Preference-Informed Fairness

Michael P. Kim* Aleksandra Korolova† Guy N. Rothblum‡ Gal Yona§

April 4, 2019

Abstract

As algorithms are increasingly used to make important decisions pertaining to individuals, algorithmic discrimination is becoming a prominent concern. The seminal work of Dwork *et al.* [ITCS 2012] introduced the notion of *individual fairness* (IF): given a task-specific similarity metric, every pair of similar individuals should receive similar outcomes. In this work, we study fairness when individuals have diverse preferences over the possible outcomes. We show that in such settings, individual fairness can be too restrictive: requiring individual fairness can lead to less-preferred outcomes for the very individuals that IF aims to protect (e.g. a protected minority group).

We introduce and study a new notion of *preference-informed* individual fairness (PIIF), a relaxation of individual fairness that allows for outcomes that deviate from IF, provided the deviations are in line with individuals' preferences. We show that PIIF can allow for solutions that are considerably more beneficial to individuals than the best IF solution. We further show how to efficiently optimize any convex objective over the outcomes subject to PIIF, for a rich class of individual preferences. Motivated by fairness concerns in targeted advertising, we apply this new fairness notion to the multiple-task setting introduced by Dwork and Ilvento [ITCS 2019]. We show that, in this setting too, PIIF can allow for considerably more beneficial solutions, and we extend our efficient optimization algorithm to this setting.

1 Introduction

Increasingly, algorithms are used to make consequential decisions about individuals. Examples range from deciding which advertisements and content users see online to automated loan and hiring decisions. In this work, we study such decision processes, which map individuals to a (potentially rich) set of outcomes. Automated decision-making comes with benefits, but it also

*Stanford University. 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. mpk@cs.stanford.edu

†University of Southern California. Part of this work completed while visiting the Weizmann Institute of Science. korolova@usc.edu

‡Weizmann Institute of Science. rothblum@alum.mit.edu. Research supported by the ISRAEL SCIENCE FOUNDATION (grant number 5219/17).

§Weizmann Institute of Science. gal.yona@weizmann.ac.il. Research supported by the ISRAEL SCIENCE FOUNDATION (grant number 5219/17).

raises substantial societal concerns (cf. [O'N17] 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. To address these concerns, 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.

The seminal work of [DHP+12] introduced the notion of *individual fairness* (IF). This notion 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 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, a divergence between their outcome distributions (say the statistical distance) 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 (such as credit scores and insurance risk scores), and in natural scenarios (a metric specified by an external regulator). Given an appropriate metric, individual fairness provides powerful protections from discrimination and prevents a "catalog of evils" (see [DHP+12]).

Accounting for individuals' preferences. Our work is motivated by settings in which individuals have diverse preferences over the possible outcomes. For example, different users might have very different preferences over which ads they would like to see. We note that although automated decisions are increasingly informed by rich knowledge about the individuals, the actual outcomes are not necessarily aligned with individuals' preferences. This could be because of the decision-maker's considerations: for example, an ad platform's need to increase revenue might lead it to show users less-preferred ads (say, because those advertisers are bidding aggressively).

One important conceptual contribution of individual fairness is distinguishing between outcome distributions that might make some (or even all) individuals unhappy (say, because they receive low benefit), from distributions that are discriminatory. For instance, when applying for a loan, an unqualified individual might be disappointed if their application is rejected; further, they might be even less happy when they see that a different qualified individual is approved for the same loan. In the eyes of the task-specific similarity metric, however, these two individuals are *different* – one is qualified, the other unqualified. Thus, individual fairness does not consider such an outcome discriminatory. Indeed, deciding to reject all applications (qualified and unqualified) might make no one happy, but it is perfectly fair since all individuals (similar or not) are treated similarly.

We argue, however, that when individuals have diverse preferences, IF can also be too restrictive: ignoring individuals' preferences can come at a high cost to *the very individuals that IF aims to protect* (e.g. a protected minority).

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. To simplify the example, we assume that there is a single metric for judging qualifications for software development roles across employers (e.g., the GPA in the CS major).[1] Consider candidates i, j and k, who are all similarly qualified.

