $R_i(d_i)$, the amount of actual work that went into the deployment of those resources, $W_i(d_i)$, and the amount of luck that went into deployment of those resources, $L_i(d_i)$. Let's assume that luck and work exhaust the possible sources of one's resources, so that $R = L + W$ (resources = luck + work). Libertarians like Rand [15] focus on each person's total contribution at the time of consent, where a wealthier person who received most of her funds through luck is still entitled to a larger share than a scrappy young upstart who works harder but contributes less to production. By contrast, *Desert-Based* approaches [17] consider rights to be proportionate to individual effort, and the effects of luck on allocations are to be minimized.[1] These two principles are summarized as:

Libertarian Principle:

$$\text{LB}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{R_x}{\sum R_i}S)$$

Desert Principle:

$$\text{DS}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{W_x}{\sum W_i}S)$$

For instance, let's say that Alice contributed 3 dollars to the production and Bob only contributed 2 dollars. Libertarians would determine that Alice has a right to $\frac{3}{5}$ of the goods, while Bob only has a right to $\frac{2}{5}$ of it, and the correct distribution outcome is (6,4). On the other hand, if it turns out that Alice only worked to generate 1 dollar of her investment, while Bob worked hard to generate all of his investment, then Desert Principles would determine that Bob is owed $\frac{2}{3}$ of the goods, while Alice is only entitled to $\frac{1}{3}$ of it, so the central planner must randomize between the equally good integer allocations of (3,7) or (4,6).

Defining normative distribution principles as optimization problems makes it relatively straightforward to insert them into machine learning algorithms. For Consequentialists, fairness principles get built into the data itself, since every set of inputs and classifications, (x_1, y_1), are now also paired with a utility value, v_1, and any learning algorithm that maximizes accuracy will be weighted for fairness considerations. For Deontologists, loss functions can be added as constraints on the goal of maximizing accuracy. This follows the important discovery of Thornton et al. [19] that Consequentialist and Deontological principles fall into natural categories within machine learning. It also allows us to see how all fairness principles aim to maximize accuracy within fairness boundaries.

4 DEONTOLOGICAL APPROACHES

Given that it is impossible to achieve equality across all metrics for the entire population, Egalitarians may attempt to segment the population based on either intended recipients or victimized groups. In the context of binary classifier evaluations, an intent-based approach could assume that the classification of a positive trait, $\hat{y} = 1$, or negative trait, $y \hat{=} 0$, is the *goal* of the model, and thus only measurements (α) that are conditional on classifications

[1] Rather than directly dividing goods proportionally to effort, *Luck Egalitarians* will indirectly prioritize effort by averaging the effects of luck. The Desert and Luck-Egalitarian principles will look identical for our purposes, so this subtlety will be ignored.

are those for which the designers are morally responsible, which have the form:

$$p(\alpha|\hat{y})$$

These include demographic parity and positive/negative predictive parity. Inequalities in other metrics are what a Kantian would call "foreseen" harms, or in this case, foreseen inequalities, for which the designers should be held morally responsible. In this respect, Northpointe adopted an intent-based position when they emphasized the equality of COMPAS in both demographic and positive predictive parity. This argument assumes that classifications of crime are the intended goals of risk-assessment, and all other results are collateral damage. These arguments are also common in debates about predictive policing, where police claim that they use statistical measures to "go where the crime is," and any unfair error rates are merely foreseen side-effects.

On the other hand, a compensation-based approach to distributive justice focuses on which group is made worse-off by the classifier, in terms of their prior states. As Nozick emphasized, this changes the question of distributive justice into a question of corrective justice. Saleiro et al. [18] adopt this position in their "bias and fairness audit toolkit" called Aequitas, which distinguishes between *assistive* interventions and *punitive* ones. Assistive interventions are those that confer benefits which the recipients would not have otherwise enjoyed, while punitive interventions impose costs or punishments where none would have previously existed. Aequitas then proposes that punitive classifiers should be evaluated by equality in FP rates (giving punishments to those who don't deserve them), while assistive classifiers should be evaluated by equality in FN rates (failing to give rewards to those who do deserve them).

