An Algorithmic Framework for Fairness Elicitation

Christopher Jung
University of Pennsylvania, Philadelphia, PA, USA

Michael Kearns
University of Pennsylvania, Philadelphia, PA, USA

Seth Neel
Harvard University, Cambridge, MA, USA

Aaron Roth
University of Pennsylvania, Philadelphia, PA, USA

Logan Stapleton
University of Minnesota, Minneapolis, MN, USA

Zhiwei Steven Wu
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

We consider settings in which the right notion of fairness is not captured by simple mathematical definitions (such as equality of error rates across groups), but might be more complex and nuanced and thus require *elicitation* from individual or collective stakeholders. We introduce a framework in which pairs of individuals can be identified as requiring (approximately) equal treatment under a learned model, or requiring ordered treatment such as "applicant Alice should be at least as likely to receive a loan as applicant Bob". We provide a provably convergent and *oracle efficient* algorithm for learning the most accurate model subject to the elicited fairness constraints, and prove generalization bounds for both accuracy and fairness. This algorithm can also combine the elicited constraints with traditional statistical fairness notions, thus "correcting" or modifying the latter by the former. We report preliminary findings of a behavioral study of our framework using human-subject fairness constraints elicited on the COMPAS criminal recidivism dataset.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Fairness, Fairness Elicitation

Digital Object Identifier 10.4230/LIPIcs.FORC.2021.2

Related Version *Full Version*: https://arxiv.org/abs/1905.10660

1 Introduction

The literature on algorithmic fairness has consisted largely of researchers proposing and showing how to impose technical definitions of fairness [18, 32, 66, 3, 4, 34, 60, 48, 29, 26, 4, 14]. Because these imposed notions of fairness are described analytically, they are typically simplistic, and often have the form of equalizing simple error statistics across groups. Our starting point is the observation that:

1. This process cannot result in notions of fairness that do not have any simple, analytic description, and
2. This process also overlooks a more precursory problem: namely, *who gets to define what is fair?*

It's unlikely that researchers alone are best fit for defining algorithmic fairness. Recent work identifies undue power imbalances [40] and biases [39] that arise when algorithm designers and researchers are the only voices in conversations around ethical design. [59] find that many machine learning practitioners are disconnected from the "organisational

and institutional realities, constraints and needs" specific to the contexts in which their algorithms are applied. Researchers may not be able to propose a concise technical definition, e.g. statistical parity, to capture the nuances of fairness in any given context. Furthermore, many philosophers hold that *stakeholders* who are affected by moral decisions and *experts* who understand the context in which moral decisions are made will have the best judgment about which decisions are fair in that context [63, 39].

To this end, we aim to allow stakeholders and experts to play a central role in the process of defining algorithmic fairness. This is aligned with recent works on *virtual democracy*, which propose and enact participatory methods to automate moral decision-making [13, 51, 31, 46, 21].

The way we involve stakeholders is motivated by two concerns:

1. We want stakeholders to have free rein over how they may define fairness, e.g. we don't want to simply have them vote on whether existing, simple constraints like statistical parity or equalized odds is best; and
2. We want non-technical stakeholders to be able to contribute, even if they may not understand the inner workings of a learning algorithm.

We hold that people often cannot elucidate their conceptions of fairness; yet, they can identify specific scenarios where fairness or unfairness occurs.[1] Drawing from individual notions of fairness like [17, 29] that are defined in terms of pairwise comparisons, we therefore aim to elicit stakeholders conceptions of fairness by asking them to compare pairs of individuals in specific scenarios. Specifically, we ask whether it's fair that one particular individual should receive an outcome that is as desirable or better than the other.

When pointing out fairness or unfairness, this kind of pairwise ranking is natural. For example, after Serena Williams was penalized for a verbal interaction with an umpire in the 2018 U.S. Open Finals, tennis player James Blake tweeted, "I have said worse and not gotten penalized. And I've also been given a "soft warning" by the ump where they tell you knock it off or I will have to give you a violation. [The umpire] should have at least given [Williams] that courtesy" [65]. Here, Blake thinks that: 1) Williams should have been judged as or less severely than he would have been in a similar situation; and 2) the umpire's decision was unfair, because Williams was judged more severely.

Thus, we ask a set of stakeholders about a fixed set of pairs of individuals subject to a classification problem. For each pair of individuals (A, B), we ask the stakeholder to choose from amongst a set of four options:

1. Fair outcomes must classify A and B the *same* way (i.e. they must either both get a favorable classification or both get an unfavorable classification).
2. Fair outcomes must give A an outcome that is equal to *or preferable to* the outcome of B.
3. Fair outcomes must give B an outcome that is equal to *or preferable to* the outcome of A
4. Fair outcomes may treat A and B differently without any constraints.

These constraints, a data distribution, and a hypothesis class define a learning problem: minimize classification error subject to the constraint that the rate of violation of the elicited pairwise constraints is held below some fixed threshold. Crucially and intentionally we elicit relative pairwise orderings of outcomes (e.g. A and B should be treated equally), but do not elicit preferences for absolute outcomes (e.g. A should receive a positive outcome). This is

[1] This is philosophically akin to a theory of moral epistemology called *moral perception*, which claims that we know moral facts (e.g. goodness or fairness) via perception, as opposed to knowing them via rules of morality (see [10]).

because *fairness* – in contrast to *justice* – is often conceptualized as a measure of equality of outcomes, rather than correctness of outcomes[2]. In particular, it remains the job of the learning algorithm to optimize for correctness subject to elicited fairness constraints.

We remark that the premise (and the foundation for the enormous success) of machine learning is that accurate decision making rules in complex scenarios cannot be defined with simple analytic rules, and instead are best derived directly from data. Our work can be viewed similarly, as deriving fairness constraints from data elicited from experts and stakeholders. In this paper, we solve the computational, statistical, and conceptual issues necessary to do this, and demonstrate the effectiveness of our approach via a small behavioral study.

1.1 Results

Our Model

We model individuals as having features in $\mathcal{X}$ and binary labels, drawn from some distribution $\mathcal{P}$. A committee of *stakeholders*[3] $u \in \mathcal{U}$ has preferences about whether one individual should be judged better than another individual. We imagine presenting each stakeholder with a set of pairs of individuals and asking them to choose one of four options for each pair, e.g. given the features of Serena Williams and Jacob Blake:

1. No constraint;
2. Williams should be treated as well as Blake or better;
3. Blake should be treated as well as Williams or better; or
4. Williams and Blake should be treated similarly.

Here, when we refer to how an individual *should be treated*, we mean the probability that an individual is given a positive label by the classifier. This may be a bit of a relaxation of these judgments, since they are not about actualized classifications, but rather the *probabilities* of positive classification. For example, we may not consider it a violation of fairness preference (2) if Williams is judged worse than Blake in a specific scenario; yet, if an ump is *more likely* to judge Williams worse than Blake in general, then this would violate this fairness preference.