[1] In a more realistic example, employers' preferences beyond this metric would take center stage. To prevent

Suppose there are three employers: X, Y and Z. When the candidates are polled for their preferences, i prefers $X > Y > Z$, j prefers $Y > Z > X$, and k prefers $Z > X > Y$ (these preferences can be due to rich and diverse factors, from geographic location to work-life balance considerations). Since i, j, k are all similarly qualified, IF requires that the candidates should receive similar interview distributions over $\{X, Y, Z\}$. Thus, IF *rules out the allocation where each candidate gets their most-preferred interview.* Is this "ruled out" outcome actually unfair?

This simplified example demonstrates that even when a suitable similarity metric is known, individual fairness can be too strict when individuals have diverse preferences over rich outcome spaces. Thus, we call for a study of fairness that accounts for individual preferences. Looking ahead, we note that these concerns extend to settings where we allow separate similarity metrics for each outcome. This is a natural setting for studying fairness in targeted advertising (different ads are subject to different fairness constraints), and was formalized and studied as the *multiple-task setting* in the work of Dwork and Ilvento [DI19]. We begin by considering the standard (single metric) setting, and defer discussion of the multiple-task setting to Section 1.3.

1.1 This Work: Preference-Informed Individual Fairness

Building on the perspective of individual fairness, 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.

Each individual is assigned an allocation (a distribution over outcomes). We assume that each individual has preferences over some or all possible allocations (we emphasize that we model preferences over distributions, not just over deterministic outcomes). PIIF establishes fairness by comparing each individual i's outcome distribution $\pi(i)$ to the outcome distribution $\pi(j)$ of every other individual j. The requirement is that i (weakly) prefers their current allocation $\pi(i)$ over an allocation that would have satisfied individual fairness with respect to $\pi(j)$. In particular, for $\pi(i)$ to be considered PIIF for i, we require that for each j there exists some alternative allocation that i could have received, denoted $\pi^j(i)$, that satisfies the individual fairness Lipschitz condition (see above) with respect to $\pi(j)$ **and** i (weakly) prefers their actual allocation to the individually-fair alternative $\pi^j(i)$. More formally, we define PIIF as follows:

Definition 1 (Preference-Informed Individual Fairness (PIIF), informal)**.** *An allocation π mapping each individual to a distribution over outcomes satisfies* ***preference-informed individual fairness*** *with respect to a divergence D, a similarity metric d, and individual preferences $\preceq$, if the following holds:*

For every pair of individuals i and j, whose outcome distributions are $\pi(i)$ and $\pi(j)$ (respectively), there exists another outcome distribution $\pi^j(i)$ such that:

discrimination, it can be quite important for the university to take an active role in the allocation. First, employers' preferences might be explicitly or implicitly discriminatory (e.g. [Das18]). Further, even if each employer's preferences are not discriminatory, Dwork and Ilvento [DI19] show that discriminatory outcomes can arise if the employers engage in unchecked competition for the students.

1. $\pi^j(i)$ *is individually fair w.r.t* $\pi(j)$: $D\left(\pi^j(i), \pi(j)\right) \leq d(i,j)$.

2. i *(weakly) prefers* $\pi(i)$ *over* $\pi^j(i)$: $\pi^j(i) \preceq_i \pi(i)$.

PIIF preserves the spirit of the core interpersonal fairness guarantee of IF: for each individual i, and 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 is that it allows for a much richer solution space*, which can lead to higher benefits (for individuals or for the decision maker, more on this below).

Referring back to the career expo example described above, the allocation where each candidate (deterministically) interviews with their preferred employer is PIIF. To see this, consider i comparing her outcomes with j's and k's under such an interview assignment. Comparing with j, i prefers outcome X to receiving outcome Y, which would satisfy the IF constraint w.r.t. j. Similarly, she prefers X to Z, which would satisfy the IF constraint w.r.t. 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.

We pause to discuss two caveats. First, we assume throughout that the individual's preferences are known. This is certainly a non-trivial assumption, but in many settings it can be in the interest of the decision-maker to know these preferences. For example, in targeted advertising, ad platforms often claim that the targeting is in line with users' interests (see Section 1.3). Second, we take the outcome space as a given. This can be problematic if the outcomes themselves are already biased (e.g. tailored to the preferences of the majority). This is a problem for individual fairness as well, and PIIF does not escape this issue (we note, though, that PIIF may alleviate the issue by allowing the minority to receive outcomes that they prefer). This calls for a further study of fairness of the outcomes themselves, which is beyond the scope of our work.

