An Overview of My Work in Welfare-Centric Fair Machine Learning

Cyrus Cousins

December 2023

1 Introduction

I see my work as a rigorous interdisciplinary challenge to the machine learning status quo. I find myself equal parts *enthralled* by the accomplishments of modern machine learning and *disturbed* by the harm caused by industrial "AI systems." As the real-world impact of machine learning continues to grow, I increasingly see fairness as *the most important topic* in machine learning. I have always been drawn to statistical learning theory to rigorously study learning and quantify overfitting, and this perspective informs my study of the complex sociotechnical issues we face today. Many real-world fairness issues are due to basic errors (e.g., overrepresentation of white males in image datasets [Karkkainen and Joo, 2021] is distribution shift), but as academics we have not made issues of fairness in machine learning easy to understand, and deep questions remain regarding *what fairness even means* and *how fairness interacts with learning*. I see *fairness* and *statistics* as as inextricably linked: fairness *on the training set* means nothing, and *overfitting to fairness* often manifests in discriminatory ways given small minority group samples. Before discussing my research in depth, I briefly state a few key points of my research philosophy and ethics.

While we can not always predict how our work will be applied, we do have a *societal responsibility* to avoid working on problems likely to have harmful impact and to favor problems likely to have positive impact. I strive to motivate my research with problems faced by applied researchers, practitioners, and society at large. We also have an *academic responsibility* to study truly novel and interesting problems, and strive for real progress in the field. It is always worth asking whether we are even solving the right problems. Are we grounding our work in the interests of practitioners and the field at large? Is the community interested in our problem, and *should they be*?

Even with a carefully-considered and well-motivated research problem, we must be cognizant of our limited perspective. Others may pose problems in different but equally valid ways, or may study overlapping problems, and while at times it's worth arguing oven these differences, it is often more helpful to acknowledge and address alternative perspectives and framings. A mathematician at heart, I see great value in making analysis and algorithms general, while making *methodologically necessary* assumptions clear. Ultimately, it is not our place to tell others, especially domain experts, *what they want to compute*; we should rather focus on *how to compute* a desideratum and *what is efficiently computable*.

2 A Theory of Fair Machine Learning

The modern zeitgeist around fair machine learning centers imposing *statistical parity constraints* [Dwork et al., 2012, Verma and Rubin, 2018] (equalized odds, equality of opportunity, outcome, etc.). While such approaches seem intuitively fair, they suffer from computational intractability [Hu and Chen, 2020], mutual incompatibility [Kleinberg et al., 2017], and can actually worsen outcomes for all groups [Hu and Chen, 2020, Jorgensen et al., 2022, 2023].

I also criticize learning with statistical parity fairness concepts on a statistical level: if we enforce fairness constraints *on the training set*, they are *not guaranteed* on the underlying distribution (i.e., we "overfit to fairness"). Moreover, if we "overconstrain" the model during training to account for this, then *for any sample size*, with *arbitrarily high probability*, (1) there may not exist any training-feasible model, and (2) the optimal training-feasible model may be greatly outperformed by a true-feasible model that is not training-feasible. Addressing these issues is quite challenging [Thomas et al., 2019, Yona and Rothblum, 2018], and can require unbounded sample complexity.