We represent these preferences abstractly as a set of ordered pairs $C_u \subseteq \mathcal{X} \times \mathcal{X}$ for each stakeholder u. If $(x, x') \in C_u$, this means that stakeholder u believes that individual x' must be treated as well as individual x or better, i.e. ideally the classifier h classifies such that $h(x') \geq h(x)$. This captures all possible responses above. For example, for Serena Williams (s) and Jacob Blake (b), if stakeholder u responds:

1. *No constraint* $\Leftrightarrow$ $(s, b) \notin C_u$ nor $(b, s) \notin C_u$;
2. *Williams as well as Blake* $\Leftrightarrow$ $(b, s) \in C_u$;
3. *Blake as well as Williams* $\Leftrightarrow$ $(s, b) \in C_u$; or
4. *Treated similarly* $\Leftrightarrow$ $(s, b) \in C_u$ and $(b, s) \in C_u$ (since if $h(b) \geq h(s)$ and $h(s) \geq h(b)$, then $h(s) = h(b)$).

[2] Sidney Morgenbesser, following the Columbia University campus protests in the 1960s, reportedly said that the police had treated him unjustly, but not unfairly. He said that he was treated unjustly because the police hit him without provocation – but not unfairly, because the police were doing the same to everyone else as well.

[3] Though we develop our formalism as a committee of stakeholders, note that it permits the special case of a single subjective stakeholder, which we make use of in our behavioral study.

We impose no structure on how stakeholders form their views nor on the relationship between the views of different stakeholders – i.e. the sets $\{C_u\}_{u \in \mathcal{U}}$ are allowed to be arbitrary (for example, they need not satisfy a triangle inequality), and need not be mutually consistent. We write $C = \cup_u C_u$.

We then formulate an optimization problem constrained by these pairwise fairness constraints. Since it is intractable to require that all constraints in C be satisfied exactly, we formulate two different "knobs" with which we can quantitatively relax our fairness constraints.

For $\gamma > 0$ (our first knob), we say that the classification of an ordered pair of individuals $(x, x') \in C$ satisfies γ-fairness if the probability of positive classification for x' plus γ is no smaller than the probability of positive classification for x, i.e. $\mathbb{E}[h(x')] + \gamma \geq \mathbb{E}[h(x)]$. In this expression, the expectation is taken only over the randomness of the classifier h. Equivalently, a γ-fairness violation corresponds to the classification of an ordered pair of individuals $(x, x') \in C$ if the difference between these probabilities of positive classification is greater than γ, i.e. $\mathbb{E}[h(x) - h(x')] > \gamma$. Thus, γ acts as a buffer on how likely it is that x' be classified worse than x before a fairness violation occurs. For example, if Blake (b) receives a good label (i.e. no penalty) 80% of the time and Williams (s) 50% of the time, then for $\gamma = 0.1$ this constitutes a γ-fairness violation for the ordered pair $(b, s) \in C$, since $\mathbb{E}[h(b) - h(s)] = 0.3 \geq 0.1 = \gamma$.

We might ask that for *no* pair of individuals do we have a γ-fairness violation:

$$\max_{(x,x') \in C} \mathbb{E}[h(x) - h(x')] \leq \gamma.$$

On the other hand, we could ask for the weaker constraint that *over a random draw of a pair of individuals*, the expected fairness violation is at most η (our second knob): $\mathbb{E}_{(x,x') \sim \mathcal{P}^2}[(h(x) - h(x')) \cdot \mathbf{1}[(x, x') \in C]] \leq \eta$. We can also combine both relaxations to ask that the in expectation over random pairs, the "excess" fairness violation, on top of an allowed budget of γ, is at most η. For example, as above, if Blake receives a good label 80% of the time and Williams 50%, for $\gamma = 0.1$, the umpire classifier would pick up 0.2 excess fairness violation for $(b, s) \in C$. In Section 2, we weight these excess fairness violations by the proportion of stakeholders who agree with the corresponding fairness constraint and mandate their sum be less than η. Subject to these constraints, we would like to find the distribution over classifiers that minimizes classification error: given a setting of the parameters γ and η, this defines a benchmark with which we would like to compete.

Our Theoretical Results

Even absent fairness constraints, learning to minimize 0/1 loss (even over linear classifiers) is computationally hard in the worst case (see e.g. [20, 19]). Despite this, learning seems to be empirically tractable in the real world. To capture the *additional* hardness of learning subject to fairness constraints, we follow several recent papers [2, 33] in aiming to develop *oracle efficient* learning algorithms. Oracle efficient algorithms are assumed to have access to an *oracle* (realized in experiments using a heuristic – see the next section) that can solve weighted classification problems. Given access to such an oracle, oracle efficient algorithms must run in polynomial time. We show that our fairness constrained learning problem is computationally no harder than unconstrained learning by giving such an oracle efficient algorithm (or reduction), and show moreover that its guarantees generalize from in-sample to out-of-sample in the usual way – with respect to both accuracy and the frequency and magnitude of fairness violations. Our algorithm is simple and amenable to implementation, and we use it in our experimental results.

Our Experimental Results

We implement our algorithm and run a set of experiments on the COMPAS recidivism prediction dataset, using fairness constraints elicited from 43 human subjects. We establish that our algorithm converges quickly (even when implemented with fast learning heuristics, rather than "oracles"). We also explore the Pareto curves trading off error and fairness violations for different human subjects, and find empirically that there is a great deal of variability across subjects in terms of their conception of fairness, and in terms of the degree to which their expressed preferences are in conflict with accurate prediction. We find that most of the difficulty in balancing accuracy with the elicited fairness constraints can be attributed to a small fraction of the constraints.

1.2 Related Work

Individual Fairness and Elicitation

Our work is related to existing notions of *individual fairness* like [17, 29] that conceptualize fairness as a set of constraints binding on pairs of individuals. In particular, the notion of metric fairness proposed in [17] is closely related, but distinct from the fairness notions we elicit in this work. In particular: 1) We allow for constraints that require that individual A be treated better than or equal to individual B, whereas metric fairness constraints are symmetric, and only allow constraints of the form that A and B be treated similarly. In this sense our notion is more general; 2) We elicit binary judgments between pairs of individuals, whereas metric fairness is defined as a Lipschitz constraint on a real valued metric. In this sense our notion is more restrictive. Though, we – along with a line of work on classification with pairwise constraints – see merit in pairwise constraints because they "can be relatively easy to collect from human feedback" [49, p. 114].

The most technically related piece of work is Rothblum and Yona [53], who first frame individual fairness in a PAC learning setting and prove similar generalization guarantees to ours for a relaxation of metric fairness. Our conceptual focus is quite different, however: for general learning problems, they prove worst-case hardness results, whereas we derive algorithms in the oracle-efficient model and evaluate them on real elicited user data. The concurrent work of [41] makes a similar observation about guaranteeing fairness with respect to an unknown metric, although their aim is the orthogonal goal of fair representation learning.