Preference-informed group fairness. While our primary focus throughout this work is on individual fairness, the space of fairness definitions is rich and an extensive literature studies so-called "group fairness" notions. We note that diverse individual preferences also arise naturally in the study of group fairness. The guiding principles behind preference-informed IF can also be applied towards relaxing group fairness notions such as statistical parity. This can, for example, unlock better outcomes for protected groups. See Section 5 for a detailed discussion.

Organization. We describe our main results in this introduction, with additional details in later sections. We discuss alternative approaches to defining fairness given individual preferences in Section 1.1.1. A discussion of the relationship between PIIF and IF is in Section 1.1.2. We give an overview on efficient optimization under PIIF in Section 1.2. We describe our results in the multiple-task setting in Section 1.3. We conclude with a further comparison to related work in Section 1.4. We elaborate on these contributions in Sections 2-4.

1.1.1 Fairness and preferences: other approaches and definitions

With the definition of PIIF in mind, we discuss alternative approaches. First, we note that two recent works have suggested incorporating individuals' preferences into fairness definitions. Balkan *et al.* [BDNP18] present *envy-freeness* (EF) as an alternative to individual fairness in the context of classification. Zafar *et al.* [ZVR$^+$17] consider fairness and preferences at the group level. We elaborate on these works in Section 1.4.

Envy-freeness. A prominent notion of fair division in the game-theoretic literature is envy-freeness [Var74, Fol67]. An allocation is envy-free (EF) if no individual i would prefer individual j's outcome to their own. We argue that this notion is more restrictive than what is necessary for ensuring non-discrimination in the settings that motivate our work (such as targeted advertising).

Returning to the loan example above, if i is not qualified for a loan but j is, then i might envy the fact that j received the loan, but not giving i the loan does not constitute discrimination.[2] In this work, we take the perspective that (given a suitable similarity metric) allocations that are individually fair provide strong protections from discrimination (even though they might not be envy-free). Armed with this perspective, our goal is *relaxing individual fairness to allow a richer set of solutions which still protect against discrimination*.

We note that PIIF is a relaxation of both individual fairness and envy-freeness: any IF allocation is PIIF, and the same is true for any envy-free allocation.

Incorporating preferences into the metric. An alternative approach to introducing a new fairness notion would be to incorporate preferences into the metric used by the IF definition. For example, we could modify the metric to consider individuals as "similar" only if their preferences are similar. However, when there is a rich set of possible outcomes, and a correspondingly rich set of possible preferences, it is not clear whether this is a sensible (or even feasible) approach.

Revisiting 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. Further, separating preferences from the metric also seems desirable from a conceptual perspective, as these two objects aim to capture very different aspects of the problem.

1.1.2 PIIF and Individual Fairness

As discussed above, PIIF aims to provide a flexible framework that relaxes IF. This allows for a richer set of solutions while simultaneously maintaining the core interpersonal discrimination-prevention guarantee of individual fairness. Unlocking a richer set of solutions can benefit the

[2] One might argue that an unqualified individual will not derive utility from a loan. Suppose, however, that getting the loan is a given, and an algorithm determines the interest rate. All applicants want lower rates, but we do not view offering lower rates to more qualified applicants as discrimination (assuming qualifications are determined in an unbiased manner).

individuals or other parties. For example, an ad platform may aim to maximize its revenue while ensuring non-discrimination. Enlarging the space of fair allocations may raise the platform's revenue.

To formalize this intuition, we consider *social welfare*, the sum of all individuals' benefits from their outcomes, as an aggregate measure of benefit to the individuals. When there are k possible outcomes, the social welfare achieved by the best IF allocation can be smaller by a factor of k than the social welfare achieved by the best PIIF allocation (this ratio is tight). The details and a discussion are in Section 3.3.