Axiomatic Theory The wellbeing of society overall and of disadvantaged or minority groups is well-studied in welfare economics [Dalton, 1920, Debreu, 1959, Gorman, 1968, Pigou, 1912] and moral philosophy [Bentham, 1789, Parfit, 1997, Rawls, 1971], and prescriptive requirements for fair systems are often encoded in law [Kumar et al., 2022, Selbst et al., 2024], so I ask, *"Why did we as computer scientists need to redefine fairness?"* To briefly summarize centuries of thought, *utilitarian welfare* measures overall wellbeing as the *sum* or *average* utility across a population [Bentham, 1789, Mill, 1863], *Rawlsian* or *egalitarian* welfare measures the *minimum* utility [Rawls, 1971, 2001], and *prioritarian* concepts lie somewhere in between [Arneson, 2000, Parfit, 1997]. Utilitarianism is criticized for not incentivizing *equitable redistribution*, and egalitarianism is criticized for ignoring all but the *most disadvantaged* groups in society. In contrast, prioritarianism encompasses various justice criteria that *prioritize* the wellbeing of the impoverished, without ignoring others, making tradeoffs between them in various ways. The *Pigou-Dalton transfer principle* [Dalton, 1920, Pigou, 1912] and the *Debreu-Gorman axioms* [Debreu, 1959, Gorman, 1968] lead all welfare functions to concord with sums of *logarithms* or *powers* of utilities, i.e., for g groups and utility vectors $\boldsymbol{s} \in \mathbb{R}_+^g$, for some $p \in \mathbb{R}$, all fairness concepts $\mathrm{M}(\boldsymbol{s})$ define a *partial order* over utility vectors that agrees with

$$\mathrm{M}(\boldsymbol{s}) = \operatorname{sgn}(p) \sum_{i=1}^{g} \boldsymbol{s}_i^p \ , \quad \text{or} \quad \mathrm{M}(\boldsymbol{s}) = \sum_{i=1}^{g} \ln(\boldsymbol{s}_i) \ . \tag{1}$$

This lays the foundation for my work, but the *mathematical context* of machine learning and estimation raises a few issues.

1) Existing analysis is almost entirely based on welfare and utility, whereas machine learning often considers loss. Simple transformations (e.g., negation) are insufficient, as we require nonnegative $\boldsymbol{s}$ in (1).

2) While directly applicable to *individual level fairness*, the theory does not gracefully handle *heterogeneous group sizes*. This is important to many *discrimination issues* facing minority groups in machine learning, such as differential performance of facial recognition or medical machine learning systems.

3) The scale of welfare functions, and thus the difficulty of approximation or estimation, varies with p. Similarly, depending on p, (1) can be very sensitive to small changes to $\boldsymbol{s}$, which complicates optimization and estimation. Moreover, since (1) is only specified *up to an ordering*, are approximations even meaningful?

While one could introduce *ad hoc* objectives to address these issues, I wanted a "natural" characterization of fair machine learning. I thus sought to show that the assumptions I carried as a computer scientist with learning and estimation in mind could be expressed as *simple axioms* from which a class of fair objectives arises.

The economics and philosophy literatures primarily treat *utility* and *wellbeing*, but in machine learning we often center loss (disutility) instead of utility. I show [Cousins, 2021] that the theory of *suffering-focused ethics* can produce a family of "malfare functions" that quantify *societal suffering* (rather than wellbeing). Given nonnegative utility, we seek to maximize a quasiconcave *welfare function* $\mathrm{W}(\boldsymbol{s})$ with $p \leq 1$ in (1), but given nonnegative disutility, we instead minimize a quasiconvex *malfare function* $\reflectbox{W}(\boldsymbol{s})$ with $p \geq 1$ in (1).

I argue that in machine learning, we usually want models to generalize to *unseen individuals*, thus we should target *group-level* fairness guarantees. However, the classical *symmetry axiom* ($\mathrm{M}(\boldsymbol{s}) = \mathrm{M}(\pi(\boldsymbol{s}))$ $\forall$ permutations π) would require all groups be treated equally, *regardless of size*. I thus introduced *group weightings* $\boldsymbol{w}$ (probability vectors, where $\boldsymbol{w}_i$ is the population frequency of group i), alongside the *weighted symmetry* ($\mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}(\pi(\boldsymbol{s}); \pi(\boldsymbol{w}))$ for all permutations π) and *weighted decomposability* axioms (if $\boldsymbol{w}$ and $\boldsymbol{w}'$ differ only on groups with equal (dis)utility, then $\mathrm{M}(\boldsymbol{s}; \boldsymbol{w}) = \mathrm{M}(\boldsymbol{s}; \boldsymbol{w}')$) to treat variably-sized groups.