Dwork et al. [17] first proposed the notion of individual metric-fairness that we take inspiration from, imagining fairness as a Lipschitz constraint on a randomized algorithm, with respect to some "task-specific metric". Since the original proposal, the question of where the metric should come from has been one of the primary obstacles to its adoption, and the focus of subsequent work. Zemel et al. [67] attempt to automatically learn a representation for the data (and hence, implicitly, a similarity metric) that causes a classifier to label an equal proportion of two protected groups as positive. Kim et al. [35] consider a group-fairness like relaxation of individual metric-fairness, asking that on average, individuals in pre-specified groups are classified with probabilities proportional to the average distance between individuals in those groups. They show how to learn such classifiers given access to an oracle which can evaluate the distance between two individuals according to the metric. Compared to our work, they assume the existence of a fairness metric which can be accessed using a quantitative oracle, and they use this metric to define a statistical rather than individual notion of fairness. Gillen et al. [23] assume access to an oracle which simply identifies fairness violations across pairs of individuals. Under the assumption that the oracle

is exactly consistent with a metric in a simple linear class, they give a polynomial time algorithm to compete with the best fair policy in an online linear contextual bandits problem. In contrast to [23], we make essentially no assumptions at all on the structure of the "fairness" constraints. Bechavod et al. [8] generalize the setting of Gillen et al. [23] by making no assumption on the underlying metric and reduce the number of calls to the fairness oracle. Ilvento [28] studies the problem of metric learning with the goal of using only a small number of numeric valued queries, which are hard for human beings to answer, relying more on comparison queries. In contrast with [28], we do not attempt to learn a metric, and instead directly learn a classifier consistent with the elicited pairwise fairness constraints.

Classification with Pairwise Constraints

Stretching back to at least Kleinberg and Tardos [37], there is a line of work that considers classification problems with pairwise constraints which define similarity or dissimilarity between the labels of two data points [49, 7, 49, 68, 6, 57]. Kleinberg and Tardos [37], for example, introduce a classification problem with pairwise equality constraints and a distance metric between pairs. The constraints we elicit differ in that they are asymmetric inequality relations between pairs, rather than equality or metric constraints. Much of this work is conceptually different from ours, as it is concerned with clustering and semi-supervised learning, e.g. [7] or [6].

Preference Elicitation, Social Choice Theory, and Virtual Democracy

Preference elicitation is a well-established area in machine learning [52, 16]. Social choice or voting theory aims to elicit and aggregate people's preferences in order to find a winner or ranking over alternatives (see [50] for an overview). The aim of this work is often to suggest aggregation methods which meet some desirable criteria, such as strategyproofness. *Virtual democracy* is a framework for eliciting normative preferences, constructing a model from these preferences – typically via voting methods – in order to automate moral decision making based on these norms [1, 51, 13, 46, 31, 5, 21]. Our work is conceptually similar: we elicit and learn from moral preferences. However, our work differs in two regards: 1) We focus specifically on fair classification, whereas virtual democracy is concerned with more general moral decisions, e.g. the kind of trolley problem scenarios in the *Moral Machine* experiment [5]; 2) As such, we can incorporate elicited preferences into a fair learning problem rather than using voting methods to aggregate them.

Human Perspectives on Algorithmic Fairness

A number of recent qualitative and HCI works have focused on understanding perspectives on algorithms and fairness from the public [25, 24, 56, 61, 62, 55] and from practitioners [27, 47, 59]. These works suggest that 1) algorithms should incorporate stakeholder input into the fairness principles they use and 2) these fairness principles are context-specific [11, 15, 43, 44, 9, 54]. These works also offer a perspective on how explaining or asking people about fairness influences how they respond [45, 64, 54, 9]. These works provide largely qualitative findings, which may be difficult to translate into specific design implications. Our work complements these prior works by offering a way to easily implement a fair algorithm based on human perspectives.

2 Problem Formulation

Let S denote a set of labeled examples $\{z_i = (x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ is a feature vector and $y_i \in \mathcal{Y}$ is a label. We will also write $S_X = \{x_i\}_{i=1}^n$ and $S_Y = \{y_i\}_{i=1}^n$. Throughout the paper, we will restrict attention to binary labels, so let $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{P}$ denote the unknown distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ denote a hypothesis class containing binary classifiers $h : \mathcal{X} \to \mathcal{Y}$. We assume that $\mathcal{H}$ contains a constant classifier (which will imply that the "fairness constrained" ERM problem that we define is always feasible). We'll denote classification error of hypothesis h by $err(h, \mathcal{P}) := \Pr_{(x,y)\sim\mathcal{P}}(h(x) \neq y)$ and its empirical classification error by $err(h, S) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(h(x_i) \neq y_i)$.

We assume there is a set of one or more stakeholders $\mathcal{U}$, such that each stakeholder $u \in \mathcal{U}$ is identified with a set of ordered pairs (x, x') of individuals $C_u \subseteq \mathcal{X}^2$: for each $(x, x') \in C_u$, stakeholder u thinks that x' should be treated as well as x or better, i.e. ideally that for the learned classifier h, the classification $h(x') \geq h(x)$ (we will ask that this hold in expectation if the classifier is randomized, and will relax it in various ways). For each ordered pair (x, x'), let $w_{x,x'}$ be the fraction of stakeholders who would like individual x to be treated as well as x': that is, $w_{x,x'} = \frac{|\{u|(x,x')\in C_u\}|}{|\mathcal{U}|}$. Note that if $(x, x') \in C_u$ and $(x', x) \in C_u$, then the stakeholder wants x and x' to be treated similarly in that ideally $h(x) = h(x')$.

In practice, we will not have direct access to the sets of ordered pairs C_u corresponding to the stakeholders u, but we may ask them whether particular ordered pairs are in this set (see Section 5 for details about how we actually query human subjects). We model this by imagining that we present each stakeholder with a random set of pairs $A \subseteq [n]^2$, and for each ordered pair (x_i, x_j), ask if x_j should not be treated worse than x_i; we learn the set of ordered pairs in $A \cap C_u$ for each u. Define the empirical constraint set $\hat{C}_u = \{(x_i, x_j) \in C_u\}_{\forall(i,j)\in A}$ and $\hat{w}_{x_i x_j} = \frac{|\{u|(x,x')\in \hat{C}_u\}|}{|\mathcal{U}|}$, if $(i, j) \in A$ and 0 otherwise. We write that $\hat{C} = \cup_u \hat{C}_u$. For brevity, we will sometimes write w_{ij} instead of w_{x_i,x_j}. Note that $\hat{w}_{ij} = w_{ij}$ for every $(i, j) \in A$.

Our goal will be to find the distribution over classifiers from $\mathcal{H}$ that minimizes classification error, while satisfying the stakeholders' fairness preferences, captured by the constraints C. To do so, we'll try to find D, a probability distribution over $\mathcal{H}$, that minimizes the training error and satisfies the stakeholders' empirical fairness constraints, $\hat{C}$. For convenience, we denote the expected classification error of D as $err(D, \mathcal{P}) := \mathbb{E}_{h\sim D}[err(h, \mathcal{P})]$ and likewise its expected empirical classification error as $err(D, S) := \mathbb{E}_{h\sim D}[err(h, S)]$. We say that any distribution D over classifiers satisfies (γ, η)-approximate subjective fairness if it is a feasible solution to the following constrained empirical risk minimization problem:

$$\min_{D\in\Delta\mathcal{H},\alpha_{ij}\geq 0} err(D, S) \tag{1}$$

$$\text{such that } \forall(i, j) \in [n]^2 : \quad \mathop{\mathbb{E}}_{h\sim D}[h(x_i) - h(x_j)] \leq \alpha_{ij} + \gamma \tag{2}$$

$$\sum_{(i,j)\in[n]^2} \frac{\hat{w}_{ij}\alpha_{ij}}{|A|} \leq \eta. \tag{3}$$

This "Fair ERM" problem, whose feasible region we denote by $\Omega(S, \hat{w}, \gamma, \eta)$, has decision variables D and $\{\alpha_{ij}\}$, representing the distribution over classifiers and the "fairness violation" terms for each pair of training points, respectively. The parameters γ and η are constants which represent the two different "knobs" we have at our disposal to quantitatively relax the fairness constraint, in an ℓ_∞ and ℓ_1 sense, respectively.