With the above "worst-case" gap in mind, we note that individuals' benefit (or lack thereof) from the richer solution space depends on *which of the solutions is chosen*. If a platform ignores user preferences, relaxing the definition might not improve user outcomes. Even worse, if the platform maximizes an objective that is anti-correlated with individuals' preferences, then enlarging the set of solutions might lower the benefit to individuals, see Section 3.5 for an example. Indeed, the principle behind PIIF is that the fairness notion should not *restrict* the solution space in a way that is harmful to individuals that it aims to protect. It does not guarantee beneficial outcomes (similarly to IF, outcomes where all individuals are unhappy are considered fair).

See Section 3.1 for a more detailed discussion on the relationship between PIIF and other fairness notions.

1.2 Efficient Optimization

Similarly to the work of Dwork et al. [DHP+12], a central question in the study of preference-informed fairness is finding a fair allocation that minimizes a (convex) target function. One of our main contributions is an efficient optimization algorithm for a rich class of individual 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 outcome space, for which PIIF requires preferences to be defined, is exponential. In realistic settings, where the number of individuals or outcomes is typically large, this soon becomes intractable. Towards efficient optimization, we focus on two rich and structured preference classes:

1. *Preferences that admit an expected utility function representation.* These are preferences for which there exists a way of assigning numerical values – utilities – to each (deterministic) outcome, such that an individual's benefit from a probabilistic allocation is simply their expected utility.

2. *Stochastic domination.* Assuming individuals have a utility per deterministic outcome, stochastic domination formalizes the following intuition: for any distribution over outcomes, only a shift of probability mass from less desirable outcomes to more desirable outcomes is considered preferable. For a concrete example, recall the interview candidate i (whose preference over deterministic outcomes is $X > Y > Z$). Stochastic domination would mean that i would prefer X with certainty over receiving X, Y each with probability 0.5. Note that this is a non-total relation (two distributions may be incomparable) and that it defines a stronger notion of preference than (1).

Theorem 1.1 (Efficient optimization subject to PIIF, informal)**.** *If every individual's preference relation comes from one of the two classes defined above, the PIIF constraints can be written as a set of (polynomially many) linear inequalities.*

This provably linear structure of the PIIF constraints implies an efficient algorithm for minimizing any convex target function subject to preference-informed individual fairness, for a rich class of individual preferences. A useful aspect of this formulation is the automatic support for any additional linear (generally, convex) constraints that the allocations should satisfy. In the career expo example, this could be a limit on the number of interview slots per employer and per candidate, and in the ad platform example, natural constraints are the budgets of the advertisers and the number of ads that can be shown to each user.

See Section 3.2 for full details on these definitions and results.

1.3 Targeted Advertising and The Multiple-Task Setting

We proceed to discuss fairness in the *multiple-task* setting [DI19]. This setting assumes a large set of outcomes (which we refer to as tasks), with similarity modeled using a separate metric for each outcome or task. This is a natural setting for studying fairness in online advertising, where different ads should be subject to different fairness constraints.[3]

In recent years, the use of targeted advertising – where ad campaigns specify targeted audiences through criteria such as age, gender, and ethnicity – has become pervasive, and issues of non-discrimination are increasingly becoming a concern. The dominant rhetoric by ad platforms is that targeted advertising benefits everyone [Zuc19]: the individuals – because they see more relevant ads, the advertisers – because they can reach audiences likely to like their product and therefore, spend their money more effectively, and the platforms – because they can earn money while working towards the benefit of both users and the advertisers. Questions of whether such advertising systems and the ad allocation algorithms run by them enable discrimination [ATV17, TM18b, AST17, TM18a, DTD15, LT18, ASB+19] and whether they actually improve individuals' outcomes [TH19, Hei12] increasingly have made their way into public discourse [Swe13, PRMT17] and are facing legal scrutiny [BTI19, Upt18]. As such, targeted advertising serves as a natural application where it is important to provide guarantees of non-discrimination, in the presence of individual preferences.

Multiple-task IF. Dwork and Ilvento [DI19] formalized individual fairness in the multiple-task setting as the requirement that task-specific individual fairness should hold, simultaneously and separately, for every task. For example, if two individuals are similar with respect to a job posting, they should see this ad with similar probability. Multiple-task IF will enforce similar treatment for these two individuals even if they differ in their preferences over the other ads in the system. At the extreme, it may "block" the solution in which every individual sees their favourite ad.