To address issues of *sensitivity* of fairness concepts to small (dis)utility changes, and ensure their *units* match those of (dis)utility (as in utilitarian and egalitarian welfare), I introduced the *multiplicative linearity* axiom ($\mathrm{M}(\alpha \boldsymbol{s}; \boldsymbol{w}) = \alpha \mathrm{M}(\boldsymbol{s}; \boldsymbol{w})$) and the *unit scale* axiom ($\mathrm{M}(\mathbf{1}; \boldsymbol{w}) = 1$). These axioms are natural almost to the point of triviality, but they are sufficient to characterize the *cardinal value* of fairness concepts, whereas (1) specifies them *only up to a partial ordering*. These novel axioms, when combined with the Debreu-Gorman axioms, characterize the *weighted power-mean family*, i.e., the class of all fairness concepts takes the form

$$\mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{s}_i^p} \text{ for } p \neq 0 \ , \quad \text{or} \quad \mathrm{M}_0(\boldsymbol{s}; \boldsymbol{w}) = \exp\left(\sum_{i=1}^{g} \boldsymbol{w}_i \ln(\boldsymbol{s}_i) \right) \ . \tag{2}$$

From these axioms alone stems the *monotonicity property*, i.e., for all $p \leq q$, we have

$$\min_{i \in 1,\dots,g} \boldsymbol{s}_i = \mathrm{M}_{-\infty}(\boldsymbol{s}) \leq \mathrm{M}_p(\boldsymbol{s}; \boldsymbol{w}) \leq \mathrm{M}_q(\boldsymbol{s}; \boldsymbol{w}) \leq \mathrm{M}_{\infty}(\boldsymbol{s}) = \max_{i \in 1,\dots,g} \boldsymbol{s}_i \ .$$

Thus power-mean justice criteria are sandwiched between the *egalitarian* minimum ($p = -\infty$) utility or maximum ($p = \infty$) disutility and the *utilitarian* arithmetic mean ($p = 1$) (dis)utility. In some sense, power-means *nonlinearly interpolate* between these extremes, where moving towards egalitarianism *magnifies the impact of inequality*, thus increasing malfare or decreasing welfare. Since *units* in power-means (2) agree with (dis)utility (e.g., $\mathrm{M}_2(\boldsymbol{s}; \boldsymbol{w})$ is measured in *dollars*, not *square dollars*), we can reasonably interpret errors (differences in $\mathrm{M}_p(\cdot; \boldsymbol{w})$) linearly, which is crucial to *approximation* and *estimation*.

Philosophical Implications to Fair Machine Learning During this foray into interdisciplinary literature, I tried to keep my results grounded in contemporary fair machine learning research. Weighted risk minimization is essentially the default approach to machine learning, and the now-standard response to machine learning bias is to train on more balanced data. While a reasonable first step, this perspective is *inherently utilitarian*, and thus *does not* incentivize equitable redistribution of harm. Minimax fair learning [Abernethy et al., 2022, Diana et al., 2021, Shekhar et al., 2021] takes the Rawlsian approach of minimizing the maximum group risk, but is thus susceptible to minority rule, and insensitive to all but the most-disadvantaged groups. By extending welfare theory to develop a novel theory of fair machine learning, I unearthed a *spectrum of objectives* that empower modelers to express and optimize their own fairness concepts. In particular, the *empirical malfare minimization* objective, given loss function ℓ and $\boldsymbol{m}_i$ labeled pairs $(\boldsymbol{x}_{i,\cdot}, \boldsymbol{y}_{i,\cdot})$ for each group i, is

$$\operatorname*{argmin}_{\theta \in \Theta} \mathrm{M}_p \left(i \mapsto \frac{1}{\boldsymbol{m}_i} \sum_{j=1}^{\boldsymbol{m}_i} \ell\big(h_\theta(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}\big); \boldsymbol{w} \right) \quad , \quad \text{where } i \mapsto f(i) \doteq \Big\langle f(1), f(2), \dots, f(g) \Big\rangle \ ,$$

for some $\boldsymbol{w}$-weighted p-power-mean. This generalizes weighted risk minimization ($p = 1$) and minimax fair learning ($p = \infty$), while contextualizing and addressing shortcomings of both approaches for $p \in (1, \infty)$.