The parameter γ defines, for any ordered pair (x_i, x_j), the maximum difference between the probabilities that x_i and x_j receive positive labels without constituting a fairness violation. The parameter α_{ij} captures the "excess fairness violation" beyond γ for (x_i, x_j).

The parameter η upper bounds the sum of these allotted excess fairness violation terms α_{ij}, each weighted by the proportion of judges who perceive they ought to be treated similarly $\hat{w}_{ij}$ and normalized with the total number of pairs presented $|A|$. Thus, η bounds the expected degree of dissatisfaction of the panel of stakeholders $\mathcal{U}$, over the random choice of an ordered pair $(x_i, x_j) \in A$ and the randomness of their classification. We iterate over all $(i,j) \in [n]^2$ (not just those in $\hat{C}$) because $\hat{w}_{ij} = 0$ if no judge prefers x_i should be classified as well as x_j.

To better understand γ and η, we consider them in isolation. First, suppose we set $\gamma = 0$. Then, *any* difference in probabilities of positive classification between pairs is deemed a fairness violation. So, if we choose $(D, \{\alpha_{ij}\})$ such that the sum of weighted differences in positive classification probabilities exceeds η, i.e.

$$\sum_{(i,j)\in[n]^2} \frac{\hat{w}_{ij}\, \mathbb{E}_{h\sim D}[h(x_i) - h(x_j)]}{|A|} > \eta,$$

then this is an infeasible solution. Second, suppose that $\eta = 0$. Then, for any $(x_i, x_j) \in C$ (for which $\hat{w}_{ij} > 0$), if the expected difference in labels exceeds γ, i.e. $\mathbb{E}_{h\sim D}[h(x_i) - h(x_j)] > \gamma$, then this is an infeasible solution.

2.1 Fairness Loss

Our goal is to develop an algorithm that will minimize its empirical error $err(D, S)$, while satisfying the empirical fairness constraints $\hat{C}$. The standard VC dimension argument states that empirical classification error will concentrate around the true classification error: we hope to show the same kind of generalization for fairness as well. To do so, we first define fairness loss with respect to our elicited fairness preferences here.

For some fixed randomized hypothesis $D \in \Delta\mathcal{H}$ and w, define γ-fairness loss between an ordered pair as $\Pi_{D,w,\gamma}((x,x')) = w_{x,x'} \max\left(0, \mathbb{E}_{h\sim D}\left[h(x) - h(x')\right] - \gamma\right)$. For a set of pairs $M \subset \mathcal{X} \times \mathcal{X}$, the γ-fairness loss of M is defined to be: $\Pi_{D,w,\gamma}(M) = \frac{1}{|M|}\sum_{(x,x')\in M} \Pi_{D,w,\gamma}((x,x'))$. This is the expected degree to which the difference in classification probability for a randomly selected pair exceeds the allowable budget γ, weighted by the fraction of stakeholders who think that x' should be treated as well as x. By construction, the empirical fairness loss is bounded by η (i.e. $\Pi_{D,w,\gamma}(M) \leq \sum_{ij} \frac{\hat{w}_{ij}\alpha_{ij}}{|A|} \leq \eta$), and we show in Section 4, the empirical fairness should concentrate around the true fairness loss $\Pi_{D,w,\gamma}(\mathcal{P}) := \mathbb{E}_{x,x'\sim\mathcal{P}^2}\left[\Pi_{D,w,\gamma}(x,x')\right]$.

2.2 Cost-sensitive Classification

In our algorithm, we will make use of a cost-sensitive classification (CSC) oracle. An instance of CSC problem can be described by a set of costs $\{(x_i, c_i^0, c_i^1)\}_{i=1}^n$ and a hypothesis class, $\mathcal{H}$. Costs c_i^0 and c_i^1 correspond to the cost of labeling x_i as 0 and 1 respectively. Invoking a CSC oracle on $\{(x_i, c_i^0, c_i^1)\}_{i=1}^n$ returns a hypothesis h^* such that $h^* \in \text{argmin}_{h\in\mathcal{H}} \sum_{i=1}^n \left(h(x_i)c_i^1 + (1 - h(x_i))\, c_i^0\right)$. We say that an algorithm is *oracle-efficient* if it runs in polynomial time assuming access to a CSC oracle.

3 Empirical Risk Minimization

In this section, we give an oracle-efficient algorithm whose pseudocode is shown in Algorithm 1 for approximately solving our (in-sample) constrained empirical risk minimization problem. Details are deferred to the full version of this paper which is available on arXiv [30]. We prove the following theorem:

Algorithm 1 No-Regret Dynamics.

Input: training examples $\{x_i, y_i\}_{i=1}^n$, bounds C_λ and C_τ, time horizon T, step sizes μ_λ and $\{\mu_\tau^t\}_T^{t=1}$,
Set $\theta_1^0 = \mathbf{0} \in \mathbb{R}^{n^2}$
Set $\tau^0 = 0$
for $t = 1, 2, \ldots, T$ **do**
  Set $\lambda_{ij}^t = C_\lambda \frac{\exp \theta_{ij}^{t-1}}{1+\sum_{i',j' \in [n]^2} \exp \theta_{i'j'}^{t-1}}$ for all pairs $(i,j) \in [n]^2$
  Set $\tau^t = proj_{[0,C_\tau]}\left(\tau^{t-1} + \mu_\tau^t \left(\frac{1}{|A|}\sum_{i,j} w_{ij}\alpha_{ij}^{t-1} - \eta\right)\right)$
  $D^t, \alpha^t \leftarrow \text{BEST}_\rho(\lambda^t, \tau^t)$
  for $(i,j) \in [n]^2$ **do**
    $\theta_{ij}^t = \theta_{ij}^{t-1} + \mu_\lambda^{t-1}\left(\mathbb{E}_{h\sim D^t}\left[h(x_i) - h(x_j)\right] - \alpha_{ij}^t - \gamma\right)$

Output: $\frac{1}{T}\sum_{t=1}^T D^t, \frac{1}{T}\sum_{t=1}^T \alpha^t$

▶ **Theorem 1.** *Fix parameters ν, C_τ, C_λ that serve to trade off running time with approximation error. There is an efficient algorithm that makes $T = \left(\frac{2C_\lambda\sqrt{\log(n)}+C_\tau}{\nu}\right)^2$ CSC oracle calls and outputs a solution $(\hat{D}, \hat{\alpha})$ with the following guarantee. The objective value is approximately optimal:*

$$err(\hat{D}, S) \leq \min_{(D,\alpha)\in\Omega(S,\hat{w},\gamma,\eta)} err(D, S) + 2\nu.$$

And the constraints are approximately satisfied: $\mathbb{E}_{h\sim\hat{D}}[h(x_i) - h(x_j)] \leq \hat{\alpha}_{ij} + \gamma + \frac{1+2\nu}{C_\lambda}, \forall (i,j) \in [n]^2$ and $\frac{1}{|A|}\sum_{(i,j)\in[n]^2} \hat{w}_{ij}\hat{\alpha}_{ij} \leq \eta + \frac{1+2\nu}{C_\tau}$.