[3] For instance, consider the following example, due to [DI19]. Suppose there are two ad campaigns, one for a high-paying tech job and another for children's toys. The similarity metric associated with the tech ad should assign a small distance to parents and non-parents of similar qualifications, whereas the metric for toys might reasonably assign significant distance between parents and non-parents; that is, in one task (marketing children's products), it may be appropriate to differentiate based on whether someone is a parent, whereas in the other task (employment opportunities), such differentiation should not be permissible.

Multiple-task PIIF. Targeted advertising is a domain where individuals naturally have preferences that are both rich and diverse. The discussion above renders it a particularly appealing test-bed for *preference-informed* fairness. To extend the definition of PIIF to the multiple-task setting, we follow the same guiding principle as in Section 1.1. We 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, $\pi^j(i)$. The main distinction is that now $\pi^j(i)$ has to satisfy *multiple-task* IF with respect to j's current allocation. The formal definition is in Section 4.

Intuitively, PIIF will in fact allow individuals that are similar with respect to an ad to see it with a different probability – so long as this happens for a "good cause": e.g. to enable the individual who sees it with a lower probability to instead see some other ad they *prefer* with a higher probability. In particular, the solution in which every user sees their favourite ad satisfies this definition (even though it may not be obtainable in practice, due to other constraints in the advertising system).

We proceed with an overview of our results for the multiple-task setting. A full description can be found in Section 4.

1.3.1 Efficient optimization in the multiple-task setting

Similarly to the single task setting results (see Section 1.2), we consider optimizing convex objectives over the allocation, subject to the multi-task PIIF constraints. Here too, the optimization can be efficient so long as individuals' preferences come from one of the two classes described in Section 1.2. As was the case there, the fairness constraints can be specified as linear equations, and this formulation can support any additional linear (generally, convex) constraints.

1.3.2 Social Welfare under IF and PIIF

The main motivation behind the definition of PIIF is that multiple-task IF may come at a high cost not only to the platform, but also to the individuals that IF aims to protect. We make this argument formal by constructing a family of instances for which requiring multiple-task IF prevents most individuals from receiving a preferred allocation. Since PIIF always considers the solution in which an individual receives their favourite allocation as fair, this implies a provable (worst-case) gap between the best social welfare that can be achieved under IF and under PIIF.

Overview of construction. The construction is inspired by a construction of [DI19] that shows the impossibility of individual fairness under naive composition. For intuition, we begin by recapping their construction, adapted 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: $d_c(i,j) = |u_i(c) - u_j(c)|$. That is, if two individuals derive similar utility from a positive outcome on the task, they are considered similar; intuitively, we'd expect that this metric is perfectly aligned with social welfare.

The construction involves two campaigns c_0 and c_S. c_0 will be a generic campaign where for all individuals $i \in \mathcal{X}$, $u_i(c_0) = 1$; we take $d_{c_0}(i,j) = 0$ for all $i,j \in \mathcal{X} \times \mathcal{X}$. c_S will be targeted where subpopulation S receives nontrivial utility, but the rest of the population receives no utility; specifically, we take $u_i(c_S) = 10$ for all $i \in S$ and $u_j(c_S) = 0$ otherwise ($j \in T$). By these utility

values, we take d_{c_S} to treat pairs within $S \times S$ similarly, pairs from $T \times T$ similarly, but for $i, j \in S \times T$, we let $d_{c_S}(i, j) = 1$ (arbitrarily large).

Given these campaigns, a natural allocation of ads to individuals, which we call $\pi^{\mathcal{W}}$, would deterministically assign $\pi^{\mathcal{W}}(i) = c_S$ to all individuals in $i \in S$; they each receive utility $u_i(c_S) = 10$ from this ad. Further, it makes sense to show the untargeted $\pi^{\mathcal{W}}(j) = c_0$ to individuals in $j \in T$ because they benefit positively from seeing c_0, whereas they get no benefit from c_S. Indeed, such an allocation maximizes the social welfare; everyone sees their favorite ad. But note that $\pi^{\mathcal{W}}$ violates IF on c_0; in particular, pairs $i, j \in S \times T$ are similar according to d_{c_0} but receive c_0 with different probabilities; that is,

$$\pi^{\mathcal{W}}(j)_{c_0} - \pi^{\mathcal{W}}(i)_{c_0} = 1 > d_{c_0}(i, j) = 0.$$

Suppose the platform is required to allocate ads in accordance with multiple-task IF; the above observation suggests that if anyone sees c_0, then everyone must see c_0 with the same probability. Intuitively, because everyone in $\mathcal{X}$ is similar according to c_0, under IF constraints 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.

