Preference-Informed Fairness

Michael P. Kim

Stanford University, CA, USA
mpk@cs.stanford.edu

Aleksandra Korolova

University of Southern California, CA, USA
korolova@usc.edu

Guy N. Rothblum

Weizmann Institute of Science, Rehovot, Israel
rothblum@alum.mit.edu

Gal Yona

Weizmann Institute of Science, Rehovot, Israel
gal.yona@weizmann.ac.il

Abstract

In this work, we study notions of fairness in decision-making systems when individuals have diverse preferences over the possible outcomes of the decisions. Our starting point is the seminal work of Dwork et al. [ITCS 2012] which introduced a notion of *individual fairness* (IF): given a task-specific similarity metric, every pair of individuals who are similarly qualified according to the metric should receive similar outcomes. We show that when individuals have diverse preferences over outcomes, requiring IF may unintentionally lead to less-preferred outcomes for the very individuals that IF aims to protect (e.g. a protected minority group). A natural alternative to IF is the classic notion of fair division, *envy-freeness* (EF): no individual should prefer another individual's outcome over their own. Although EF allows for solutions where all individuals receive a highly-preferred outcome, EF may also be overly-restrictive for the decision-maker. For instance, if many individuals agree on the best outcome, then if any individual receives this outcome, they all must receive it, regardless of each individual's underlying qualifications for the outcome.

We introduce and study a new notion of *preference-informed individual fairness* (PIIF) that is a relaxation of both individual fairness and envy-freeness. At a high-level, PIIF requires that outcomes satisfy IF-style constraints, but allows for deviations provided they are in line with individuals' preferences. We show that PIIF can permit outcomes that are more favorable to individuals than any IF solution, while providing considerably more flexibility to the decision-maker than EF. In addition, we show how to efficiently optimize any convex objective over the outcomes subject to PIIF for a rich class of individual preferences. Finally, we demonstrate the broad applicability of the PIIF framework by extending our definitions and algorithms to the multiple-task targeted advertising setting introduced by Dwork and Ilvento [ITCS 2019].

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains

Keywords and phrases algorithmic fairness

Digital Object Identifier 10.4230/LIPIcs.ITCS.2020.16

Funding *Michael P. Kim*: Part of this work completed while visiting the Weizmann Institute of Science. Supported, in part, by a Google Faculty Research Award, CISPA Center for Information Security, and the Stanford Data Science Initiative.
Aleksandra Korolova: Part of this work completed while visiting the Weizmann Institute of Science. Another part completed while visiting the Simons Institute for the Theory of Computing.
Guy N. Rothblum: Research supported by the ISRAEL SCIENCE FOUNDATION (grant number 5219/17).
Gal Yona: Research supported by the ISRAEL SCIENCE FOUNDATION (grant number 5219/17).

© Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona;
licensed under Creative Commons License CC-BY

11th Innovations in Theoretical Computer Science Conference (ITCS 2020).
Editor: Thomas Vidick; Article No. 16; pp. 16:1–16:23
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Acknowledgements This work grew out of conversations during the semester on Societal Concerns in Algorithms and Data Analysis (SCADA) hosted at the Weizmann Institute of Science. The authors thank Omer Reingold for helpful conversations, which influenced our understanding and the presentation of the work.

1 Introduction

Increasingly, algorithms are used to make consequential decisions about individuals. Examples range from determining which content users see online to deciding which applicants are considered in lending and hiring decisions. Automated decision-making comes with benefits, but it also raises substantial societal concerns (cf. [26] for a recent perspective). One prominent concern is that these algorithms might discriminate against individuals or groups in a way that violates laws or social and ethical norms [1, 29, 10, 7]. Thus there is an urgent need for frameworks and tools to mitigate the risks of algorithmic discrimination. A growing literature attempts to tackle these challenges by exploring different fairness criteria and ways to achieve them.

One prominent framework for establishing fairness in algorithmic decision-making systems comes from the seminal work of Dwork et al. [12], which introduced the notion of *individual fairness* (IF). IF relies on a task-specific similarity metric that specifies, for every pair of individuals, how similar they are with respect to the task at hand. Given such a metric, individual fairness requires that similar individuals (according to the metric) be treated similarly, i.e., assigned similar outcome distributions. This is formalized via a Lipschitz condition, requiring that for any two individuals i and j, the distance between their outcome distributions is bounded by their distance according to the metric. Although coming up with a good metric can be challenging, metrics arise naturally in prominent existing examples (e.g. credit or insurance risk scores), and in natural scenarios (e.g. a metric specified by an external regulator). Given an appropriate metric, individual fairness provides powerful protections from discrimination.

Accounting for individuals' preferences

Our work is motivated by settings in which individuals may hold diverse preferences over the possible outcomes. Natural examples of such settings include recommendation systems on professional employment websites where job-searchers have diverse considerations (geography, work-life balance, company culture, etc.) that affect their interest in potential employers, and targeted advertising systems where different users have a wide variety of preferences over the subset of ads they'd like to see out of an enormous set of possibilities. While the metric-based IF constraints prevent myriad forms of discrimination that can arise in automated decision-making systems, we argue that when individuals have different preferences over outcomes, IF can be too restrictive. Specifically, we show that in such settings, ignoring individuals' preferences (as IF does) can come at a high cost to *the very individuals that IF aims to protect.*

We illustrate this observation using a simple example. Consider a university organizing a career expo focused on software developer positions. The university would like to assign each graduating student to (at most) a single interview slot with a prospective employer. To prevent discrimination, the university would like to enforce individual fairness. For simplicity, we assume that there is an unbiased metric for judging qualifications for software development roles across employers based on GPA in the CS major. Consider candidates i, j, and k, who are all similarly qualified, and suppose there are three employers X, Y, and

Z. When the candidates are polled for their preferences, i prefers $X \succ Y \succ Z$, j prefers $Y \succ Z \succ X$, and k prefers $Z \succ X \succ Y$ (possibly due to geographic and work-life balance considerations). Despite the diversity of their preferences, since i, j, and k are all similarly qualified, IF requires that the candidates receive similar distributions over interviews with the employers X, Y, and Z. Thus, IF *rules out the allocation where each candidate gets their most-preferred interview.*

This toy example demonstrates that IF can be overly-restrictive, preventing some solutions where every individual is very happy with their outcome. Moreover, under IF, even the most socially-conscious decision-maker may be forced to disregard the preferences of some groups of individuals in order to satisfy the constraints. For example, if a decision-maker is required by IF to give similar members of majority and minority populations similar outcomes, then the decision-maker may choose the IF solution that gives everyone the outcome preferred by the majority, running the risk of ignoring the preferences of historically-marginalized groups of individuals.

Faced with this shortcoming of IF, we consider alternative notions of fairness that may be better suited to handle settings where individuals hold rich preferences over outcomes. The most natural alternative notion is *envy-freeness* (EF) [32, 14], a classic game-theoretic concept of fair division. A set of outcomes is said to be envy-free if no individual prefers the outcome given to any other individual over their own. At first glance, EF seems like a promising solution concept that addresses the concerns raised about IF: the decision where every individual receives their most-preferred outcome is EF. Indeed, Balcan et al. [4] recently presented EF as an alternative to IF in the context of fair classification.

However, we argue that EF may also be overly-restrictive, constraining the decision-maker in unreasonable ways. Returning to the example of the career expo, suppose another individual ℓ has similar preferences to i, $(X \succ Y \succ Z)$, but ℓ has a significantly lower GPA than i. Consider an allocation where i receives their most-preferred interview X, but ℓ does not receive any interview. In this case, ℓ envies i so this solution does not satisfy EF; nevertheless, the solution is reasonable from a fairness perspective. Since i has a much better GPA than ℓ, it doesn't seem unfair to give i the interview with X over ℓ, especially if the interview spots are limited.

This expanded example highlights the need for distinguishing between outcome distributions that might make some (or even all) individuals *unhappy*, from distributions that are *unfairly discriminatory*; articulating this distinction was an important conceptual contribution of the definition of individual fairness [12]. Indeed, the unqualified individual ℓ might be unhappy that they do not receive an interview; further, they might be even less happy when they see that the qualified individual i received an interview with their top choice X. In the eyes of the task-specific similarity metric, however, these two individuals are *different* – according to their GPAs, one is qualified, the other – unqualified. Thus, IF does not consider such an outcome discriminatory. Furthermore, deciding to assign no one to interviews (qualified and unqualified alike) might make no one happy, but it is not unfairly discriminatory, since all individuals are treated similarly.

In this work, we adopt the perspective that given a suitable metric, solutions that are individually fair provide strong protections from discrimination, even though they might not be envy-free. Armed with this perspective, we seek to relax the IF requirements to allow for a richer set of solutions, while still providing meaningful protections against discrimination.

1.1 This Work: Preference-Informed Fairness

Building on the perspective from [12], we propose and study the notion of *preference-informed individual fairness* (PIIF). Our guiding principle is:

> *Allocations that deviate from individual fairness may be considered fair,*
> *provided the deviations are in line with individuals' preferences.*

Before describing PIIF, we establish some notation. We model a decision-maker's policy π as a mapping from individuals to allocations, i.e., distributions over outcomes. We assume that each individual i has preferences over the possible allocations, where $p \succeq_i q$ denotes that i (weakly) prefers allocation p to allocation q. To discuss notions of individual fairness, we assume that D is a divergence where $D(p, q)$ measures some distance between two allocations p, q (e.g. the total-variation distance), and d is the task-specific metric where $d(i, j)$ specifies the similarity between individuals i and j.