3.1 Outline of the Solution

We frame the problem of solving our constrained ERM problem (equations (1) through (3)) as finding an approximate equilibrium of a zero-sum game between a primal player and a dual player, trying to minimize and maximize respectively the Lagrangian of the constrained optimization problem.

The Lagrangian for our optimization problem is

$$\mathcal{L}(D, \alpha, \lambda, \tau) = err(D, S) + \sum_{(i,j)\in[n]^2} \lambda_{ij}\left(\mathop{\mathbb{E}}_{h\sim D}[h(x_i) - h(x_j)] - \alpha_{ij} - \gamma\right) + \tau\left(\frac{1}{|A|}\sum_{(i,j)\in[n]^2} w_{ij}\alpha_{ij} - \eta\right)$$

For the constraint in equation (2), corresponding to the γ-fairness violation for each ordered pair of individuals (x_i, x_j), we introduce a dual variable λ_{ij}. For the constraint (3), which corresponds to the η-fairness violation over all pairs of individuals, we introduce a dual variable of τ. For brevity, we define vectors $\lambda \in \Lambda$ and α which are made up of all the multipliers λ_{ij} and the excess fairness violation allotments α_{ij}, respectively. The primary player's action space is $(D, \alpha) \in (\Delta\mathcal{H}, [0,1]^{n^2})$, and the dual player's action space is $(\lambda, \tau) \in (\mathbb{R}^{n^2}, \mathbb{R})$.

Solving our constrained ERM problem is equivalent to finding a minmax equilibrium of $\mathcal{L}$:

$$\operatorname*{argmin}_{(D,\alpha)\in\Omega(S,\hat{w},\gamma,\eta)} err(D,S) = \operatorname*{argmin}_{D\in\Delta\mathcal{H},\alpha\in[0,1]^{n^2}} \max_{\lambda\in\mathbb{R}^{n^2},\tau\in\mathbb{R}} \mathcal{L}(D,\alpha,\lambda,\tau)$$

Because $\mathcal{L}$ is linear in terms of its parameters, Sion's minimax theorem [58] gives us

$$\min_{D\in\Delta\mathcal{H},\alpha\in[0,1]^{n^2}} \max_{\lambda\in\mathbb{R}^{n^2},\tau\in\mathbb{R}} \mathcal{L}(D,\alpha,\lambda,\tau) = \max_{\lambda\in\mathbb{R}^{n^2},\tau\in\mathbb{R}} \min_{D\in\Delta\mathcal{H},\alpha\in[0,1]^{n^2}} \mathcal{L}(D,\alpha,\lambda,\tau).$$

By a classic result of Freund and Schapire [22], one can compute an approximate equilibrium by simulating "no-regret" dynamics between the primal and dual player. "No-regret" meaning that the average *regret* –or difference between our algorithm's plays and the single best play in hindsight– is bounded above by a term that converges to zero with increasing rounds.

In our case, we define a zero-sum game wherein the primary player's plays from action space $(D,\alpha)\in(\Delta\mathcal{H},[0,1]^{n^2})$, and the dual player's plays from action space $(\lambda,\tau)\in(\mathbb{R}^{n^2}_{\geq 0},\mathbb{R}_{\geq 0})$. In any given round t, the dual player plays first and the primal second. The primal player can simply best respond to the dual player (see Algorithm 2).

However, since the dual player plays first, they cannot simply best respond to the primal player's action. The dual player has to anticipate the primal player's best response in order to figure out what to play. Ideally, the dual player would enumerate every possible primal play and calculate the best dual response. However, this is intractable. So, the dual player updates dual variables $\{\lambda,\tau\}$ according to *no-regret* learning algorithms (exponentiated gradient descent [36] and online gradient descent [69], respectively).

The time-averaged play of both players converges to an approximate equilibrium of the zero-sum game, where the approximation is controlled by the regret of the dual player. This approximate equilibrium corresponds to an approximate saddle point for the Lagrangian $\mathcal{L}$, which is equivalent to an approximate solution to the Fair ERM problem.

We organize the rest of this section as follows. First, for simplicity, we show how the primal player updates $\{D,\alpha\}$ (even though the dual player plays first). Second, we show how the dual player updates $\{\lambda,\tau\}$. Finally, we prove that these updates are no-regret and relate the regret of the dual player to the approximation of the solution to the Fair ERM problem.

3.2 The Primal Player's Best Response

In each round t, given the actions chosen by the dual player (λ^t,τ^t), the primal player needs to best respond by choosing (D^t,α^t) such that $(D^t,\alpha^t)\in\operatorname{argmin}_{D\in\Delta\mathcal{H},\alpha\in[0,1]^{n^2}}\mathcal{L}(D,\alpha,\lambda^t,\tau^t)$. We can separate the optimization problem into two as shown above in Algorithm 2: one optimization over hypothesis D and one over violation factor α. As for D^t, the primal player can update the hypothesis D by leveraging a CSC oracle. Given λ^t, we can set the costs as follows:

$$c_i^0 = \frac{1}{n}\mathbb{E}_{h\sim D}\left[\mathbb{1}(y_i\neq 0)\right], \qquad c_i^1 = \frac{1}{n}\mathbb{E}_{h\sim D}\left[\mathbb{1}\left(y_i\neq 1\right)\right] + (\lambda_{ij}^t - \lambda_{ji}^t).$$

Then, $D^t = h^t = CSC\left(\{(x_i,c_i^0,c_i^1)\}_{i=1}^n\right)$ (we note that the best response is always a deterministic classifier h^t).

As for α^t, we can show that the primal player sets $\alpha_{ij}^t = 1$ if $\tau^t\frac{w_{ij}}{|A|} - \lambda_{ij}^t \leq 0$ and 0 otherwise. We defer its derivation and proofs to the full version of this paper which is available on arXiv [30].

Algorithm 2 Best Response, $BEST_\rho(\lambda, \tau)$, for the primal player.