This welfare-centric approach centers *fairness* and *societal impact*, which forces modelers to consider not their own goals, but rather the impact of their model on others. In contrast, *fairness constraints* amend existing machine learning objectives, and can be tuned to be of secondary importance (industry incentivizes *profit optimization*, considering *fairness* only insofar as profits or reputation are harmed). I see this radical perspective on machine learning objectives as invaluable in the discussion of regulation, rights, and responsibilities surrounding AI systems.

A Concrete Example Let us now consider a simple example of the tradeoffs we must consider in a social planning problem. Suppose we have to select a single type of drink to serve to three guests from the below options. This illustrates issues of mutual unsatisfiability and statistical estimation that may arise with constraint-based methods, but do understand that it serves as a metaphor for similar tradeoffs that need to be made in machine learning, e.g., which features to consider in a linear model, or what basic patterns to learn to recognize in the lower layers of a neural network.

Option	$\boldsymbol{s}$	$\mathrm{M}_{-\infty}(\boldsymbol{s})$ $\min_i \boldsymbol{s}_i$	$\mathrm{M}_0(\boldsymbol{s})$ $\sqrt[3]{\prod_i \boldsymbol{s}_i}$	$\mathrm{M}_1(\boldsymbol{s})$ $\frac{1}{3}\sum_i \boldsymbol{s}_i$	Parity Gap $\max_i \boldsymbol{s}_i - \min_i \boldsymbol{s}_i$
Water	$\langle 1, 1, 1 \rangle$	**1**	1	1	**0**
Kefir	$\langle 2, 3, 1 \rangle$	**1**	$\boldsymbol{\sqrt[3]{6} \approx 1.82}$	2	2
Kombucha	$\langle 7, 2, 0 \rangle$	0	0	**3**	7

Observe that Water is the "fairest" choice in terms of the *parity gap* (max - min utility), but everyone would weakly prefer Kefir, and Kombucha has the highest total utility. Fairness constraints on the parity gap could force to selection of Water, to no one's benefit, and constraints on the parity gap may be mutually unsatisfiable with constraints on other fairness statistics, such as utility variance or Atkinson's indices. In contrast, optimizing power-means will never prefer dominated outcomes, and optimizing egalitarian welfare $\mathrm{M}_{-\infty}(\cdot)$ captures the essence of the parity gap seeking to prevent harm to the most disadvantaged group without causing such issues.

Moreover, power-means directly encode fairness concepts, and generally preserve convex or concave curvature for optimization purposes, whereas many parity constraints are non-convex, and thus may fracture a convex feasible parameter set into one or more non-convex regions.

As for statistical concerns, the convergence of power-mean estimates is now well-understood [Cousins, 2021, 2022, 2023b], whereas parity-constrained optimization must consider estimated objective and estimated constraints. Estimating $\boldsymbol{s}$ from bounded continuous random variables, for no sample size can we determine whether the parity gap of Kefir is < 2, thus optimizing for, say, utilitarian welfare under a parity gap constraint of < 2 is not possible *for any sample size*.