To specify which parity rates are important, Aequitas needs to designate one of the classification values to be the undesirable result, θ. For a risk-assessment algorithm, where $\hat{y} = 1$ means dangerous, then $\theta = 1$. For a loan-eligibility algorithm, where $\hat{y} = 1$ means eligible, then $\theta = 0$. The rates that we care about then have the form:

$$p[(\hat{y} = \theta)|(y = (1 - \theta))]$$

This formula correctly produces FP rates for risk-assessment algorithms and FN rates for loan eligibility algorithms. Another way of describing the compensation view is making use of a "Non-Maleficence" principle that imposes egalitarianism across harms but not benefits. As Saleiro et al. state: "…[providing] assistance to individuals who are false positives will not hurt them, but missing individuals could be harmful to them."

There are two primary objections to compensation approaches. First, it allows for arbitrarily large inequalities for beneficial outcomes; a risk-assessment classifier may be judged acceptable when it contains a large disparity between black and white dangerous prisoners who are mistakenly set free. Second, it assumes a dubious distinction between "better-off" and "worse-off," which may break down when considering all the effects of a classifier. As Consequentialists will note, setting dangerous prisoners free may not make them worse-off, but it certainly makes the public worse-off.

Rather than enforcing Egalitarianism across a select group, a Libertarian principle bases allocations entirely on the resources that each person contributed towards production. One way of translating this into the context of binary classifiers is to think of the prevalence of a trait in some population as that group's "investment" in the classification. If the prior distribution of trait y in Group x is $(y_x = 1)$, then the contribution of Group x is:

$$p(y_x = 1|x)$$

In our sample data, the expectation of reoffense for white prisoners is 48.3% (290/600), while the expectation of reoffense for black prisoners is 70% (140/200). This is one way of thinking about the value of R_x from the normative Libertarian principle. If R_a for the population of white prisoners is 48%, and R_b for the population of black prisoners is 70%, then each group is entitled to success rates that are *at least* as fair as their initial contributions. For instance, if the gap between black and white prisoners on FP or FN rates exceeds the size of this original gap in the target trait, then the Libertarian may call that classifier unfair. However, inequality *within* that range is not unfair, since it is proportional to the original inequality of the trait in the populations. Even though that original inequality may be the result of bad luck or historical oppression, Libertarians do not view our classifier as responsible for mitigating these factors.

In classic Libertarianism, *any* variance from the proportions of original contributions is judged to be unfair. If Alice contributes 3 dollars to production, and Bob only contributes 2, then a 50-50 split of the goods would be just as wrong as a 70-30 split, since the ideal distribution is 60-40. However, in the case of classifiers, the ideal is to achieve 100% success for both groups in predictive measures and 0% for both groups in error measures. Thus, from a fairness perspective, we only care about ensuring that inequality between groups does not *exceed* the pre-existing inequalities in the target trait. Formally, this can be achieved by partitioning rates that exceed this prior ratio from those that don't. Let $d^* \in d_i$ be the subset of rates that exceed the original ratio of contributions between groups:

$$d^* \in d_i : \frac{d^*_x}{\sum d^*_i} > \frac{R_x}{\sum R_i}$$

The Libertarian "cost" that we are seeking to minimize becomes the difference between these unfair rates and the original contribution ratio. Once we minimize this difference, then the model can be optimized for accuracy, and any remaining inequalities between groups are "unfortunate, but not unfair." The goal of the Libertarian is not to create more equality, but simply *not to enlarge* the pre-existing inequalities of our society.