Input: training examples $S = \{x_i, y_i\}_{i=1}^n$, $\lambda \in \Lambda$, $\tau \in \mathcal{T}$, CSC oracle CSC
for $i = 1, \ldots, n$ **do**
  if $y_i = 0$ **then**
    Set $c_i^0 = 0$
    Set $c_i^1 = \frac{1}{n} + \sum_{j \neq i} \lambda_{ij} - \lambda_{ji}$
  else
    Set $c_i^0 = \frac{1}{n}$
    Set $c_i^1 = \sum_{j \neq i} \lambda_{ij} - \lambda_{ji}$
$D = CSC(S, c)$
for $(i, j) \in [n]^2$ **do**

$$\alpha_{ij} = \begin{cases} 1: & \tau \frac{w_{ij}}{|A|} - \lambda_{ij} \leq 0 \\ 0: & \tau \frac{w_{ij}}{|A|} - \lambda_{ij} > 0. \end{cases}$$

Output: D, α

3.3 The Dual Player's No-regret Updates

In order to reason about convergence, we need to restrict the dual player's action space to lie within a bounded ℓ_1 ball, defined by the parameters C_τ and C_λ that appear in our theorem – and serve to trade off running time with approximation quality: $\Lambda = \left\{\lambda \in \mathbb{R}_+^{n^2} : \|\lambda\|_1 \leq C_\lambda\right\}, \mathcal{T} = \{\tau \in \mathbb{R}_+ : \|\tau\|_1 \leq C_\tau\}$. The dual player will use exponentiated gradient descent [36] to update λ and online gradient descent [69] to update τ, where the reward function will be defined as: $r_\lambda(\lambda^t) = \sum_{(i,j) \in [n]^2} \lambda_{ij}^t \left(\mathbb{E}_{h \sim D}\left[h(x_i) - h(x_j)\right] - \alpha_{ij} - \gamma\right)$ and $r_\lambda(\tau^t) = \tau^t \left(\frac{1}{|A|} \sum_{(i,j) \in [n]^2} w_{ij} \alpha_{ij} - \eta\right)$. We defer its derivation and proofs to the full version of this paper which is available on arXiv [30].

4 Generalization

In this section, we show that fairness loss generalizes out-of-sample. (Error generalization follows from the standard VC-dimension bound, which – because it is a uniform convergece statement is unaffected by the addition of fairness constraints. See the supplement for the standard statement.)

Proving that the fairness loss generalizes doesn't follow immediately from a standard VC-dimension argument for several reasons: it is not linearly separable, but defined as an average over non-disjoint *pairs* of individuals in the sample. The difference between empirical fairness loss and true fairness loss of a randomized hypothesis $D \in \Delta\mathcal{H}$ is also a non-convex function of the supporting hypotheses h, and so it is not sufficient to prove a uniform convergence bound merely for the base hypotheses in our hypothesis class $\mathcal{H}$. We circumvent these difficulties by making use of an ε-net argument, together with an application of a concentration inequality, and an application of Sauer's lemma. Briefly, we show that with respect to fairness loss, the continuous set of distributions over classifiers have an ε-net of sparse distributions. Using the two-sample trick and Sauer's lemma, we can bound the number of such sparse distributions. The end result is the following generalization theorem:

▶ **Theorem 2.** *Let S consists of n i.i.d points drawn from $\mathcal{P}$ and let M represent a set of m pairs randomly drawn from $S \times S$. Then we have:*

$$\Pr_{\substack{S\sim\mathcal{P}^n \\ M\sim(S\times S)^m}}\left(\sup_{D\in\Delta\mathcal{H}}\left|\Pi_{D,w,\gamma}(M) - \mathop{\mathbb{E}}_{(x,x')\sim\mathcal{P}^2}\left[\Pi_{D,w,\gamma}(x,x')\right]\right| > 2\varepsilon\right)$$
$$\leq \left(8\cdot\left(\frac{e\cdot 2n}{d}\right)^{dk}\exp\left(\frac{-n\varepsilon^2}{32}\right) + \left(\frac{e\cdot 2n}{d}\right)^{dk'}\exp\left(-8m\varepsilon^2\right)\right),$$

where $k' = \frac{2\ln(2m)}{\varepsilon^2} + 1$, $k = \frac{\ln(2n^2)}{8\varepsilon^2} + 1$, and d is the VC-dimension of $\mathcal{H}$.

To interpret this theorem, note that the right hand side (the probability of a failure of generalization) begins decreasing exponentially fast in the data and fairness constraint sample parameters n and m as soon as $n \geq \Omega(d\log(n)\log(n/d))$ and $m \geq \Omega(d\log(m)\log(n/d))$.

5 A Behavioral Study

The framework and algorithm we have provided can be viewed as a tool to elicit and enforce a notion of fairness defined by a collection of stakeholders. In this section, we describe preliminary results from a human-subject study we performed in which pairwise fairness preferences were elicited and enforced by our algorithm. We note that the subjects included in our empirical study were not stakeholders affected by the algorithm we used (the COMPAS algorithm). Thus, our results should not be interpreted as cogent for any policy modifications to the COMPAS algorithm. We instead report our empirical findings primarily to showcase the performance of our algorithm and to act as a template for what should be reported if our framework were applied with relevant stakeholders (for example, if fairness preferences about COMPAS data were elicited from inmates).[4]

5.1 Data

Our study used the COMPAS recidivism data gathered by ProPublica [5] in their celebrated analysis of Northepointe's risk assessment algorithm [42]. This data consists of defendants from Broward County in Florida between 2013 to 2014. For each defendant the data consists of sex (male, female), age (18-96), race (African-American, Caucasian, Hispanic, Asian, Native American), juvenile felony count, juvenile misdemeanor count, number of other juvenile offenses, number of prior adult criminal offenses, the severity of the crime for which they were incarcerated (felony or misdemeanor), as well as the outcome of whether or not they did in fact recidivate. Recidivism is defined as a new arrest within 2 years, not counting traffic violations and municipal ordinance violations.

5.2 Subjective Fairness Elicitation

We implemented our fairness framework via a web app that elicited subjective fairness notions from 43 undergraduates at a major research university. After reading a document describing the data and recidivism prediction task, each subject was presented with 50 randomly chosen

[4] We omit such an empirical study due to the difficulty of accessing such stakeholders and leave this for future work.

[5] The data can be accessed on ProPublica's Github page here. We cleaned the data as in the ProPublica study, removing any records with missing data. This left 5829 records, where the base rate of two-year recidivism was 46%.

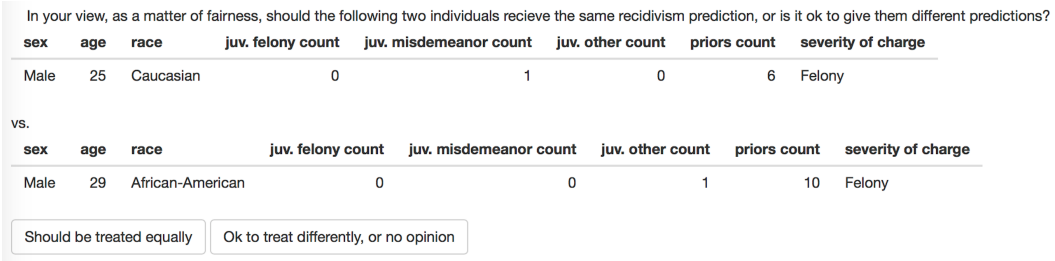

Figure 1 Screenshot of sample subjective fairness elicitation question posed to human subjects.

pairs of records from the COMPAS data set and asked whether in their opinion the two individuals should treated (predicted) equally or not. Importantly, the subjects were shown only the features for the individuals, and not their actual recidivism outcomes, since we sought to elicit subjects' fairness notions regarding the predictions of those outcomes. While absolutely no guidance was given to subjects regarding fairness, the elicitation framework allows for rich possibilities. For example, subjects could choose to ignore demographic factors or criminal histories entirely if they liked, or a subject who believes that minorities are more vulnerable to overpolicing could discount their criminal histories relative to Caucasians in their pairwise elicitations.