Applications I first applied this theory during my doctoral studies to develop a concept of fair-PAC (FPAC) learning. I ask, *"given a concept class $\mathcal{H}$ and g per-group probability distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_g$, is it possible to ε (approximately) δ (probably) recover the optimal h^* with finite sample complexity for any power-mean malfare function? Is the sample complexity of fair learning polynomial?"* I answer [Cousins, 2021] affirmatively, and moreover show that, statistically speaking, for *finite-class classification*, PAC and FPAC learnability are equivalent. As for *computational complexity*, I find that standard conditions for efficient convex optimization (e.g., SVM, GLM, etc.) also suffice for malfare objectives. These results stem from Lipschitz-continuity of power-mean malfare functions (i.e., $p \geq 1$). I show [Cousins, 2023b] that power-mean welfare functions are Lipschitz continuous iff $p < 0$, but only Hölder continuous for $p \in (0, 1)$. I then generalize FPAC learning to optimization of arbitrary families of malfare or welfare functions, and show that, for power-mean welfare, sample complexity may depend *exponentially* on $\min(\frac{1}{p}, \min_{i \in 1,\ldots,g} \boldsymbol{w}_i)$.

In practice, data distributions greatly impact learnability, and in fair machine learning, each group can have their own data distribution and sample size. Moreover, in the real world, most machine learning is profit-driven, data are *actively collected* at a cost, and fairness is a tertiary reputational concern. I thus ask [Cousins, 2022], *"How can we optimally allocate sampling effort to efficiently accomplish our goals?"* I show that the *fairness concept*, *model class*, and *data distributions* interact in complicated ways, but *progressive-sampling* algorithms can *actively sample* based on *estimated greedy improvement* to optimize a given fair objective with near-optimal sample complexity.

Feeling as though I've only scratched the surface, I endeavor to explore these ideas in the broader machine learning and algorithmic fairness context. In Cousins et al. [2022], Michael Littman, Kavosh Asadi, and I study *fair reinforcement learning*, where each group i provides *noisy reward feedback* $R_i(s, a)$ to an agent, who optimizes the welfare of per-group value functions (i.e., expected γ-geometrically discounted reward). In our parlance, we seek

$$\operatorname*{argmax}_{\pi} \mathrm{M}_p \left(i \mapsto \mathop{\mathbb{E}}_{\pi, s} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \pi(s_t)) \right] ; \boldsymbol{w} \right) \quad . \tag{3}$$

We give an algorithm that, with high probability, takes polynomially many *exploration actions* before always producing ε-optimal policies. Cardinal welfare theory also sees applications beyond machine learning. During my postdoctoral studies, I collaborated with Yair Zick and Vignesh Viswanathan to analyze such objectives in fair allocation settings. We show sufficient conditions for efficient optimization in restricted classes of submodular valuation functions [Cousins et al., 2023c,d].

Elicitation and working from partial information In Mazzetto et al. [2021], for convex classifiers, Alessio Mazzetto, Dylan Sam, Stephen Bach, Eli Upfal, and I replace *known class labels* with *confidence sets* $\mathcal{S} \subseteq \triangle_c^m$ over *feasible labelings*, and show that we can efficiently adversarially train c-class classifiers. In Dong and Cousins [2022], Evan Dong and I show the same for adversarial training over unknown *group labels*, i.e., $\mathcal{S} \subseteq \triangle_g^m$, with minimax fairness objectives, i.e., for m unlabeled training points $\boldsymbol{x}$, labels $\boldsymbol{y}$, and unknown group IDs $\boldsymbol{z}$, we pose

$$\operatorname*{argmin}_{\theta \in \Theta} \max_{\boldsymbol{y} \in \mathcal{S}} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{c} \boldsymbol{y}_j \cdot \ell(h_\theta(\boldsymbol{x}_i), j) \quad , \quad \text{or} \quad \operatorname*{argmin}_{\theta \in \Theta} \max_{\boldsymbol{z} \in \mathcal{S}} \max_{i \in 1,\ldots,g} \frac{\sum_{j=1}^{m} \boldsymbol{z}_{j,i} \cdot \ell(h_\theta(\boldsymbol{x}_j), \boldsymbol{y}_j)}{\sum_{j=1}^{m} \boldsymbol{z}_{j,i}} \quad . \tag{4}$$

In both works, we apply statistical learning techniques to induce *linear constraints* on class labels $\boldsymbol{y}$ or group ID labels $\boldsymbol{z}$, and then *adversarially optimize* subject to said constraints. This involves estimating label frequencies on a *labeled dataset*, then constructing an uncertainty set $\mathcal{S}$ to respect these statistics (to within

probabilistic error), making these *semisupervised methods*. Crucially, the amount of labeled data to construct $\mathcal{S}$ scales with the *number of statistical constraints*, whereas the *generalization error* of the *trained model* scales with the $\mathcal{L}_1$ radius of the $\mathcal{S}$, the complexity of the model class $\mathcal{H}$, and the amount of *unlabeled data*.