Using this notation, we can restate the notions of IF and EF as follows. A policy π is *individually-fair* (IF) if for all pairs of individuals i, j, the Lipschitz condition $D(\pi(i), \pi(j)) \leq d(i, j)$ is satisfied.[1] A policy π is *envy-free* (EF) if for all individuals i, for all other individuals j, $\pi(i) \succeq_i \pi(j)$.

Preference-informed individual fairness

As in both IF and EF, PIIF establishes fairness by comparing the allocation of each individual i to the allocation of every other individual j. For each such comparison, PIIF requires that either $\pi(i)$ satisfies individual fairness with respect to $\pi(j)$ or i prefers their allocation $\pi(i)$ over some alternative allocation that would have satisfied individual fairness with respect to $\pi(j)$. More technically, for π to be considered PIIF for each individual i, we require that for every other individual j there exists some alternative allocation $p^{i;j}$ that i could have received that satisfies the IF Lipschitz condition with respect to $\pi(j)$ and where i (weakly) prefers their actual allocation $\pi(i)$ to the IF alternative $p^{i;j}$.

▶ **Definition 1** (PIIF). *A policy π that maps individuals to allocations satisfies Preference-Informed Individual Fairness with respect to a divergence D, a similarity metric d, and individual preferences $\{\succeq_i\}$, if for every individual i, for every other individual j, there exists an alternative allocation $p^{i;j}$ such that:*

- *$p^{i;j}$ is individually fair w.r.t $\pi(j)$:* $D\left(p^{i;j}, \pi(j)\right) \leq d(i, j)$.
- *i (weakly) prefers $\pi(i)$ over $p^{i;j}$:* $\pi(i) \succeq_i p^{i;j}$.

We emphasize that, in general, $p^{i;j} \neq p^{j;i}$; that is, the alternative chosen for i with respect to j's allocation need not be the same as that chosen for j with respect to i. Figure 1 provides a succinct summary of the definitions of IF, EF, and PIIF.

PIIF preserves the spirit of the core interpersonal fairness guarantee of IF: for each individual i, for every individual j who is similar to i, either i's outcome distribution is similar to j's, or i receives an even better (more-preferred) outcome distribution. The main advantage of PIIF over IF is that it allows for a much richer solution space, which can lead to preferable outcomes for individuals. Further, PIIF does not restrict the allocations unnecessarily; as in IF, the constraints only bind when a pairs of individuals are sufficiently similar according to the metric. In other words, PIIF – unlike EF – permits solutions that may be disappointing to some individuals (i.e. where i envies j) but should not be considered discriminatory (because i and j are substantially different according to the task at hand).

[1] Throughout, we assume that d and D are scaled appropriately to be in the same "units." That is, without loss of generality, we assume the relevant Lipschitz constant in the IF-style constraints is 1.

Individual Fairness (IF)
for every i, for every j:
$D(\pi(i), \pi(j)) \leq d(i, j)$

Envy-Freeness (EF)
for every i, for every j:
$\pi(i) \succeq_i \pi(j)$

Preference-Informed Individual Fairness (PIIF)
for every i, for every j, there exists $p^{i;j}$ s.t.
$D(p^{i;j}, \pi(j)) \leq d(i, j)$
$\pi(i) \succeq_i p^{i;j}$

Figure 1 Summary of individual fairness notions.

Referring back to the career expo example, we note that the allocation where the three qualified candidates i, j, and k (deterministically) interview with their preferred employer is PIIF. To see this, consider i comparing their outcome to those of j and k under such an interview assignment. Comparing with j, i prefers outcome X to receiving outcome Y, which would satisfy the IF constraint with respect to j. Similarly, she prefers X to Z, which would satisfy the IF constraint with respect to k. Indeed, since i receives her preferred outcome, one can argue that there is no discrimination against i in the allocation. Similar reasoning applies to j and to k. In fact, the allocation where each individual deterministically receives their preferred outcome is always PIIF, a property we find desirable for a fairness definition. Further, consider the allocation of ℓ, who we assumed was significantly less qualified than i (and thus, j and k). If ℓ is sufficiently dissimilar to all other candidates, then the scheduler can assign ℓ to any interview and still satisfy PIIF. To see this, note that if $d(\ell, i)$ is sufficiently large, we can always take $p^{\ell;i} = \pi(\ell)$, and the constraints for individual ℓ with respect to i will be satisfied (with identical arguments when comparing ℓ to j and k).

1.2 Our Contributions

Our running example illustrates that in many reasonable situations (involving rich and diverse individual preferences over outcomes), the existing notions of individual fairness and envy-freeness may not capture an appropriate notion of fairness or may unnecessarily constrain the decision-maker. In high-stakes domains, such as employment and personalized content selection, both limitations are significant and may hinder adoption of fairness-conscious decision-making. We propose PIIF as a relaxation of IF that addresses the identified shortcomings of existing notions while still providing meaningful protections against discrimination. We view this as an important conceptual contribution in its own right.

With the motivation and definition for PIIF in place, we provide a comprehensive characterization of the relationship between PIIF and other individual notions of fairness. In Section 2, we show formally that PIIF can be viewed as a relaxation of both IF and EF; that is, any solution that satisfies either IF or EF also satisfies PIIF. Further, we demonstrate that PIIF is a non-trivial relaxation of both notions, by proving that there exist settings in which PIIF solutions cannot be captured by IF or EF constraints alone, for *any* choice of metrics d, D and preferences.

To introduce PIIF, we have argued qualitatively that relaxing IF to PIIF allows for more preferable outcomes for individuals. We quantify these claims by comparing the *social welfare* of a decision-maker's policy achievable under PIIF and under IF. In Section 3, we show optimal bounds on the ratio of the best social welfare under PIIF to that under IF; the ratio can grow *linearly* in the number of individuals classified or in the number of possible outcomes grows.

With the definition and properties of PIIF in place, we turn our attention to the algorithmic question of how to achieve PIIF. In Section 4, we show that for a rich family of individual preferences, there is an efficient algorithm to minimize a convex objective subject to PIIF. The result follows by observing that for structured classes of preferences, the set of PIIF constraints is convex. In particular, to optimize over PIIF, we can augment the convex program defined for IF in [12] to capture the additional preference constraints. As such, optimization subject to PIIF is only slightly more complex than optimization subject to IF.

Finally, we demonstrate the versatility of the PIIF framework, by applying preference-informed fairness in the context of targeted advertising (as studied by [13]). Recent empirical findings demonstrate that the ad allocation algorithms run by online advertising platforms may result in discrimination [11, 24, 1] and are thus facing legal scrutiny [31, 29, 6]. As such, developing formal frameworks for understanding fairness in such advertising systems is of great importance. In Section 5, we extend our definition of PIIF and our results to the multiple-task setting defined [13] to model fairness desiderata for the domain of large-scale targeted advertising. We show that in this practically-motivated setting, IF still may restrict the social welfare considerably compared to PIIF *even when the individuals' similarity and preferences are perfectly aligned!* The ratio of the best social welfare under PIIF to that of IF grows *linearly* in the number of tasks.

Organization

Sections 2-5 contain the technical details and proofs of our major contributions with some results deferred to the appendix. We conclude in Section 6 with comparisons to other related works and a discussion of the strengths and limitations of the current approach of preference-informed fairness as well as directions for future investigations.

2 Preference-Informed Individual Fairness

Preliminaries

Given a set of individuals $\mathcal{X}$, we consider policies that assign every individual to an outcome in the set C. We allow randomized allocation rules $\pi : \mathcal{X} \to \Delta(C)$, where for each individual $i \in \mathcal{X}$, their allocation $\pi(i) \in \Delta(C)$ represents a distribution over outcomes $c \in C$. We model individuals' preferences by assuming that every individual $i \in \mathcal{X}$ has a reflexive and transitive binary relation $\succeq_i$ that encodes their preferences over allocations in $\Delta(C)$; for $p, q \in \Delta(C)$, we use $p \succeq_i q$ to denote that i (weakly) prefers p to q.[2] We use $\succeq$ to denote the set of individuals' preference relations, $\succeq = \{\succeq_i\}_{i \in \mathcal{X}}$.

One important structured class of preference relations are those that admit a *utility function*. Here, we assume each individual $i \in \mathcal{X}$ has a real-valued function over allocations $u_i : \Delta(C) \to \mathbb{R}$, where $u_i(\pi(i))$ represents the utility to individual i from the allocation given by π. Given such a utility function, $p \succeq_i q$ if and only if $u_i(p) \geq u_i(q)$.

With this technical notation in place, for completeness, we restate the three definitions of fairness.

▶ **Definition 2** (Individual Fairness). *Given a divergence $D : \Delta(C) \times \Delta(C) \to [0, 1]$ and a similarity metric $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$, a policy $\pi : \mathcal{X} \to \Delta(C)$ is (D, d)-individually fair if for every two individuals $i, j \in \mathcal{X} \times \mathcal{X}$, the following Lipschitz condition holds.*

$$D(\pi(i), \pi(j)) \leq d(i, j) \tag{1}$$

[2] $\succeq_i$ need not be *total* nor *antisymmetric* over $\Delta(C)$.