For each subject, the pairs they identified to be treated equally were taken as constraints on error minimization with respect to the actual recidivism outcomes over the entire COMPAS dataset, and our algorithm was applied to solve this constrained optimization problem, using a linear threshold heuristic as the underlying learning oracle [33]. We ran our algorithm with $\eta = 0$ and variable γ in Equations (1) through (3), which represents the strongest enforcement of subjective fairness – the difference in predicted values must be at most γ on *every* pair selected by a subject. Because the issues we are most interested in here (convergence, tradeoffs with accuracy, and heterogeneity of fairness preferences) are orthogonal to generalization – and because we prove VC-dimension based generalization theorems – for simplicity, the results we report are in-sample.

5.3 Results

Since our algorithm relies on a learning heuristic for which worst-case guarantees are not possible, the first empirical question is whether the algorithm converges rapidly on the behavioral data. We found that it did so consistently; a typical example is Figure 2a, where we show the trajectories of model error vs. fairness violation for a particular subject's data for variable values of the input γ (horizontal lines). After 1000 iterations, the algorithm has converged to the optimal errors subject to the allowed γ.

Perhaps the most basic behavioral questions we might ask involve the extent and nature of subject variability. For example, do some subjects identify constraint pairs that are much harder to satisfy than other subjects? And if so, what factors seem to account for such variation?

Figure 2b shows that there is indeed considerable variation in subject difficulty. For each of the 43 subjects, we have plotted the error vs. fairness violation Pareto curves obtained by varying γ from 0 (pairs selected by subjects must have identical probabilistic predictions of recidivism) to 1.0 (no fairness enforced whatsoever). Since our model space is closed under probabilistic mixtures, the worst-case Pareto curve is linear, obtained by all mixtures of the error-optimal model and random predictions. Easier constaint sets are more convex. We see

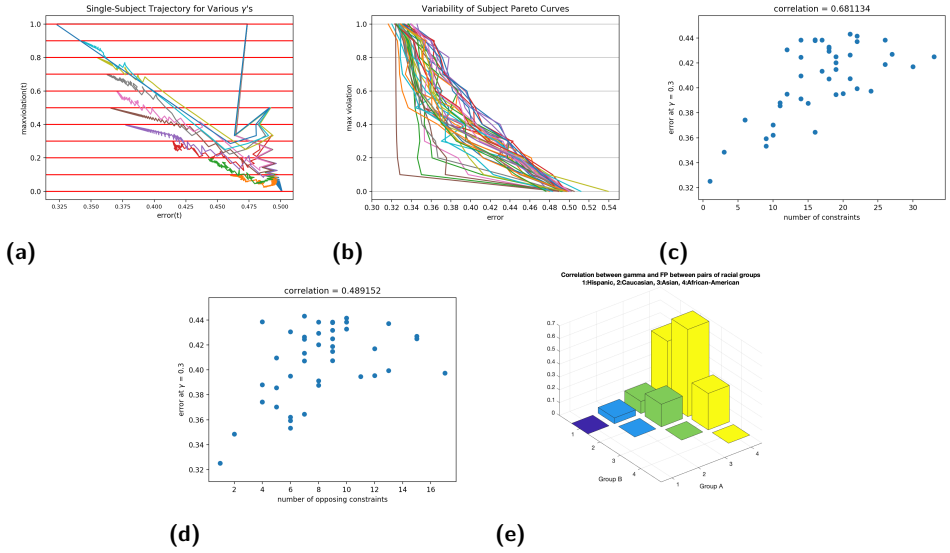

Figure 2 (a) Sample algorithm trajectory for a particular subject at various γ. (b) Sample subjective fairness Pareto curves for a sample of subjects. (c) Scatterplot of number of constraints specified and number of opposing constraints vs. error at $\gamma = 0.3$. (d) Scatterplot of number of constraints where the true labels are different vs. error at $\gamma = 0.3$. (e) Correlation between false positive rate difference and γ for racial groups.

in the figure that both extremes are exhibited behaviorally – some subjects yield linear or near-linear curves, while others permit huge reductions in unfairness for only slight increases in error, and virtually all the possibilities in between are realized as well. [6]

Since each subject was presented with 50 random pairs and was free to constrain as many or as few as they wished, it is natural to wonder if the variation in difficulty is explained simply by the number of constraints chosen. In Figure 2c we show a scatterplot of the the number of constraints selected by a subject (x axis) versus the error obtained (y axis) for $\gamma = 0.3$ (an intermediate value that exhibits considerable variation in subject error rates) for all 43 subjects. While we see there is indeed strong correlation (approximately 0.69), it is far from the case that the number of constraints explains all the variability. For example, amongst subjects who selected approximately 16 constraints, the resulting error varies over a range of nearly 8%, which is over 40% of the range from the optimal error (0.32) to the worst fairness-constrained error (0.5). More surprisingly, when we consider only the "opposing" constraints, pairs of points with different true labels, the correlation (0.489) seems to be weaker. Enforcing a classifier to predict similarly on a pair of points with different true labels should increase the error, and yet, it is less correlated with error than the raw number of constraints. This suggests that the variability in subject difficulty is due to the nature of the constraints themselves rather than their number or disagreement with the true labels.

It is also interesting to consider the collective force of the 1432 constraints selected by all 43 subjects together, which we can view as a "fairness panel" of sorts. Given that there are already individual subjects whose constraints yield the worst-case Pareto curve, it is unsurprising that the collective constraints do as well. But we can exploit the flexibility of our optimization framework in Equations (1) through constraint (3), and let $\gamma = 0.0$ and vary only η, thus giving the learner discretion in which subjects' constraints to discount or discard at a given budget η. In doing so we find that the unconstrained optimal error

[6] The slight deviations from true convexity are due to approximate rather than exact convergence.

can be obtained while having the average (exact) pairwise constraint be violated by only roughly 25%, meaning roughly that only 25% of the collective constraints account for all the difficulty.

Finally, we can investigate the extent to which behavioral subjective fairness notions align with more standard statistical fairness definitions, such as equality of false positive rates. For instance, for each subject and a pair of racial groups, we take the absolute difference in false positive rates of the classifier at $\gamma \in \{0.0, 0.1, \ldots, 1.0\}$ and calculate the correlation coefficient between realized values of γ (which measure violation of subjective unfairness) and the false positive rate differences. Figure 2e shows the average correlation coefficient across subjects for each pair of racial groups. We note that subjective fairness correlates with a smaller gap between the false positive rates across Caucasians and African Americans: but correlates substantially less for other pairs of racial groups.