Desert-based approaches reject this focus on the prior prevalence of a trait within the population, since this can be the result of unjust circumstances (either historical oppression or luck). In binary classifiers, advocates of Desert often employ a modified version of demographic parity called *conditional demographic parity*, where the representation of protected groups in the predicted categories must be equal, conditional on certain factors deemed legitimate [9]. If $W(X)$ are the set of traits in the data which are the result of work, then the classifiers that Desert-theorists care about are:

$$p(\hat{y} = (1, 0)|W(X))$$

For a risk-assessment algorithm, the traits in $W(X)$ might be those based on legitimate factors like prior convictions, rather than the

education level of one's parents, even if the latter turns out to be an extremely accurate predictor of the target variable.

The challenge for Desert theorists is to specify exactly what the legitimate restrictions in $W(X)$ ought to be. The broadest restriction may simply be those people who are actually well-qualified for the classification, where $W(X) = (y = 1)$, which is what Hardt et al. [7] call *equality of opportunity*:

$$p(\hat{y} = 1|y = 1)$$

Hardt motivates this idea in a blog post where he demonstrates the problem with demographic parity for targeted marketing:

> Consider, for example, a luxury hotel chain that renders a promotion to a subset of wealthy whites (who are likely to visit the hotel) and a subset of less affluent blacks (who are unlikely to visit the hotel). The situation is obviously quite icky, but demographic parity is completely fine with it so long as the same fraction of people in each group see the promotion.

Rather than merely ensuring that the promotion is offered to both white and black consumers in equal numbers, Hardt emphasizes that what matters is offering it to the white and black consumers *who are likely to visit the hotel* in equal numbers. For our sample risk-assessment algorithm, when considering *the prisoners who are actually dangerous*, the model predicts high-risk at a rate of 93% for white prisoners and 64% for black prisoners, and thus fails to satisfy equality of opportunity.

There are other possible ways for Desert-theorists to restrict the data for conditional parity. For example, we could look at those prisoners who had a certain number of priors, and demand that the predictions of white prisoners with 3 priors should be the same as predictions for black prisoners with 3 priors. Dwork et al.'s [5] "fairness through awareness" model will measure individual distances rather than group outcomes. Dwork explicitly cites the work of Luck Egalitarians like Roemer as philosophical inspiration. Under her model, each individual is assigned some distance in a metric space that evaluates desert, and the way to evaluate the fairness of models is by the average distance between individuals from each group within that metric space. For instance, a white prisoner may have been born to a more privileged background through pure luck, so even if both a white and black prisoner have the same number of prior convictions (say, three priors), by averaging or nullifying the effects of their background privileges, we may rate the "effort" of the white prisoner less than that of the black prisoner, and consequently, predict different levels of risk. There is a case to be made that focusing on individual effort will also be a better predictor of future behaviors, but this is more of a Consequentialist argument. The Desert theorist is purely concerned with reward and punishment for past behaviors.

The central objection to Desert principles is that the factors labeled as "luck" and "work" cannot be properly distinguished. For instance, those who are hard-working have this character trait due to a collection of factors in their history (parenting, role-models, education, genetics, etc.) that are not themselves earned through work.

5 CONSEQUENTIALIST APPROACHES

Consequentialists will listen to the debates about various fairness metrics with some degree of skepticism, since social justice is not dependent on equality or inequality between groups, but rather, on the *effects* of inequality on the happiness of individuals in those groups. Thus, a metric like demographic parity is itself uninteresting to a Utilitarian, since mere inequality of representation in the categories of $\hat{y} = 1$ or $\hat{y} = 0$ does not tell us anything about the relationship between classifications and total utility. Instead, if we could produce estimates of the relative weights of each outcome based on overall social cost, then we could simply design the model to optimize social cost rather than equality.