▶ **Definition 3** (Envy Freeness). *Given a set of preferences $\succeq$, a policy $\pi : \mathcal{X} \to \Delta(C)$ is $\succeq$-envy-free if for all individuals $i \in \mathcal{X}$, and for all other individuals $j \in \mathcal{X}$,*

$$\pi(i) \succeq_i \pi(i) \tag{2}$$

▶ **Definition 4** (Preference-Informed Individual Fairness). *Given a divergence $D : \Delta(C) \times \Delta(C) \to [0, 1]$, a similarity metric $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$, and a set of preferences $\succeq$, a policy $\pi : \mathcal{X} \to \Delta(C)$ is $(D, d, \preceq)$-PIIF if for all individuals $i \in \mathcal{X}$, for all other individuals $j \in \mathcal{X}$, there exists an allocation $p^{i;j} \in \Delta(C)$ such that:*

$$D\left(p^{i;j}, \pi(j)\right) \le d(i, j) \tag{3}$$

$$\pi(i) \succeq_i p^{i;j} \tag{4}$$

Often, the divergence D, metric d, and preferences $\succeq$ will be fixed. In these contexts, we use $\Pi^{\mathrm{IF}}, \Pi^{\mathrm{EF}}, \Pi^{\mathrm{PIIF}}$ to denote the set of IF, EF, and PIIF solutions, respectively.

2.1 PIIF relaxes IF and EF

We have argued informally that PIIF captures the appealing aspects of both IF (strong discrimination protections) and EF (respecting the preferences of individuals) without being overly prescriptive in a way that might hurt individuals or the decision-maker. Our first result formalizes these claims, by characterizing PIIF as a relaxation of both IF and EF. We show that any policy that is either IF or EF is also PIIF.

▶ **Proposition 5.** *Fixing a divergence, the metric, and preferences, $\Pi^{\mathrm{IF}} \subseteq \Pi^{\mathrm{PIIF}}$ and $\Pi^{\mathrm{EF}} \subseteq \Pi^{\mathrm{PIIF}}$.*

As solution concepts, both IF and EF are always feasible, but for very different reasons: for IF, any allocation that treats all individuals identically is feasible; for EF, the allocation that gives everyone their most-preferred outcome is envy-free. Thus, both of these extreme solutions will also be feasible for PIIF. In general, PIIF will be a strict relaxation of these concepts that allows for interpolation between the notions. Intuitively, more diverse preferences of individuals tend to give rise to richer sets of PIIF solutions compared to IF, and nontrivial metrics d (i.e., further from the all-zeros "metric") give rise to richer sets of PIIF solutions compared to EF. Given the right framing, the proof of this result is almost immediate.

Proof. To see that an IF policy π satisfies PIIF, for each i, we take $p^{i;j} = \pi(i)$ for all j. Consider an allocation $\pi \in \Pi^{\mathrm{IF}}$. From the perspective of any individual $i \in \mathcal{X}$, when comparing to individual $j \in \mathcal{X}$, if $p^{i;j} = \pi(i)$, then, by the fact that π satisfies IF, condition (3) is satisfied. By reflexivity of $\succeq_i$, (4) is also satisfied, so $\pi \in \Pi^{\mathrm{PIIF}}$.

To see that an EF policy π satisfies PIIF, for each i, we take $p^{i;j} = \pi(j)$ for all j. Consider an allocation $\pi \in \Pi^{\mathrm{EF}}$. From the perspective of any $i \in \mathcal{X}$, when comparing to $j \in \mathcal{X}$, if $p^{i;j} = \pi(j)$, then, condition (3) is satisfied trivially because $D(\pi(j), \pi(j)) = 0$. Since π satisfies EF, we know that $\pi(j) \preceq_i \pi(i)$, so condition (4) also holds; thus $\pi \in \Pi^{\mathrm{PIIF}}$. ◀

PIIF generalizes IF and EF

We remark that this intuition also shows that PIIF is a *generalization* of both IF and EF; that is, both notions can be "implemented" as special cases of PIIF. To implement IF, we can set all individual's preference relation $\succeq_i$ to be the trivial reflexive relation, where for all allocations p, $p \succeq_i p$, and for all nontrivial pairs $p \neq q$, p and q are incomparable. To implement EF, we simply take $d(i,j) = 0$ for all i, j pairs. In other words, we can think of the set of IF solutions as those where we require the alternative allocation for i compared to j to be i's actual allocation $p^{i;j} = \pi(i)$, and we can think of the set of EF solutions as those where we require the alternative allocation for i compared to j to be j's allocation $p^{i;j} = \pi(j)$.

PIIF is a meaningful relaxation of IF and EF

A natural question to ask is whether we need to introduce a new definition of individual fairness. In particular, we might hope that we could "implement" PIIF using IF with a metric that incorporates preferences or with EF with preferences that incorporate distances. We argue that when there is a rich set of possible outcomes and a correspondingly-rich set of possible preferences, such an approach is infeasible. In particular, PIIF captures constraints that could not be cast within the language of IF or EF alone.

To build intuition, we revisit the career expo example: suppose that two similarly qualified individuals i and j have a similar top choice (say, X), but disagree on their second choice (i prefers Y, whereas j prefers Z). Do these individuals have similar preferences or divergent ones? Intuitively, a fair assignment could give them similar probabilities of seeing X, but different probabilities of seeing Y and Z. Individual fairness treats all outcomes symmetrically for all individuals, and does not let us make such distinctions. The following proposition strengthens this intuition, demonstrating that there are in fact settings in which EF preferences cannot be encoded using any IF metric, and vice versa. Note that this implies that PIIF – a relaxation of both notions – captures constraints that cannot be cast within the language of IF or EF alone.

▶ **Proposition 6.** *There exists a set of preferences $\succeq$ such that for any choice of divergence D and metric d*

$$\Pi^{\succeq\text{-EF}} \neq \Pi^{(D,d)\text{-IF}}.$$

There exists a divergence-metric pair D, d such that for any choice of preferences $\succeq$,

$$\Pi^{(D,d)\text{-IF}} \neq \Pi^{\succeq\text{-EF}}.$$

Proof. In both constructions, we will assume there are two disjoint groups of individuals $S, T \subseteq \mathcal{X}$. Consider two outcomes $p, q \in C$. Suppose $\succeq$ is such that for some $i \in S$ and $j \in T$, $p \succ_i q$ and $q \succ_j p$. Consider any D and d: if $D(p,q) \leq d(j,i)$, then assigning p to j and q to i will be (D,d)-IF, but it is not $\succeq$-EF; otherwise, if $D(p,q) > d(i,j)$, then assigning p to S and q to T will not be (D,d)-IF, even though it is $\succeq$-EF. Thus, no D, d can capture $\succeq$-EF.

Now take D to be total variation distance and consider a metric d where $d(i,j) = 0$ for $i, j \in S \times S$ and $T \times T$, and $d(i,j) = 1$ for $i, j \in S \times T$. Under this metric, assigning any fixed allocation to everyone in S and any (potentially-different) fixed allocation to everyone in T is (D,d)-IF. Consider some $\succeq$. If there is some $i \in S$ such that $p \succ_i q$ or if there is some $j \in T$ such that $q \succ_j p$, then the (D,d)-IF allocation that assigns q to every $i \in S$ and p to every $j \in T$ is not $\succeq$-EF.

Thus, for all individuals in $i \in S \cup T$, $\succeq_i$ must be either the relation, where $p \equiv q$ or the trivial reflexive relation where p and q are incomparable. Suppose $i, j \in S \times S$ both have $\succeq_i = \succeq_j = \equiv$. Then, the solution that assigns p to i and q to j is $\succeq$-EF, but violates the Lipschitz condition of (D,d)-IF. On the other hand, if there is some $i \in S$ that holds the trivial reflexive relation, then the (D,d)-IF solution that assigns p to all of S and q to all of T will not satisfy $\succeq$-EF, because $p \not\succeq_i q$. ◀

2.2 Metric Envy-Freeness

We arrived at PIIF by starting with the metric-based IF as a strong notion of nondiscrimination and relaxing the notion to incorporate individuals' preferences and allow for a richer set of solutions while providing a meaningful protections against discrimination. A conceptually-different approach towards these goals would start with preference-based EF, but allow the decision-maker some freedom by incorporating distances between individuals. In particular, consider the following relaxation of EF, which we call *Metric Envy-Freeness* (MEF), that intuitively captures the idea that no individual should envy the allocation of any other *similar* individual.

▶ **Definition 7.** *Suppose each individual $i \in \mathcal{X}$ has a utility function $u_i : \Delta(C) \to \mathbb{R}$; let $\mathcal{U} = \{u_i\}_{i \in \mathcal{X}}$. Given a similarity metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, a policy $\pi : X \to \Delta(C)$ satisfies $(d, \mathcal{U})$-metric-envy-freeness if for every individual $i \in \mathcal{X}$, for every other individual $j \in \mathcal{X}$,*

$$u_i(\pi(i)) \geq u_i(\pi(j)) - d(i,j)$$

This definition starts with the envy-freeness constraint for utility-based preferences, but then relaxes the constraint between i and j by their distance according to the metric. For the metric-utility comparison of MEF to be meaningful, we assume that utilities and metric distances are normalized to one another; without loss of generality, assume that each utility and metric distance is bounded in $[0,1]$. For each pair i, j, the notion interpolates between two extremes based on the value of $d(i,j)$: if $d(i,j) = 0$, then envy-freeness binds; when $d(i,j) = 1$, the allocation i receives is not constrained by the allocation j receives.

As $d(i,j) \geq 0$ for all pairs of individuals, MEF is clearly a relaxation of EF. That said, it's not immediately obvious how MEF relates to IF or PIIF. While conceptually different, we show that MEF captures a closely-related notion of fairness to PIIF, in the special case where preferences are given by structured utility functions. To relate MEF to PIIF, we need to assume the following Lipschitz conditions.