We leave a more complete investigation of our behavioral study for future work, including the detailed nature of subject variability and further comparison of behavioral subjective fairness to standard algorithmic fairness notions.

6 Discussion and Limitations

We provide a framework to involve non-technical stakeholders into the process of defining algorithmic fairness. Our approach offers a means for stakeholders to encode their more nuanced, contextual principles of fairness into a model. By eliciting pairwise fairness preferences, our approach is designed to be easily understood by laypeople, even if they do not understand the technicalities of the learning algorithm. We provide theoretical guarantees, as well as preliminary experiments to demonstrate the functionality of our algorithm.

Here, we anticipate a criticism: biases may be embedded into the model if the stakeholders we elicit from are biased. To address this, we clarify that our approach aims to easily elicit and operationalize what stakeholders think is fair and unbiased. As such, if stakeholders truthfully report their preferences, then our approach should produce a model which these stakeholders believe is unbiased. In this sense, we consider what is biased and what is fair in our context to be subjective. While this point of view may be uncomfortable within the algorithmic fairness community, it is a well-established argument within ethics and normative economics [38]. Furthermore, if the design goal is for the model to achieve some objective standard of neutrality, then we believe this should be addressed at the level of choosing which stakeholders to elicit from and how they are elicited – see Section 1.2 for related qualitative and HCI works which consider how to elicit fairness preferences from people. We consider this to be an important direction for future work.[7]

References

1 Matthew Adler. Aggregating moral preferences. *Economics and Philosophy*, 32:283–321, 2016.

2 Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

3 Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 60–69, 2018. URL: `http://proceedings.mlr.press/v80/agarwal18a.html`.

[7] See, for example, a follow-up interview study which elicits pairwise fairness beliefs about risk assessment tools used in the Allegheny County child welfare system [12].

4 Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 120–129, 2019. URL: http://proceedings.mlr.press/v97/agarwal19d.html.

5 Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563:59–64, 2018.

6 Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *ICML*, 2018.

7 Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining, (SDM-2004), pp. , Lake Buena Vista, FL*, pages 333—-344, 2004.

8 Yahav Bechavod, Christopher Jung, and Steven Z. Wu. Metric-free individual fairness in online learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/80b618ebcac7aa97a6dac2ba65cb7e36-Abstract.html.

9 Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 377. ACM, 2018.

10 Lawrence Blum. *Moral Perception and Particularity*. Cambridge University Press, 1994.

11 Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 41, 2019. doi:10.1145/3290605.3300271.

12 Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2021. arXiv:2102.01196.

13 Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for artificial intelligence. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2018.

14 Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. arXiv:1808.00023.

15 Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 275–285, New York, NY, USA, 2019. ACM. doi:10.1145/3301275.3302310.

16 Carmel Domshlak, Eyke Hüllermeier, Souhila Kaci, and Henri Prade. Preferences in ai: An overview. *Artificial Intelligence*, 175(7):1037—-1052, 2011. doi:10.1016/j.artint.2011.03.004.

17 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.

18 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012. doi:10.1145/2090236.2090255.

19 Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

20 Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

21 Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

22 Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *COLT*, volume 96, pages 325–332. Citeseer, 1996.

23 Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

24 Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 903–912. International World Wide Web Conferences Steering Committee, 2018. doi:10.1145/3178876.3186138.

25 Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

26 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

27 Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 600, 2019. doi:10.1145/3290605.3300830.

28 C Ilvento. Metric learning for individual fairness. Manuscript submitted for publication, 2019.

29 Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

30 Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. An algorithmic framework for fairness elicitation. *arXiv preprint*, 2019. arXiv:1905.10660.

31 Anson Kahng, Min Kyung Lee, Ritesh Noothigattu, Ariel D. Procaccia, and Christos-Alexandros Psomas. Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3173–3182, 2019.

32 Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

33 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

34 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2569–2577, 2018. URL: http://proceedings.mlr.press/v80/kearns18a.html.

35 Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852, 2018.

36 Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63, 1997.

37 Jon Kleinberg and Éva Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and markov random fields. In *P40th Annual Symposium on Foundations of Computer Science, New York, NY, USA*, pages 14—-23, 1999. doi:10.1109/SFFCS.1999.814572.

38 James Konow. Is fairness in the eye of the beholder? an impartial spectator analysis of justice. *Social Choice and Welfare*, 33:101—-127, 2009.

39 Felicitas Kraemer, Kees van Overveld, and Martin Peterson. Is there an ethics of algorithms? *Ethics and Information Technology*, 13:251—-260, 2011.

40 Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. Participatory approaches to machine learning. International Conference on Machine Learning Workshop, 2020.

41 Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *CoRR*, abs/1907.01439, 2019. arXiv:1907.01439.

42 Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm, March 2019. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

43 Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018.

44 Min Kyung Lee and Su Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1035–1048. ACM, 2017.

45 Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3365–3376. ACM, 2017.

46 Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):181:1–181:35, 2019.

47 John Monahan, Anne Metz, and Brandon L Garrett. Judicial appraisals of risk assessment in sentencing. *Virginia Public Law and Legal Theory Research Paper*, No. 2018-27, 2018.

48 Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

49 Nam Nguyen and Rich Caruana. Improving classification with pairwise constraints: A margin-based approach. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 113–124, 2008.

50 Noam Nisan. Introduction to mechanism design (for computer scientists). In *N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, Algorithmic Game Theory*. Cambridge University Press, 2007.

51 Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd Conference on Artificial Intelligence, (AAAI)*, pages 1587–1594, 2018.

52 Gabriella Pigozzi, Alexis Tsoukiàs, and Paolo Viappiani. Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 77:361–401, 2016.

53 Guy N Rothblum and Gal Yona. Probably approximately metric-fair learning. *arXiv preprint*, 2018. arXiv:1803.03242.

54 Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, and Michael Carl Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Vienna, Austria, July 12–18, 2020*, 2020.

55 Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare?: Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99—-106. ACM, 2019.

56 Nicholas Scurich and John Monahan. Evidence-based sentencing: Public openness and opposition to using gender, age, and race as risk factors for recidivism. *Law and Human Behavior*, 40(1):36, 2016.

57 Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. In *Neural Computation*, pages 1–35, 2021. doi:10.1162/neco_a_01373.

58 Maurice Sion et al. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

59 Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

60 Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

61 AJ Wang. Procedural justice and risk-assessment algorithms, 2018. doi:10.2139/ssrn.3170136.

62 Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

63 Pak-Hang Wong. Democratizing algorithmic fairness. *Philosophy & Technology*, 33:225–244, 2020.

64 Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 656. ACM, 2018.

65 Ariana Yaptangco. Male tennis pros confirm serena's penalty was sexist and admit to saying worse on the court. *Elle*, 2017. URL: http://www.elle.com/culture/a23051870/male-tennis-pros-confirm-serenas-penalty-was-sexist-and-admit-to-saying-worse-on-the-court/.

66 Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW*, pages 1171–1180. ACM, 2017.

67 Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

68 H. Zeng and Y. Cheung. Semi-supervised maximum margin clustering with pairwise constraints. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):926–939, 2012. doi:10.1109/TKDE.2011.68.

69 Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.

A Missing Details And Proofs

We defer all the missing details and proofs to the full version of this paper which is available on arXiv [30].