In collaboration with Justin Payan and Yair Zick, I applied similar methods to reviewer assignment [Cousins et al., 2023b], wherein the goal is to match n_1 papers to n_2 reviewers to optimize total assigned affinity. Realistically, reviewers can't bid on all papers, so we rely on other sources of information (e.g., keyword or NLP similarity), which leaves low confidence on the *quality* of the review a match would produce. We thus *adversarially optimize* total affinity over an *uncertainty set* $\mathcal{S} \subseteq \mathbb{R}^{n_1 \times n_2}$ is of large matrices, and bound *weighted square error* of affinities from some predicted centroid (e.g., TPMS [Charlin and Zemel, 2013] or some such NLP-based score), which yields *ellipsoidal confidence sets*. This is statistically efficient, and also computationally convenient for the adversary (convex SOCP). Axis-aligned ellipsoids also arise naturally as *Gaussian contours*, where the affinity of each paper-reviewer pair has some mean and variance, which the optimal allocation nonlinearly incorporates.

We are currently working on [Cousins et al., 2023a] generalizing to objectives to include soft robust (expected + worst-case welfare) and fairness constraints (e.g., by paper sector), as well as methods of generating smaller (more accurate and more precise) uncertainty sets using collaborative filtering and sophisticated generalization bounds (rather than *ad hoc* fixed predictive models and simple tail bounds). We are also considering temporal aspects to this problem: in an online setting (rolling review), we may need to allocate papers to reviewers now, while considering that reviewers may also be needed next month to review a new crop of papers, which results in difficult planning-style decision making under various models of uncertainty about the future.

Playing on the ideas of fairness, robustness, and uncertainty laid out in these works, I draw deep philosophical connections between them in Cousins [2023a]. The *original position* or *veil of ignorance* argument for egalitarianism of John Rawls [1971, 2001] states that our concept of fairness, justice, or welfare should be decided from behind a veil of ignorance, and thus our preferred world should consider everyone impartially (invariant to our identity). This can be posed as a zero-sum game, where a Dæmon constructs a world, and an adversarial Angel then places the Dæmon within their world. This game incentivizes the Dæmon to maximize the minimum utility over all people (i.e., to maximize *egalitarian welfare*). Thus egalitarianism arises from extreme *risk aversion* or *robustness*, and I show that by weakening the adversarial Angel, milder forms of robust objective arise, which I argue are effective *robust proxies* for *fair* learning or allocation tasks. In particular, utilitarian, Gini, and power-mean welfare and malfare concepts all arise from modified adversarial games. This has philosophical implications for the understanding each of these concepts, and, exploiting the duality between fairness and robustness, I show that these robust fairness concepts can all be efficiently optimized under mild conditions via standard maximin optimization techniques.

3 Current and Future Work

My prior work leaves many unanswered questions, both practical and theoretical, regarding welfare-centric fair machine learning. I now discuss ongoing and future work that extends the scope and usability of such methods.