▶ **Definition 8** (Lipschitz utility). *A utility function $u : \Delta(C) \to \mathbb{R}$ is ℓ-Lipschitz with respect to $D : \Delta(C) \times \Delta(C) \to \mathbb{R}^+$ if $|u(p) - u(q)| \leq \ell \cdot D(p,q)$.*

Lipschitz utility functions are quite natural. For instance, taking D to be the total variation distance, if individuals' preferences admit an expected utility function, where each outcome has utility in $[0,1]$, then individuals' utilities will be 1-Lipschitz. In other words, individuals' utilities are not highly sensitive to very small changes in the allocation they receive.

▶ **Definition 9** (Reverse-Lipschitz utility). *A utility function $u : \Delta(C) \to \mathbb{R}$ is ℓ-reverse-Lipschitz with respect to $D : \Delta(C) \times \Delta(C) \to \mathbb{R}^+$ if $\frac{1}{\ell} \cdot D(p,q) \leq |u(p) - u(q)|$.*

Reverse-Lipschitz utility functions are less natural. This assumption implies that no pair of outcomes is valued very similarly. One natural setting where the reverse-Lipschitz condition holds nontrivially is in the case of binary outcomes, where each individual prefers one outcome over the other. Under these assumptions, we can show the following relationship between MEF and PIIF.

▶ **Theorem 10.** *Suppose $D : \Delta(C) \times \Delta(C) \to \mathbb{R}^+$ is a divergence, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a similarity metric, and $\mathcal{U} = \{u_i\}$ is a family of utility functions. Let $\succeq^{\mathcal{U}}$ denote the family of preferences induced by $\mathcal{U}$. Consider a policy $\pi : \mathcal{X} \to \Delta(C)$. For some constant $\ell \geq 1$:*

- *Suppose for all $i \in \mathcal{X}$, u_i is ℓ-Lipschitz with respect to D. Then,*

$$\Pi^{(D,d,\succeq^{\mathcal{U}})\text{-PIIF}} \subseteq \Pi^{(\ell\cdot d,\mathcal{U})\text{-MEF}}.$$

- *Suppose for all $i \in \mathcal{X}$, u_i is ℓ-reverse-Lipschitz with respect to D. Then,*

$$\Pi^{(d,\mathcal{U})\text{-MEF}} \subseteq \Pi^{(D,\ell\cdot d,\succeq^{\mathcal{U}})\text{-PIIF}}.$$

Proof. To see that PIIF implies MEF, we start with a policy π that satisfies $(D, d, \succeq^{\mathcal{U}})$-PIIF. To establish MEF, we compare the utility of individual i on their allocation $\pi(i)$ to that of another individual $\pi(j)$; we denote by $p^{i;j}$ the alternative allocation for i that satisfies the PIIF constraints.

$$\begin{aligned} u_i(\pi(i)) &\geq u_i(p^{i;j}) && (5)\\ &\geq u_i(\pi(j)) - \left[u_i(\pi(j)) - u_i(p^{i;j})\right] \\ &\geq u_i(\pi(j)) - \ell \cdot D\left(\pi(j), p^{i;j}\right) && (6)\\ &\geq u_i(\pi(j)) - \ell \cdot d(i,j) && (7)\end{aligned}$$

where (5) follows by the fact that π satisfies PIIF; (6) follows by the Lipschitz condition; and (7) follows again from the fact that π satisfies PIIF. Thus, $u_i(\pi(i)) \geq u_i(\pi(j)) - \ell \cdot d(i,j)$, so π is $(\ell \cdot d, \mathcal{U})$-MEF.

To see that MEF implies PIIF, we start with a policy π that satisfies $(d,\mathcal{U})$-MEF. To establish PIIF, we consider an arbitrary pair of individuals i and j, and exhibit an allocation $p^{i;j}$ that satisfies the PIIF conditions with respect to $\pi(i)$ and $\pi(j)$. Comparing individual i to individual j, we consider two cases.

First, suppose $u_i(\pi(i)) \geq u_i(\pi(j))$ and take $p^{i;j} = \pi(j)$. In this case, the Lipshitz constraint is trivially satisfied,

$$D(p^{i,j}, \pi(j)) = D(\pi(j), \pi(j)) = 0 \leq \ell \cdot d(i,j)$$

and the assumption that $u_i(\pi(i)) \geq u_i(\pi(j))$ implies that

$$\pi(i) \succeq_i \pi(j) = p^{i;j},$$

so the PIIF constraints are satisfied.

Next, suppose $u_i(\pi(i)) < u_i(\pi(j))$ and take $p^{i;j} = \pi(i)$. In this case, the preference condition is trivially satisfied,

$$\pi(i) \succeq_i \pi(i) = p^{i;j}$$

and the Lipschitz condition follows as

$$\begin{aligned} D(p^{i;j}, \pi(j)) &= D(\pi(i), \pi(j)) \\ &\leq \ell \cdot (u_i(\pi(j)) - u_i(\pi(i))) && (8)\\ &\leq \ell \cdot d(i,j) && (9)\end{aligned}$$

where (8) follows by the reverse-Lipschitz condition and (9) follows by the fact that π satisfies MEF. Thus, π is $(D, \ell \cdot d, \succeq^{\mathcal{U}})$-PIIF. ◀

Relating PIIF and MEF

Theorem 10 shows that under appropriate assumptions, we can relate the notions of PIIF and MEF. We interpret the theorem to say that, in most settings we would apply preference-informed fairness, MEF is a strictly more relaxed notion than PIIF; that is, every PIIF solution will be MEF, but there will be MEF solutions that do not satisfy PIIF.

Specifically, as we remarked earlier, it is reasonable to assume that individuals' utility functions are 1-Lipschitz with respect to D. In this case, small changes in the allocation according to D cannot result in dramatically different utilities, and Theorem 10 says that $(D, d, \succeq^{\mathcal{U}})$-PIIF implies $(d, \mathcal{U})$-MEF.

In general, with a rich set of outcomes, we do not expect that individuals' utility functions will be reverse-Lipschitz with respect to D. As such, there are cases when $(d, \mathcal{U})$-MEF will be considerably more relaxed than $(D, d, \succeq^{\mathcal{U}})$-PIIF, allowing for deviations from (D, d)-IF that do not give utility improvements for all individuals. To see this point, consider the following example with two individuals i, j, where $d(i, j) = 0.5$ and two outcomes p, q, with the utility functions of i and j defined as follows.

	p	q
u_i	1.0	0.5
u_j	0.5	1.0

We take D to be the total variation distance between allocations. It's easy to verify that all allocations that treat the individuals identically and the welfare-maximizing solution where i receives p and j receives q will be MEF; this can be seen by noting that each of these solutions satisfies EF, thus, also MEF. In fact, in this example, the utility functions are sufficiently Lipschitz to guarantee that all PIIF solutions will be MEF.

On the other hand, consider the allocation where i receives q and j receives p, deterministically. This solution is not IF, because $1 = D(q, p) > d(i, j) = 0.5$, nor EF, because both individuals envy the others' solution. Further, the solution is not PIIF. To see this, note that because each individual receives their least favorite outcome, the only surrogate $p^{i;j}$ that can be chosen for individual i such that $q \succeq_i p^{i;j}$ is $p^{i;j} = q$ (similarly, $p^{j;i} = p$ for individual j). As the actual allocation is not IF, one of the PIIF constraints will be violated.

Still, this allocation does satisfy MEF. Specifically, $0.5 = u_i(q) \geq u_i(p) - d(i, j) = 1.0 - 0.5$ and $0.5 = u_j(p) \geq u_j(q) - d(j, i) = 1.0 - 0.5$. In other words, in this instance, MEF allows for a solution that is not permitted by any of the other notions of individual fairness we consider. In particular, the allocation deviates from individual fairness in a way that does not help either individual. This example suggests that in reasonable settings, MEF is a strictly weaker concept than PIIF and may be too permissive.

3 Welfare under IF, EF, and PIIF

In order to motivate PIIF, we argued that IF may be overly-restrictive and limit the "quality" of solutions from the perspective of individuals. In this section, we formalize this claim, showing that PIIF admits solutions that can be significantly more preferable to individuals. We quantify the idea of individual quality using the game-theoretic notion of social welfare.

We restrict our attention to the common setting where individuals hold preference relations that admit a utility function. In this setting, given a policy π, we can track the *social welfare* of the policy, defined to be the overall utility experienced by individuals.

▶ **Definition 11.** *Given a policy $\pi : \mathcal{X} \to \Delta(C)$, social welfare $\mathcal{W}(\pi)$ is the sum of the individuals' utilities under π.*

$$\mathcal{W}(\pi) = \sum_{i \in \mathcal{X}} u_i(\pi(i)) \tag{10}$$

For a collection of allocations Π, we let $\mathcal{W}^*(\Pi) = \max_{\pi \in \Pi} \mathcal{W}(\pi)$ denote the optimal social welfare achievable by any allocation in Π, and let $\mathcal{W}^* = \mathcal{W}^*(\Delta(C)^{\mathcal{X}})$ denote the optimal (unconstrained) social welfare.

An important special case of preferences that admit a utility function are those that admit an *expected utility function.* Using such preferences is a standard approach in economics for modeling decision-making in the presence of uncertainty.[3]

▶ **Definition 12** (Expected utility representation)**.** *A preference relation $\succeq$ admits an expected utility representation if and only if there exists a function $u : C \to \mathbb{R}^+$, such that for any two allocations $p, q \in \Delta(C)$,*

$$p \succeq q \iff \sum_{c \in C} p_c \cdot u(c) \geq \sum_{c \in C} q_c \cdot u(c) \tag{11}$$

Individual fairness may restrict social welfare

Using the notation introduced above, note that $\mathcal{W}^*(\Pi^{\text{PIIF}}) = \mathcal{W}^*$ because the welfare-maximizing allocation is feasible for PIIF. Here, we aim to understand how much IF may restrict the best social welfare compared to PIIF, by relating $\mathcal{W}^*(\Pi^{\text{IF}})$ to $\mathcal{W}^*$.