If α is the average rate of each category: TP_x, FP_x, TN_x, FN_x, and $(w_\alpha)(u_\alpha)$ are the average weighted utilities of each category, then Utilitarians will simply select the model which maximizes the sum of utilities:

$$\sum(w_\alpha)(u_\alpha)(\alpha)$$

While Prioritarians will maximize the minimum:

$$min(w_\alpha)(u_\alpha)(\alpha)$$

The utilities for each category will differ depending on the type of classifier. For loan eligibility algorithms, the negative utility produced by FPs is much worse than the negative utility produced by FNs, since the former represent a complete loss of investment. However, for risk-assessment algorithms in criminal justice, the negative utility of FNs may be much worse than the negative utility of FPs, simply on the grounds of the amount of violence prevented. The suffering caused by keeping an innocent person incarcerated is likely less than the suffering caused by allowing an extremely dangerous prisoner to go free. Deontologists will reject this entire project of calculating how many innocent people in jail we are willing to accept for clean streets, but if there is no distinction being made between the total costs for both prisoners and the public, then employing a common currency of "hedonic values" is a necessary requirement.

Let's start by assigning utilities (u_a) to each category of outcomes in our sample classifier, including both the utilities of prisoners and the public. There are notorious challenges in "interpersonal measurements of utility," but for now these values will merely be stipulated. Say that releasing dangerous prisoners will produce an average of 400 units of harm to the public in the form of violent crime, and that detaining any prisoners will cause 100 units of harm to them and their families, with an additional 100 units of suffering for peaceful prisoners unnecessarily incarcerated. Outcomes involving peaceful prisoners (FP and FN) have no impacts on public utility. A map of these utilities is represented in Table 1:

Table 1: Sample Utilities for Risk-Assessment Outcomes

Outcome	Prisoner Utility	Public Utility	Total Utility
TP	-100	400	300
FP	-200	0	-200
FN	100	-400	-300
TN	100	0	100

Fairness in Consequentialist approaches is found in how we assign weights, (w_a), to each group outcome based on relative social cost. For instance, there may be a greater social cost associated with keeping peaceful black prisoners in jail than peaceful white prisoners, since this perpetuates a historical cycle of deprivation in the black community. Similarly, there may be a greater social benefit to releasing peaceful black prisoners. In our sample data, the proportion of black reoffenders to their group is 1.4 times larger than the proportion of white prisoners to their group (70%:48%). In that case, we might weigh the negative social cost of continued false positives for black prisoners to be 1.4 times worse, and the positive social benefit of true negatives for black prisoners to be 1.4 times better. This weighting feature is the primary tool for achieving greater parity in Consequentialist approaches. Using these values, Figure 4 represents the utilities for our sample classifier.

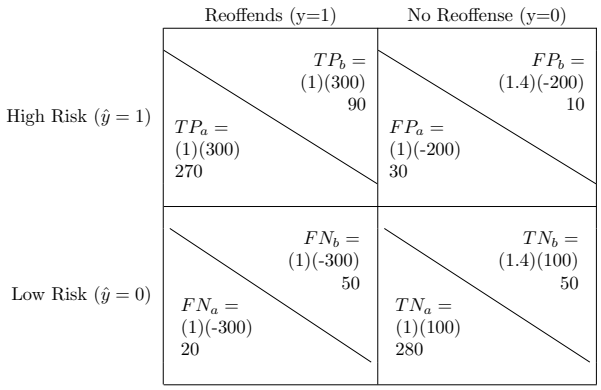

Figure 3: Contingency Table with Weighted Utilities.

Consequentialism has the benefit of incorporating fairness directly into the data, and optimizing for overall benefit. Because of the extra benefit for TNs and extra cost for FPs within black prisoners, optimizing for utility will also produce increased parity between groups. Consistent with the concerns of Corbett-Davies et al. [4], this parity will be balanced against the social cost of sacrificing excessive amounts of public safety.

There are many objections to Consequentialism. We have already considered the problem of interpersonal comparisons of utility. Another objection maintains that, even if it we could assign specific utilities to each category, it may still be unrealistic to do so systematically for all protected groups throughout the data. Consequentialists have sometimes responded to these objections by advocating a form of *Rule Utilitarianism*, which employs heuristics that have proven effective at maximizing happiness within a small-scale environment. A Consequentialist might propose something similar with fairness metrics. Hu and Chen [8] have recently demonstrated the formal possibility of translating each of the ML fairness metrics into a corresponding utility calculation. If one could simulate the effects of different models on total utility, and show that fairness metric M is instantiated in the class of models that ranks highest, then this would be a strong Consequentialist argument for the use of M as a default metric.