Convex Optimization I am currently supervising several undergraduate projects on convex optimization of nonlinear malfare objectives. One such project applies *biased stochastic gradient descent* (bias due to the nonlinearity of malfare) to construct efficient first-order optimization routines that balance per-iterate cost-savings against a larger number of required iterations. In another student project, we observe that power-means of smooth or strongly convex per-group risk functionals are not in general smooth or strongly convex. However, we show that proximal gradient descent updates leveraging the structure of the malfare objective, i.e., if $\mathrm{R}_i(\theta)$ is the risk of group i for model parameters θ, the proximal operator

$$\theta^{(t+1)} \leftarrow \operatorname*{argmin}_{\theta \in \Theta} \mathrm{M}_p \left(i \mapsto \mathrm{R}_i(\theta^{(t)}) + \nabla_{\theta^{(t)}} \mathrm{R}_i(\theta^{(t)}) \cdot (\theta - \theta^{(t)}) \,;\, \boldsymbol{w} \right) + \frac{\gamma^{(t)}}{2} \left\| \theta - \theta^{(t+1)} \right\|_2^2 \,,$$

can yield convergence rates in terms of the smoothness or strong convexity properties *of per-group risk functionals* (rather than the entire objective), and $\mathbf{O}(\frac{1}{\varepsilon})$ iterations may suffice to ε-optimize the objective

(as opposed to $\mathbf{O}(\frac{1}{\varepsilon^2})$ iterations in general). Finally, a third student project studies the *differential-privacy implications* of fair training. Is fair training *equally private for all*, or are smaller groups "less private" (i.e., can we provide some $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\delta}_i$ for the privacy loss w.r.t. changing one sample from group i)?

Fairness Concept Elicitation I now describe an ongoing research effort that automates aspects of *fairness concept selection*, which arises in fair machine learning and allocation settings. Axiomatic theory only characterizes fairness concepts *up to the power-mean family*, and within this class, reasonable people can disagree, thus we can not normatively argue a modeler "should" adopt some fairness concept. Understanding or expressing one's fairness concept requires critical quantitative thought about a fundamental qualitative human process, and for systems that impact large numbers of people, it's worth asking, *"whose fairness concept should be optimized?"* In collaboration with Yair Zick and several of our students, the goal is to empower modelers by *interactively eliciting* human fairness concepts to within ε error, by issuing *binary queries* as to which of two outcomes is preferable. To measure distance between fairness concepts M and M$'$, we take the *supremum distance*

$$\Delta(\mathrm{M}, \mathrm{M}') \doteq \sup_{\boldsymbol{s} \in [0,1]^g} |\mathrm{M}(\boldsymbol{s}) - \mathrm{M}'(\boldsymbol{s})| \quad .$$

Bounding this metric ensures that, assuming unit-bounded utility, the true and elicited welfare functions essentially agree. We show that, for power-means, binary queries elicit halfspace cuts on p, thus $\boldsymbol{\Theta}(\log n)$ queries are *necessary and sufficient* (via binary search), where n is the number of distinct p in an ε grid w.r.t. $\Delta(\cdot, \cdot)$ distance. To bound n, we define the *minimal additive upper-bound*

$$\Delta^{\uparrow}(\mathrm{M}_p(\cdot; \boldsymbol{w}), \mathrm{M}_q(\cdot; \boldsymbol{w})) \doteq \int_p^q \sup_{\boldsymbol{s} \in [0,1]^g} \frac{\partial}{\partial r} \mathrm{M}_r(\boldsymbol{s}; \boldsymbol{w}) \, \mathrm{d}r \quad ,$$

which is the smallest upper-bound to $\Delta(\mathrm{M}_p(\cdot; \boldsymbol{w}), \mathrm{M}_q(\cdot; \boldsymbol{w}))$ for which the triangle inequality holds with equality. We show that the ε search grid on the interval $[p, q]$ contains $n \leq \frac{1}{\varepsilon} \Delta^{\uparrow}(\mathrm{M}_p(\cdot; \boldsymbol{w}), \mathrm{M}_q(\cdot; \boldsymbol{w}))$ points, with asymptotic equivalence $\lim_{\varepsilon \to 0} \varepsilon n = \Delta^{\uparrow}(\mathrm{M}_p(\cdot; \boldsymbol{w}), \mathrm{M}_q(\cdot; \boldsymbol{w}))$, thus binary search requires $\boldsymbol{\Theta}(\log \frac{1}{\varepsilon} + \log \Delta^{\uparrow}(\mathrm{M}_p(\cdot; \boldsymbol{w}), \mathrm{M}_q(\cdot; \boldsymbol{w})))$ binary elicitation queries. We are currently seeking grant funding for this project, and hope to extend our analysis to other classes of fair objective while considering human factors.