Intuitively, social welfare can be hurt significantly by requiring IF compared to PIIF when many individuals are considered similar, but there is a diversity of preferences over outcomes. Formalizing this intuition, we argue that without any assumptions about the class of utility functions of individuals, the ratio between the best social welfare under IF and PIIF can grow with the number of individuals. Further, even under the stronger assumption that individuals' preferences admit an expected utility representation, the ratio can grow with the number of outcomes.

▶ **Theorem 13.** *There exists a family of instances such that*

$$\frac{\mathcal{W}^*}{\mathcal{W}^*\left(\Pi^{\text{IF}}\right)} \geq |\mathcal{X}|\,.$$

Additionally, there exists a family of instances where individuals' preferences admit an expected utility representation and

$$\frac{\mathcal{W}^*}{\mathcal{W}^*\left(\Pi^{\text{IF}}\right)} \geq |C|\,.$$

Proof. The two claims of the theorem follow by similar constructions. Suppose every individual is considered similar according to d; that is, for all $i, j \in \mathcal{X} \times \mathcal{X}$, $d(i, j) = 0$. This means that any IF solution must assign every individual the same distribution over outcomes. To show the gaps, we will compare the best IF solution (i.e. constant allocation) to the welfare-maximizing solution, which is feasible under PIIF.

[3] Although not all preference relations admit this form, a rich class of preferences do. For example, different levels of tolerance towards risk can be captured within this framework (e.g., a risk-averse individual would have a utility function u which is concave). Von Neumann and Morgensterm [33] provide a complete characterization of this class of preference relations.

To begin, suppose we allow individuals to specify arbitrary utility functions; let each individual $i \in \mathcal{X}$ hold a distinct $p_i \in \Delta(C)$ (i.e., $p_i \neq p_j$ for all $i \neq j$) such that $u_i(p_i) = 1$ and $u_i(q) = 0$ for any $q \neq p_i$. In this case, the optimal social welfare is $\mathcal{W}^* = |\mathcal{X}|$. For any fixed allocation, however, the best social welfare is to choose a distribution over the set of $\{p_i\}$. Any such distribution will achieve welfare 1; thus, $\mathcal{W}^*(\Pi^{\mathrm{IF}}) = 1$.

Now, suppose every individual is required to specify an expected utility representation. Let each individual choose some $c \in C$, such that a $1/|C|$-fraction of individuals prefer each outcome c; let $u_i(c) = 1$ for their preferred outcome and $u_i(c) = 0$, otherwise. Again, the optimal social welfare is $\mathcal{W}^* = |\mathcal{X}|$. Under any policy that assigns every individual the same fixed allocation p, the social welfare is given by $\sum_{i \in \mathcal{X}} \sum_{c \in C} p_c \cdot u_i(c) = \frac{|\mathcal{X}|}{|C|} \cdot \sum_{c \in C} p_c = \frac{|\mathcal{X}|}{|C|}$. Thus, $\mathcal{W}^*(\Pi^{\mathrm{IF}}) = \frac{|\mathcal{X}|}{|C|}$. ◀

We note that the gaps demonstrated in Theorem 13 are optimal in their settings. In particular, any constant allocation will be IF, and we can always recoup a $1/|\mathcal{X}|$ fraction of the social welfare with a constant allocation tailored for the individual with the highest utility. Further, in the case where preferences admit an expected utility representation, we can choose the constant allocation on the $c \in C$ of maximum welfare. Finally, we note that one unsatisfying aspect of these constructions is that they rely on the fact that all individuals are similar according to d. In Section 5.1, we show that in the multiple-task setting of [13], such gaps exist even with nontrivial metrics that seem to be aligned with social welfare.

PIIF does not guarantee social welfare

Because PIIF is a relaxation of IF, the best social welfare achievable under PIIF is always at least that of IF. That said, because PIIF is a strict relaxation of IF, it does not necessarily guarantee that *every* allocation's social welfare improves under PIIF. In particular, when the decision-maker seeks to optimize a utility function that runs against individuals' utilities within the set of PIIF solution, the obtained social welfare may be arbitrarily worse under the PIIF constraints than IF constraints.

Suppose that the decision-maker has an additive utility function of the form $f(\pi) = \sum_{i \in \mathcal{X}} f_i(\pi(i))$, for $f_i : \Delta(C) \to \mathbb{R}$. Let $\pi_f^{\mathrm{IF}} = \mathrm{argmax}_{\pi \in \Pi^{\mathrm{IF}}} f(\pi)$ and $\pi_f^{\mathrm{PIIF}} = \mathrm{argmax}_{\pi \in \Pi^{\mathrm{PIIF}}} f(\pi)$ denote the optimal IF (resp., PIIF) solution in terms of $f(\cdot)$.

▶ **Proposition 14.** *There exists a family of instances and an additive utility function f such that*

$$\mathcal{W}(\pi_f^{\mathrm{PIIF}}) = 0 < \mathcal{W}(\pi_f^{\mathrm{IF}}).$$

Proof. Suppose there are two disjoint classes of individuals, S and T, which each make up half of $\mathcal{X}$, and all individuals are similar; for all $i, j \in \mathcal{X} \times \mathcal{X}$, $d(i, j) = 0$. Suppose there are three outcomes $C = \{p, q, r\}$. Consider the utility functions defined as follows.

	p	q	r
$f_{i \in S}$	$1/2 + \varepsilon$	1	0
$f_{j \in T}$	$1/2 + \varepsilon$	0	1
$u_{i \in S}$	1	0	0
$u_{j \in T}$	1	0	0

Under IF, the decision-maker must treat all individuals identically. Given this constraint, the outcome that maximizes f is allocating p deterministically to everyone. This allocation π_f^{IF} achieves $\mathcal{W}(\pi_f^{\mathrm{IF}}) = 1$. Without the constraint that all individuals' allocations are identical, the allocation that assigns q to individuals from S and r to individuals from T maximizes f. In fact, this allocation will be feasible for PIIF: note that every individual experiences 0 utility from everyone's allocation, so no one envies anyone else; under PIIF, envy-free solutions are feasible. Thus, $\mathcal{W}(\pi_f^{\text{PIIF}}) = 0$. In fact, this reveals that the proposition holds not only for PIIF, but also for any notion that relaxes EF. ◀

This construction demonstrates that the PIIF constraints alone do not guarantee improved social welfare compared to the IF constraints. We remark, however, that this is in line with our initial motivation for PIIF: decoupling the objective of provably preventing discrimination from the objective of ensuring beneficial outcomes in aggregate. Finally, we note that if this is a concern, it can be addressed within our framework by adding a constraint to the optimization program, discussed in detail in Section 4, that ensures the social welfare is above some baseline. In particular, an appropriate individually fair solution could act as this baseline, by first computing the social welfare obtained by it and then requiring that the resulting PIIF solution has at least this social welfare. At the extreme, we could even add such a constraint on the utility experienced by each individual; thus, obtaining the guarantee that any deviations from an IF solution are optimal, from the individuals' perspective, compared to some benchmark IF solution.

4 Optimization subject to PIIF

As we have argued, satisfying PIIF is always feasible: on the one hand, we can take any IF solution, including a trivial policy that treats all individuals identically; alternatively, we can take any EF solution, including the welfare-maximizing policy that gives everyone their most-preferred allocation. In this section, we study the question of efficient optimization of a decision-maker's utility function subject to PIIF constraints. As is standard in much of learning and optimization, we frame this task as the following minimization problem:

$$\begin{aligned} \underset{\pi:\mathcal{X}\to\Delta(C)}{\text{minimize}} \quad & f(\pi) \\ \text{subject to} \quad & \pi \in \Pi^{\text{PIIF}} \end{aligned}$$

In this section, we answer the question of feasibility of efficient optimization in the positive when f is convex, and the preferences arise from a structured, but rich class of relations.

4.1 Structured preferences

In principle, PIIF can be instantiated with any notion of preference. Without assuming anything about the preferences, however, the PIIF constraints could be difficult to handle: the space of allocations, over which the PIIF constraints are defined, is exponential. In realistic settings, where the number of individuals or outcomes is large, this exponential dependence may be intractable. Towards efficient optimization, we focus on two rich and structured preference classes.

First, we include the prominent class of preferences that admit an expected utility representation, as defined in Definition 12. Additionally, we include the class of *stochastic domination* preferences. Stochastic domination formalizes the intuition that for any distribution over outcomes, a shift of probability mass from less desirable outcomes to more desirable outcomes is considered preferable. Viewing an allocation $p \in \Delta(C)$ as a discrete probability distribution, we denote by $c \sim p$ an outcome randomly sampled from p.

▶ **Definition 15** (Stochastic domination)**.** *For an individual with a utility function $u : C \to [0, M]$ and for any two allocations $p, q \in \Delta(C)$, p stochastically dominates q if*

$$p \succeq q \iff \forall x \in [0, M], \quad \mathop{\mathbf{Pr}}_{c \sim p}[u(c) \geq x] \geq \mathop{\mathbf{Pr}}_{c \sim q}[u(c) \geq x]$$

That is, an allocation p is (weakly) preferred over q if for every possible level of utility x, the probability of achieving at least x is no worse under p than it is under q. Note that stochastic domination represents an interesting example of a non-total preferences, as two allocations may be incomparable.[4]

4.2 Efficient optimization subject to PIIF

Here, we prove that when individuals' preferences are of the forms defined above, the PIIF constraints admit efficient optimization. Formally, the following theorem demonstrates that when the divergence over allocations D is taken to be total variation distance D_{tv}, and assuming oracle access to the individual-fairness metric d, we can write the PIIF constraints as a set of (polynomially-many) linear inequalities; thus, we can efficiently minimize any convex objective f.