Now suppose that $|S| = \frac{1}{10} \cdot |\mathcal{X}|$. Then, given the utilities defined above, the social welfare from showing c_0 deterministically to all individuals in $\mathcal{X}$ is equal to the social welfare of showing c_S to individuals in S and showing no ads to T; in both cases, the average social welfare is 1. Note, however, that the average social welfare of $\pi^{\mathcal{W}}$ is 1.9: $\frac{1}{10} \cdot 10$ from the members of S and $\frac{9}{10} \cdot 1$ from the members of T.

The theorem follows by extending this construction beyond the case of two campaigns and two subgroups. By allowing further targeting and more ad campaigns to participate, the gap in social welfare can grow considerably, proportional to the number of targeted campaigns in the system. The full construction and proof is given in Section 4.4.

1.4 Further Related Work

Two recent works have recently suggested incorporating notions of individuals' *preferences* into the fairness definitions. [BDNP18] present *envy-freeness* (EF) as an alternative to the metric-based individual fairness notion of [DHP+12] 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.

Another work that considers preferences in the context of fairness is [ZVR+17]. They consider two notions of fairness at the group level: treatment parity and impact parity. Their main contribution is a relaxation of both definitions, which allows for any solution in which every protected group is "better off" *on average* (in terms of the fraction of positive predictions the group members receives) to also be considered fair. 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 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.

2 Preliminaries

In this work, we consider allocation problems; given a set of individuals $\mathcal{X}$, our task is to assign every individual to an outcome in the set $\mathcal{C}$. We allow randomized allocation rules $\pi : \mathcal{X} \to \Delta(\mathcal{C})$, where for each individual $i \in \mathcal{X}$, their allocation $\pi(i) \in \Delta(\mathcal{C})$ represents a distribution over outcomes $c \in \mathcal{C}$. Typically, we will study how to select an allocation rule π to maximize a linear objective subject to fairness constraints.

2.1 Individuals' preferences

An important part of our framework is the modeling of individuals' *preferences* over the outcome space. We do so by assuming that every individual $i \in \mathcal{X}$ has a reflexive and transitive binary relation $\preceq_i$ that encodes their preferences over allocations in $\Delta(\mathcal{C})$; for $p, q \in \Delta(\mathcal{C})$, we use $p \preceq_i q$ to denote that i (weakly) prefers q to p.[4] We use $\preceq$ to denote the set of individuals' preference relations, $\preceq \triangleq \{\preceq_i\}_{i \in \mathcal{X}}$.

In much of our discussion, we focus specifically on preference relations that admit a *utility* function $u_i : \Delta(\mathcal{C}) \to \mathbb{R}$ for each individual $i \in \mathcal{X}$, where $u_i(\pi(i))$ represents the utility to individual i from their allocation given by π. In particular, in such a setting, $p \preceq_i q$ if and only if $u_i(p) \leq u_i(q)$.

The *social welfare* of an allocation π is the sum of the individuals' expected utilities under π.

$$\mathcal{W}(\pi) = \sum_{i \in \mathcal{X}} u_i(\pi(i)) \tag{1}$$

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 some allocation in Π.

2.2 Envy-freeness

Given individual preferences, a classic notion of fairness is *envy-freeness* (EF). Originally studied in the context of fair division, EF remains a prominent approach to fairness in problems of goods allocation (e.g. cake-cutting [RW98], rent-division [Su99]). In our context, envy-freeness captures the idea that every individual values their allocation at least as much as any other individual's allocation; that is, no one envies anyone else's allocation.