6 CONCLUDING THOUGHTS

Companies that employ ML algorithms understandably wish to avoid negative headlines accusing them of discrimination. The problem is that *any* ML algorithm applied to a trait that is unequally distributed across protected groups can be accused of discrimination in some sense. Assuming the company initially designs its model to produce equality in success rates, then tomorrow's headline might read:

ALGORITHM MAKES MORE MISTAKES FOR *B* THAN *A*

If the company adjusts its model to produce equality in error rates, then the headline could read:

ALGORITHM APPROVES MORE DESERVING MEMBERS OF *B* THAN DESERVING MEMBERS OF *A*

When the company tries to produce equality of opportunity, the next headline may read:

ALGORITHM PROVIDES MORE LUCKY BREAKS FOR *B* THAN *A*

Faced with these hard choices, companies may respond by remaining deliberately vague about the details of their models, or deferring to a future industry standard, or just entirely abandoning a commitment to fairness. Yet I would encourage the groups developing ML algorithms not to become discouraged. Instead, engineers and computer scientists should realize that designing a model involves making difficult moral commitments. Attempting to ignore these choices, or just "average across them," will simply produce irresponsible results.

The answer to: "is this model fair to group x?" will always be: "fair according to which normative principle?" An emphasis on the *intended* outputs of a model will care more about equality in success rates, while an emphasis on those who are *disadvantaged* by a model will care more about equality in error rates. Concerns about proportionality will lead one to care more about matching the pre-existing prevalence of a trait within the groups, or the amount of that prevalence for which individuals can be held responsible. Measuring the overall effects on everyone affected by the model (not just the rights of a few) will lead to incorporating fairness metrics within a general calculation of social costs and benefits. If we select one approach, then others will suffer. But this is the nature of moral choices, and the only responsible way to mitigate negative headlines is to develop a consistent response to them, rather than ignore them.

ACKNOWLEDGMENTS

The author would like to express a great thanks to Alexandra Chouldechova for her thoughtful comments and suggestions during all stages of writing.

REFERENCES

[1] Matthew Adler. 2011. *Well-Being and Fair Distribution: beyond cost-benefit analysis.* Oxford University Press, New York, NY.
[2] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 8 (2018), 1–11.
[3] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2017), 153–163.

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision-Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Halifax, NS, Canada, 797–806.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proc. 3rd ITCS*. 214–26.

[6] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Certifying and Removing Disparate Impact. In *Proc. 21st SIGKDD*. ACM.

[7] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proc. 29th NIPS*. 3315–23.

[8] Lily Hu and Yiling Chen. 2018. Welfare and Distributional Impacts of Fair Classification. (2018). arXiv preprint arXiv:1807.01134.

[9] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35 (2013), 613–644.

[10] Immanuel Kant. 1788. *Critique of Practical Reason*. Oxford University Press.

[11] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proc. 8th ITCS*. ITCS.

[12] John Stuart Mill. 1861. *Utilitarianism*. Hackett, New York, NY.

[13] Herve Moulin. 2003. *Fair Division and Collective Welfare*. MIT Press, Cambridge, MA.

[14] Robert Nozick. 1974. *Anarchy, State, and Utopia*. Basic Books, New York, NY.

[15] Ayn Rand. 1961. *The Virtue of Selfishness*. Penguin Press, New York, NY.

[16] John Roemer. 1971. *Theories of Distributive Justice*. Harvard University Press, Cambridge, MA.

[17] John Roemer. 1998. *Equality of Opportunity*. Harvard University Press, Cambridge, MA.

[18] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. (2018). arXiv preprint arXiv:1811.05577.

[19] Sarah Thornton, Selina Pan, Stephen Erlien, and Christian Gerdes. 2016. Incorportating Ethical Considerations into Automated Vehicle Control. In *IEEE Transactions on Intelligent Transportation Systems*. 1–11.