▶ **Theorem 16.** *Let $\succeq = \{\succeq_i\}_{i \in \mathcal{X}}$ be the set of individuals' preferences. If every $\succeq_i$ is either the stochastic domination relation or admits an expected utility representation, then the set of $(D_{\text{tv}}, d, \succeq)$-PIIF allocations forms a convex polytope in $\mathbb{R}^k$, where $k = \text{poly}(|\mathcal{X}|, |C|)$.*

Proof. We specify the PIIF constraints using the following variables: for all $i \in \mathcal{X}$, let $\pi(i) \in \Delta(C)$ be a vector denoting the actual allocation; for every pair of individuals $(i, j) \in \mathcal{X} \times \mathcal{X}$, let $p^{i;j} \in \Delta(C)$ be a vector denoting the alternative allocation for i when comparing to j. We argue that the PIIF constraints given in (3) and (4) can each be written as linear inequalities over these variables.

First, since D is taken to be the total variation distance we can translate (3) as $\frac{1}{2} \cdot \sum_{c \in C} \left| p_c^{i;j} - \pi(j)_c \right| \leq d(i, j)$. This can be written as $2 \cdot |C| + 1$ linear inequalities (with the introduction of $|C|$ additional variables representing the absolute values).

Next, we turn to the constraint given in (4). First, consider the case of the preference relations admitting an expected utility form. Let u_i be the utility function for individual i. By definition,

$$\pi(i) \succeq_i p^{i;j} \iff \sum_{c \in C} \pi(i)_c \cdot u_i(c) \geq \sum_{c \in C} p_c^{i;j} \cdot u_i(c).$$

Thus, for every $i \in \mathcal{X}$, the PIIF constraint given in (4) with respect to $j \in \mathcal{X}$ can be written as a linear inequality in the variables $p^{i;j}$ and $\pi(i)$.

Next, we consider the case of the stochastic domination preference relation. We introduce some notation as follows. Fix an individual i and their allocation, $\pi(i)$. Suppose $|C| = k$, and that the outcomes in C are labeled in decreasing order according to i's preferences: $u_i(c_0) \geq u_i(c_1) \geq \cdots \geq u_i(c_{k-1})$. With this ordering in place, we have that for any allocation $p \in \Delta(C)$ and every rank $r \in [k]$, $\mathbf{Pr}_{c \sim p}[u_i(c) \geq u_i(c_r)] = \sum_{t=1}^{r} p_t$. Thus, for each $i \in \mathcal{X}$ we can write the stochastic domination condition as k linear inequalities for each $j \in \mathcal{X}$, where

$$\pi(i) \succeq_i p^{i;j} \iff \forall r \in [k] : \sum_{t \in [r]} \pi(i)_t \geq \sum_{t \in [r]} p_t^{i;j}.$$

Importantly, this demonstrates that for this preference relation, the constraint given in (4) can be enforced using an additional $O(|C|)$ linear constraints, one for every $r \in [k]$. ◀

[4] We remark that this preference notion is a special case of the statistical concept of *first-order* stochastic domination [16, 5].

Other notions of preference

Theorem 16 focuses on the case in which individuals' preferences satisfy one of the two forms discussed above and formalized in Definitions 12 and 15. Naturally, however, not all preference relations satisfy one of these two forms. Appealing examples include preferences where the individual deems some of the outcomes to be substitutes (i.e., interested in exactly one) or complements (i.e., only interested in the complete set) or possibly preferences that value diversity of outcomes. We leave the question of whether PIIF admits efficient optimization over such non-convex preferences as an interesting direction for future research.

5 Fairness in Targeted Advertising: Multiple-Task PIIF

In this section, we extend the definition and study of preference-informed individual fairness to the *multiple-task* setting, formalized and studied by Dwork and Ilvento [13]. This setting was introduced as a model in which to study fairness in targeted advertising, a form of online advertising where ad platforms allow advertisers to specify the characteristics of users they would like to reach, and then make algorithmic decisions as to which users will see which ads based on the advertiser specifications, predictions of ad relevance to individuals, and the ad platform's revenue objectives. Targeted advertising has become pervasive and increasingly moderates individuals' exposure to opportunities. In recent years, numerous concerns have been raised about its fairness and discrimination implications, ranging from concerns about discriminatory advertiser targeting practices enabled by the platforms [3, 30, 2] to concerns about the ad delivery and allocation algorithms run by the platforms introducing bias where none was intended by the advertiser [11, 24, 1, 31, 29]. As part of a lawsuit settlement, the most prominent targeted advertising platform, Facebook, has begun to take steps to ensure advertisers cannot discriminate in their targeting practices [28]. However, the question of how to ensure that the ad delivery and allocation algorithms do not lead to discrimination is wide open [1, 25], in part due to lack of agreement over fairness definition(s) and ad platforms' concerns that existing definitions will restrict allocations in ways that significantly impact their revenue. As such, the multiple-task setting in the presence of individual preferences provides an important model to investigate formal guarantees of non-discrimination without being overly-restrictive for the decision-maker.

In the multiple task setting, we think of the set of outcomes C as arising from a collection of distinct tasks, e.g. deciding whether to show an ad for a user of each ad campaign $c \in C$. Importantly, in this setting, a separate fairness metric d_c is specified for each task (ad campaign), which naturally models real-world concerns in advertising, where different types of ads (e.g. housing, employment, product) are subject to different regulations and standards of fairness.

▶ **Definition 17** (Multiple-task IF). *An allocation $\pi : \mathcal{X} \to \Delta(C)$ is said to be $(D, \{d_1, \dots d_k\})$-*individually fair *in the multiple-task setting if for every two individuals $i, j \in \mathcal{X} \times \mathcal{X}$, the task-specific Lipschitz condition holds for each task:*

$$\forall c \in C : \quad D(\pi(i)_c, \pi(j)_c) \leq d_c(i, j).$$

In this setting, and particularly its application to ad delivery in the targeted advertising context, the benefits of a preference-informed approach to ensuring fairness become particularly salient. For instance, consider the following example, due to [13]. Suppose there are two ad campaigns, one for a high-paying tech job and another for childrens' toys. The ad-specific metrics capture the fact that differentiating based on a particular criteria could be permissible in some cases and not in others. For example, the metric associated with the tech ad should assign a small distance to individuals of similar qualifications regardless of their status as a parent, whereas the metric for toys might reasonably assign significant distance between parents and non-parents. However, under Multiple-task IF, a parent that is qualified for the tech job ad but is interested in toys must see the tech ad with the same probability as a qualified non-parent – an overly restrictive requirement. PIIF in the multiple-task setting addresses precisely this issue.

▶ **Definition 18** (Multiple-task PIIF). *An allocation $\pi : \mathcal{X} \to \Delta(C)$ satisfies $(D, \{d_1, \ldots d_k\}, \preceq$)-preference-informed individual fairness in the multiple-task setting if for all individuals $i \in \mathcal{X}$, for all other individuals $j \in \mathcal{X}$, there exists an allocation $p^{i;j} \in \Delta(C)$ such that:*

$$\forall c \in C : \quad D\left(p_c^{i;j}, \pi(j)_c\right) \leq d_c(i,j)$$
$$\pi(i) \succeq_i p^{i;j}$$

Again, in the multiple-task setting, the preference-informed extension of IF will require that for every individual $i \in \mathcal{X}$, when comparing to every other individual $j \in \mathcal{X}$, the individual i prefers their actual allocation to some alternative allocation, $p^{i;j}$. The main distinction is that now $p^{i;j}$ has to satisfy multiple-task IF with respect to j's current allocation.

Efficient optimization. Our results regarding efficient optimization subject to PIIF from the single-task setting (Section 4) directly extend to the multiple-task setting. In particular, given the ad-specific metrics, individuals' utilities and the advertisers' bids, the platform can efficiently compute the revenue- (or social welfare-) maximizing PIIF allocation.

An interesting direction for future work is relaxing the full information assumption. In particular, an online model, in which allocations are determined on a per-user basis, could naturally be more applicable, as well as allow for the preferences to be "discovered" through the allocation procedure (see [15] for a similar approach wrt the metric itself). This may necessitate investigation of non-trivial tradeoffs, as learning individuals' preferences requires some exploration, which may be at odds with ensuring fair treatment.

5.1 Fairness and social welfare in the multiple-task setting

The construction of Proposition 13 demonstrates that in the single-task setting, the gap between the best social welfare obtainable under IF and PIIF can be large even under very structured classes of preferences. This construction can be generalized to the multiple-task setting; however, an unconvincing aspect of it is the requirement that every individual is identical according to the metric. In such a setting, it's not surprising that IF is overly-constrained.

Here, we describe a family of instances in the multiple-task setting where the per-task similarity is *perfectly aligned with individuals' utilities*; that is, if two individuals benefit similarly from an outcome c, then they are similar. In such instances, we'd expect that the metric constraints would be perfectly aligned with social welfare. Still, we show that this intuition does not carry through for multiple-task IF: for a set of tasks, there are instances where the optimal social welfare under PIIF approaches a factor $|C|$ larger than the best IF solution.