References

Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. In *International Conference on Machine Learning*, volume 162, 2022.

Richard J Arneson. Luck egalitarianism and prioritarianism. *Ethics*, 110(2):339–349, 2000.

Jeremy Bentham. An introduction to the principles of morals and legislation. *University of London: the Athlone Press*, 1789.

Laurent Charlin and Richard Zemel. The Toronto paper matching system: an automated paper-reviewer assignment system. 2013.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Cyrus Cousins. Uncertainty and the social planner's problem: Why sample complexity matters. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Cyrus Cousins. Algorithms and analysis for optimizing robust objectives in fair machine learning. In *Columbia Workshop on Fairness in Operations and AI*. Columbia University, 2023a.

Cyrus Cousins. Revisiting fair-PAC learning and the axioms of cardinal welfare. In *Artificial Intelligence and Statistics (AISTATS)*, 2023b.

Cyrus Cousins, Kavosh Asadi, and Michael L. Littman. Fair E^3: Efficient welfare-centric fair reinforcement learning. In *5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022.

Cyrus Cousins, Elita Lobo, Justin Payan, and Yair Zick. Fair resource allocation under uncertainty. In *Columbia Workshop on Fairness in Operations and AI*. Columbia University, 2023a.

Cyrus Cousins, Justin Payan, and Yair Zick. Into the unknown: Assigning reviewers to papers with uncertain affinities. In *Proceedings of the 16th International Symposium on Algorithmic Game Theory*, 2023b.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. Dividing good and better items among agents with submodular valuations. In *International Conference on Web and Internet Economics*. Springer, 2023c.

Cyrus Cousins, Vignesh Viswanathan, and Yair Zick. The good, the bad and the submodular: Fairly allocating mixed manna under order-neutral submodular preferences. In *International Conference on Web and Internet Economics*. Springer, 2023d.

Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

Gerard Debreu. Topological methods in cardinal utility theory. *Cowles Foundation Discussion Papers*, 76, 1959.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Evan Dong and Cyrus Cousins. Decentering imputation: Fair learning at the margins of demographics. In *Queer in AI Workshop @ ICML*, 2022.

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, 2012.

William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

Mackenzie Jorgensen, Elizabeth Black, Natalia Criado, and Jose Such. Supposedly fair classification systems and their impacts. *Proceedings of the 2nd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies at IJCAI*, 2022.

MacKenzie Jorgensen, Hannah Richert, Elizabeth Black, Natalia Criado, and Jose Such. Not so fair: The impact of presumably fair machine learning models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2023.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 67, page 43. Schloß Dagstuhl–Leibniz-Zentrum für Informatik, 2017.

I Elizabeth Kumar, Keegan E Hines, and John P Dickerson. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with US fair lending regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 357–368, 2022.

Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H. Bach, and Eli Upfal. Adversarial multiclass learning under weak supervision with performance guarantees. In *International Conference on Machine Learning (ICML)*, pages 7534–7543. PMLR, 2021.

John Stuart Mill. *Utilitarianism*. Parker, Son, and Bourn, London, 1863.

D Parfit. Equality and priority. *Ratio (Oxford)*, 10(3):202–221, 1997.

Arthur Cecil Pigou. *Wealth and welfare*. Macmillan and Company, limited, 1912.

John Rawls. *A theory of justice*. Harvard University Press, 1971.

John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Andrew D Selbst, Suresh Venkatasubramanian, and I Elizabeth Kumar. Deconstructing design decisions: Why courts must interrogate machine learning and other technologies. *Ohio State Law Journal*, pages 23–22, 2024.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34, 2021.

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR, 2018.