▶ **Theorem 19.** *For any constant $\varepsilon > 0$, there is a sufficiently large $|\mathcal{X}|$ such that there exists a distribution of multiple-task instances where for each task $c \in C$, $d_c(i,j) = |u_i(c) - u_j(c)|$ and*

$$\frac{\mathcal{W}^*}{\mathcal{W}^*\left(\Pi^{\mathrm{IF}}\right)} \geq |C| - \varepsilon.$$

Intuition. Our proof is inspired by a construction of [13], which shows the impossibility of multiple-task IF under "naive composition." We begin by adapting their construction to our setting. Suppose there are two subpopulations of individuals $S \subseteq \mathcal{X}$ and $T = \mathcal{X} \setminus S$. We assume that each task-specific similarity metric d_c is determined by individuals' utility: $d_c(i,j) = |u_i(c) - u_j(c)|$. Additionally, suppose there are two ad campaigns c_0 and c_S. c_0 is a generic campaign where for all individuals $i \in \mathcal{X}$, $u_i(c_0) = 1$; thus, $d_{c_0}(i,j) = 0$ for all $i,j \in \mathcal{X} \times \mathcal{X}$. c_S is targeted where subpopulation S receives nontrivial utility, but the rest of the population receives no utility; thus, d_{c_S} treats pairs within $S \times S$ similarly, pairs from $T \times T$ similarly, but for $i,j \in S \times T$, is arbitrarily large, say $d_{c_S}(i,j) = 1$.

Given these campaigns, a natural allocation of ads to individuals, which we call $\pi^{\mathcal{W}}$, deterministically assigns $\pi^{\mathcal{W}}(i) = c_S$ to all individuals in $i \in S$ since they receive positive utility from c_S. Further, it assigns the untargeted campaign $\pi^{\mathcal{W}}(j) = c_0$ to individuals in $j \in T$ because they benefit positively from seeing c_0, whereas they get no benefit from c_S. Indeed, $\pi^{\mathcal{W}}$ maximizes the social welfare; everyone sees their favorite ad. But $\pi^{\mathcal{W}}$ violates multiple-task IF on c_0; that is, for $i,j \in S \times T$, $\left|\pi^{\mathcal{W}}(j)_{c_0} - \pi^{\mathcal{W}}(i)_{c_0}\right| = 1$ but $d_{c_0}(i,j) = 0$. Intuitively, under multiple-task IF, because everyone in $\mathcal{X}$ is similar according to c_0, the platform must decide whether it is more beneficial to show c_S to the individuals in S at the expense of not being able to show c_0 to the individuals outside of T. The proposition follows by extending this construction beyond the case of two campaigns and two subgroups, and carefully constructing utility functions for individuals.

Proof. Suppose $|C| = n$. Let $t \in \mathbb{N}$ be some constant. Suppose the universe of individuals $\mathcal{X} = S_0 \cup S_1 \cup \ldots S_{n-1}$ is partitioned into disjoint subpopulations ($S_\ell \cap S_m = \varnothing$ for $\ell \neq m$) for $|\mathcal{X}| \geq t^n$. The subpopulations will become progressively smaller as ℓ increases; for each $\ell > 0$, $\frac{|S_\ell|}{|\mathcal{X}|} = 1/t^\ell$ and let $\frac{|S_0|}{|\mathcal{X}|} = 1 - \sum_{\ell=1}^{n-1} \frac{|S_\ell|}{|\mathcal{X}|}$.

Notationally, for all $\ell \in [n]$, let $T_\ell = \bigcup_{m \geq \ell} S_m$. We construct individuals' utilities as follows: for each $\ell \in [n]$, for each individual $i \in \mathcal{X}$,

$$u_i(c_\ell) = \begin{cases} t^\ell & \text{if } i \in T^\ell \\ 0 & \text{otherwise} \end{cases}$$

First, note that for individuals $i \in S_\ell$, c_ℓ maximizes their utility $u_i(c_\ell) = t^\ell$. So consider the welfare-maximizing allocation that assigns every individual in S_ℓ to campaign c_ℓ; this allocation satisfies PIIF. The average social welfare can then be written as:

$$\begin{aligned} \frac{1}{|\mathcal{X}|} \cdot \sum_{\ell=0}^{n-1} \sum_{i \in S_\ell} u_i(c_\ell) &= \frac{|S_0|}{|\mathcal{X}|} + \sum_{\ell=1}^{n-1} \frac{|S_\ell|}{|\mathcal{X}|} \cdot t^\ell \\ &= \left(1 - \sum_{\ell=1}^{n-1} t^{-\ell}\right) + \sum_{\ell=1}^{n-1} t^{-\ell} \cdot t^\ell \\ &= n - \sum_{\ell=1}^{n-1} t^{-\ell} \end{aligned}$$

Next, consider similarity metrics defined by the utilities: For each task c_ℓ, we take $d_\ell(i,j) = |u_i(c_\ell) - u_j(c_\ell)|$. By the definition of u_i, under these similarity metrics, every pair of individuals $i,j \in T_\ell \times T_\ell$ are considered similar $d_\ell(i,j) = 0$. We fix D to be the total variation distance, in other words, $D(p_c, q_c) = |p_c - q_c|$. As such, any allocation π that satisfies IF must show c_ℓ to every individual $i \in T_\ell$ with some fixed probability $\alpha_\ell = \pi(i)_{c_\ell}$. We can compute the expression for the average social welfare of any such assignment as a function of the α_ℓ.

$$
\begin{aligned}
\frac{1}{|\mathcal{X}|} \cdot \sum_{\ell=0}^{n-1} \alpha_\ell \cdot \sum_{i \in T_\ell} u_i(c_\ell) &= \alpha_0 \cdot \frac{|S_0|}{|\mathcal{X}|} + \sum_{\ell=1}^{n-1} \alpha_\ell \cdot t^\ell \cdot \frac{|T_\ell|}{|\mathcal{X}|} \\
&= \alpha_0 \cdot \left(1 - \sum_{\ell=1}^{n-1} t^{-\ell}\right) + \sum_{\ell=1}^{n-1} \alpha_\ell \cdot t^\ell \cdot \sum_{m=\ell}^{n-1} t^{-m} \qquad (12) \\
&\leq \alpha_0 + \sum_{\ell=1}^{n-1} \alpha_\ell + \sum_{\ell=1}^{n-1} \alpha_\ell \cdot \sum_{m=\ell+1}^{n-1} t^{-m+\ell} \\
&\leq \left(\sum_{\ell=0}^{n-1} \alpha_\ell\right) \cdot \left(1 + \sum_{m=2}^{n-1} t^{-m}\right) \qquad (13) \\
&= 1 + \sum_{m=2}^{n-1} t^{-m} \qquad (14)
\end{aligned}
$$

where (12) follows by expanding $\frac{|T_\ell|}{|\mathcal{X}|}$ in terms of $\frac{|S_m|}{|\mathcal{X}|} = t^{-m}$; (13) applies Hölder's inequality; and (14) uses the fact that individuals $i \in S_{n-1}$ are members $i \in T_\ell$ for all $\ell \in [n]$, so the sum of the probabilities $\sum_\ell \alpha_\ell \leq 1$.

Given a desired ε, we can take t large enough, the ratio between the social welfares exceeds $n - \varepsilon$. ◀

Optimality under IF

Intuitively, this construction highlights the fact that allowing further targeting and more ad campaigns to participate allows the gap in social welfare between the best IF and PIIF solutions to grow considerably. Note that this gap applies even if the platform's objective is to optimize social welfare, so the proof also shows a gap in worst-case utility achievable by the decision-maker under IF.

We remark that a corollary of our result is that the Dwork-Ilvento "RandomizeThen-Classify" mechanism [13] for composition under multi-task IF achieves worst-case optimal performance (in terms of both social welfare and utility to the platform). In particular, [13] give an algorithm (in a setting with limited information modeling "competitive composition") that allocates a fixed distribution $p \in \Delta(C)$ to all individuals – thus, satisfying IF – that achieves a $1/|C|$-fraction of the best unconstrained utility. Our result shows that no IF solution, even with full information, can achieve a better fraction of the achievable utility.

6 Discussion

In this section, we review additional related work, note some possible extensions within the preference-informed fairness framework, and conclude with a discussion of the strengths and limitations of our current approach.

6.1 Further related works

Since [12], a number of recent works have aimed to extend the "fairness through awareness" framework, including [27, 23, 15, 22, 20]. These works focus on translating the theoretical IF framework into practically-motivated settings.

Three recent works have suggested incorporating notions of individuals' *preferences* into the fairness definitions. First, [4] present EF as an alternative to IF and study its learning-theoretic properties. Their focus is on the question of generalization: given a classifier that is envy-free on a sample, is it approximately envy-free on the underlying distribution? Their main technical result is a positive answer to this question, when learning from a particular structured family of classifiers. An interesting open question is whether the generalization results for IF from [27] and EF from [4] can be combined to give generalization for PIIF.

Second, [34] considers two notions of fairness at the (weaker) group level: treatment parity and impact parity. Their main contribution is a relaxation of both definitions, allowing for solutions where every protected group is "better off" *on average.* From a technical perspective, achieving their notion requires solving a non-convex optimization problem even in the simple case of linear classifiers for two disjoint groups. Our approach is different in that it focuses on defining both fairness and preferences at the *individual* level. This allows for a significantly stronger fairness guarantee, as well as a much more general framework that supports any notion of benefit or preference individuals may have. Importantly, our notion provably admits efficient optimization for a rich class of preference relations.

Finally, independent of our work, [9] study and quantify trade-offs between individual fairness and utility in an online version of the targeted advertising problem. [9] also observe that IF can come at a high cost to utility in the multiple-task setting of [13], but propose a different relaxation. Under their notion, every individual i chooses a subset of the outcomes $S_i \subseteq C$ and is guaranteed that their probability of seeing an ad from S_i is greater than the probability of every other individual of seeing an ad from S_i. Such a guarantee can be viewed as a variant of EF over a more restricted class of preferences than those we consider. The main distinction from PIIF is that this notion ignores the distance metrics entirely and in this sense resembles envy-freeness more than individual-fairness.

6.2 Preference-informed group fairness

In this work, our focus was on incorporating individual preferences into the metric-based individual fairness framework Dwork et al. [12]. The space of fairness definitions, however, is large, and different definitions may be more appropriate in different contexts.

A different approach for defining fairness, often referred to as "group fairness," proceeds as follows. A protected attribute, such as race or gender, induces a partition of the individuals into a small number of groups. For simplicity, we focus on the case where there is a single protected group, S, where the rest of the population is denoted $T = \mathcal{X} \setminus S$. A classifier is considered fair if it achieves parity of some statistical measure across these groups. Group fairness notions are typically weaker than individual notions of fairness: they only provide a guarantee for the "average" member of the protected groups and might allow blatant unfairness towards a single individual or even large subgroups; indeed, the shortcomings of group notions motivated the original work on "fairness through awareness" and subsequent works [12, 21, 17, 23]. Although group fairness notions can be fragile, they are widely studied and used due to their simplicity and due to the fact that they are easier to enforce and implement (for example, they do not require a task-specific similarity metric).

In principle, much of the reasoning behind our argument for incorporating preferences into IF [12] also extends to group-fairness notions. In this section, we show how we might augment a common group notion, called *Statistical Parity* (SP), to incorporate preferences. When there is a clearly "desirable" outcome, SP aims to protect the group S by guaranteeing equal average exposure to the desired outcome.

▶ **Definition 20** (Statistical parity). *A binary classifier $h : \mathcal{X} \to \{\pm 1\}$ satisfies (exact) statistical parity with respect to S if*

$$\Pr_{i \sim S}[h(i) = 1] = \Pr_{i \sim T}[h(i) = 1]$$

In our context, when individuals have diverse preferences over the outcome space, enforcing SP may again come at a cost to members of S, the group that SP aims to protect. As a concrete example, suppose everyone in $\mathcal{X}$ prefers the outcome $+1$, with the exception of some fraction of S, denoted S', who prefer the outcome -1. In this case, the statistical parity constraints prevents the solution where $h(i) = +1$ for $i \in X \setminus S'$ and $h(i) = -1$ for $i \in S$, which from the individuals' perspective is optimal.

Building on this intuition, we extend the set of classifiers we deem fair. Assuming every individual $i \in X$ has a preference relation over $\{\pm 1\}$ (or even distributions over $\{\pm 1\}$), *preference-informed statistical parity* (PISP) allows deviations from SP, as long as they are aligned with the individuals' preferences.

▶ **Definition 21** (Preference-informed statistical parity). *A binary classifier $h : X \to \{\pm 1\}$ satisfies preference-informed statistical parity with respect to S if there exists an alternative classifier, $h' : X \to \{\pm 1\}$, such that:*

$$\forall j \in T, h'(j) = h(j)$$
$$\forall i \in S, h(i) \succeq_i h'(i)$$
$$\Pr_{i \sim S}[h'(i) = 1] = \Pr_{i \sim T}[h'(i) = 1]$$

That is, fixing the outcomes members of T receive under h, every single member of S prefers their current outcome over what they would have received under a classifier satisfying statistical parity. Importantly, the guarantee is still with respect to the preferences of the *individual* members of S.

We conclude with several remarks regarding PISP. First, note that PISP only enriches the set of solutions that satisfy SP; any classifier that satisfies SP also satisfies PISP, by taking the alternative $h' = h$. The classifier welfare-maximizing classifier, where each individual is assigned their favorite outcome, is considered fair. For example, revisiting our example above, the classifier that gives $+1$ to $\mathcal{X} \setminus S'$, and -1 to S' is fair, because the alternative classifier that gives *everyone* $+1$ satisfies the PISP constraints. Finally, we argue that PISP maintains the core of the fairness guarantee of SP. For example, consider the classifier that assigns $+1$ to members of T and -1 to members of S. This classifier benefits the members of T in a way that is *not* aligned with the preferences of all members in S; rightfully, it does not satisfy PISP, because $i \in S \setminus S'$ are harmed.

6.3 Revisiting the assumptions underlying PIIF

Three main assumptions underlie our work. The first is that the outcome space is taken as a given. This could be problematic if the outcomes themselves are biased, e.g., tailored to the preferences of the majority, or worse yet, harmful to the minority. A biased outcome space would also present a problem for both IF and EF, which PIIF does not escape entirely. Still, PIIF may ameliorate the issue, by allowing the minority to receive outcomes that they prefer. We see a study of fairness of the outcomes themselves as an exciting direction for further inquiry.

The second assumption is that any deviation from an IF solution that is aligned with individuals' preferences should still be considered fair. This assumption follow from the perspective that "fair" allocation algorithms should protect the welfare of individuals; this perspective naturally extends the perspective underlying IF that similar individuals should be treated similarly. As discussed in [18, 19], in other (legal) settings, the notion of "fairness" may necessarily imply "treatment as equal," and notions of individual fairness may not apply. In such settings, the societal notion of fairness may require going against individual preferences. Handling such settings lies beyond the scope of our current work that focuses on the computer science notions of individual fairness, in which the objective is to provide strong protections from discrimination to the individuals themselves.

The third assumption is that the individuals' preferences are known. This is certainly the most nontrivial technical assumption we make; nevertheless, there are established techniques for learning utility-functions from observed behaviour [8]. We also note that accurately learning preferences could often be in the interest of the decision-maker (for example, ad platforms often claim that their implementations of targeted advertising are in line with users' interests [35]). Still, any practical estimation of the preferences runs the risk of injecting further bias during the learning process (for example, if the minority's preferences are estimated with lesser accuracy) and, therefore, mandates special attention in future research.

References

1 Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2019. arXiv:1904.02095.

2 Julia Angwin, Noam Scheiber, and Ariana Tobin. Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads. *ProPublica*, Dec 20, 2017. URL: https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men.

3 Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica*, Nov. 21, 2017. URL: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

4 Maria-Florina Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-Free Classification. *arXiv preprint*, 2018. arXiv:1809.08700.

5 Vijay S Bawa. Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, 2(1):95–121, 1975.

6 Katie Benner, Glenn Thrush, and Mike Isaac. Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says. *The New York Times*, Mar 28, 2019. URL: https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html.

7 Miranda Bogen. All the Ways Hiring Algorithms Can Introduce Bias. *Harvard Business Review*, May 6, 2019. URL: https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias.

8 Urszula Chajewska, Daphne Koller, and Dirk Ormoneit. Learning an agent's utility function by observing behavior. In *ICML*, pages 35–42, 2001.

9 Shuchi Chawla, Christina Ilvento, and Meena Jagadeesan. Individual Fairness in Sponsored Search Auctions. *arXiv preprint*, 2019. arXiv:1906.08732.

10 Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, Oct 10, 2018.

11 Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

12 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 (ITCS)*, pages 214–226. ACM, 2012.

13 Cynthia Dwork and Christina Ilvento. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 33:1–33:20, 2019.

14 Duncan K Foley. *Resource allocation and the public sector.* PhD thesis, Yale University, 1967.

15 Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online Learning with an Unknown Fairness Metric. *NeurIPS*, 2018.

16 Josef Hadar and William R Russell. Rules for ordering uncertain prospects. *The American economic review*, 59(1):25–34, 1969.

17 Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. *ICML*, 2018.

18 Deborah Hellman. Two concepts of discrimination. *Virgina Law Review*, 102:895, 2016.

19 Deborah Hellman. Measuring Algorithmic Fairness. *Virginia Law Review*, ssrn.3418528, 2019.

20 Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and Enforcing Subjective Individual Fairness. *arXiv*, 2019. `arXiv:1905.10660`.

21 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *ICML*, 2018.

22 Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Average Individual Fairness: Algorithms, Generalization and Experiments. *arXiv*, 2019. `arXiv:1905.10607`.

23 Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness Through Computationally-Bounded Awareness. *NeurIPS*, 2018.

24 Anja Lambrecht and Catherine E Tucker. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *SSRN*, ssrn.2852260, 2018.

25 Laura Murphy. Facebook's Civil Rights Audit – Progress Report, June 30, 2019. URL: `https://fbnewsroomus.files.wordpress.com/2019/06/civilrightaudit_final.pdf`.

26 Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2017.

27 Guy Rothblum and Gal Yona. Probably Approximately Metric-Fair Learning. In *ICML*, 2018.

28 Sheryl Sandberg. Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising. *Facebook Newsroom*, Mar 19, 2019. URL: `https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/`.

29 Ariana Tobin. HUD sues Facebook over housing discrimination and says the company's algorithms have made the problem worse. *ProPublica*, Mar 28, 2019. URL: `https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms`.

30 Ariana Tobin and Jeremy B. Merrill. Facebook Is Letting Job Advertisers Target Only Men. *ProPublica*, Sept 18, 2018. URL: `https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men`.

31 Upturn. Upturn Amicus Brief in Onuoha v. Facebook, Nov 16, 2018. URL: `https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf`.

32 Hal Varian. Efficiency, equity and envy. *Journal of Economic Theory*, 9:63–91, 1974.

33 John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition).* Princeton University Press, 2007.

34 Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

35 Mark Zuckerberg. The Facts About Facebook. *The Wall Street Journal*, Jan 24, 2019. Opinion in The Wall Street Journal `https://www.wsj.com/articles/the-facts-about-facebook-11548374613`.