Definition 2. *An allocation $\pi : \mathcal{X} \to \Delta(\mathcal{C})$ is envy-free with respect to individuals' preferences $\preceq$ if for all individuals $i \in \mathcal{X}$, and for all other individuals $j \in \mathcal{X}$,*

$$\pi(j) \preceq_i \pi(i) \tag{2}$$

Very recently, [BDNP18] also considered EF in the context of classification. [BDNP18] motivate their study of EF in this context by arguing that individuals' utilities may be easier to estimate

[4] We remark that $\preceq_i$ need not be *total* nor *antisymmetric* over $\Delta(\mathcal{C})$. That is, according to individual i, two allocations could be incomparable under $\preceq_i$; separately, two allocations $p \neq q$ could be equally preferable (i.e., $p \preceq_i q$ and $q \preceq_i p$).

than the similarity metric required by [DHP+12]. Important to our study, unlike IF, the solution that maximizes social welfare is feasible under EF. Still, in the context of classification, EF may not always be an appropriate fairness requirement. For example, if everyone's preferences are identical (e.g. a binary task in which one outcome is clearly more desirable), the constraint in (2) requires that all individuals receive the same distribution over outcomes, rendering the output of any such classifier useless. See Section 1.1.1 for a further discussion.

2.3 Individual fairness

[DHP+12] introduced a notion of *individual fairness* (IF) for classification tasks. At its heart is a task-specific similarity metric that specifies, for every two individuals, how similar they are with respect to the specific task at hand. Given such a metric, similar individuals should be treated similarly, i.e., assigned similar allocations.

Definition 3 (Individual fairness)**.** *A (probabilistic) allocation $\pi : \mathcal{X} \to \Delta(\mathcal{C})$ is said to be (D, d)-*individually fair *if for every two individuals $i, j \in \mathcal{X} \times \mathcal{X}$, the following Lipschitz condition holds:*

$$D(\pi(i), \pi(j)) \leq d(i, j) \tag{3}$$

In this definition, $D : \Delta(\mathcal{C}) \times \Delta(\mathcal{C}) \to \mathbb{R}^+$ is a divergence that captures similarity between allocations and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is the task-specific similarity metric that captures similarity between individuals. Like [DHP+12], we assume that both D and d are appropriately normalized so that direct comparison is meaningful; we will assume that the distances are normalized to $[0, 1]$.

3 Preference-Informed Individual Fairness

Given an appropriate metric, individual fairness provides powerful protections from discrimination. In settings where individuals have diverse preferences over the possible outcomes, however, we have argued that ignoring individuals' preferences can come at cost to the very individuals that it aims to protect. In this section, we propose and study the notion of *preference-informed individual fairness* (PIIF). Unlike EF, it makes sense even in a context where individuals' preferences are aligned with one another; we elaborate on the relationship between PIIF and individual fairness and envy-freeness in Section 3.1.

The definition of PIIF, given below, formalizes the idea that allocations that deviate from individually-fair solutions may be considered fair, provided the deviations are aligned with individuals' preferences:

Definition 4 (Preference-Informed Individual Fairness)**.** *An allocation $\pi : \mathcal{X} \to \Delta(\mathcal{C})$ satisfies $(D, d, \preceq)$-**preference-informed individual fairness** if for all individuals $i \in \mathcal{X}$, for all other individuals $j \in \mathcal{X}$, there exists another allocation $\pi^j(i) \in \Delta(\mathcal{C})$ such that:*

$$D\left(\pi^j(i), \pi(j)\right) \leq d(i, j) \tag{4}$$

$$\pi^j(i) \preceq_i \pi(i) \tag{5}$$

Importantly, we argue that PIIF maintains the core of the interpersonal fairness guarantee of IF; see Section 1.1 for a full treatment, as well as an intuitive interpretation of the above definition.

3.1 Relations of PIIF to other fairness criteria

In this section, we take a closer look at the relationship between PIIF and two other fairness notions: individual fairness (IF) and envy-freeness (EF). We argue that PIIF captures the appealing properties of *both* these concepts.

The following proposition demonstrates that PIIF is a relaxation of both individual fairness and envy-freeness.

Proposition 3.1. *Fixing a divergence, the metric and preferences, let* $\Pi^{IF}, \Pi^{EF}, \Pi^{PIIF}$ *denote the set of IF, EF, and PIIF solutions, respectively. Then,* $\Pi^{IF} \subseteq \Pi^{PIIF}$ *and* $\Pi^{EF} \subseteq \Pi^{PIIF}$.

Proof. First, consider an allocation $\pi \in \Pi^{\text{IF}}$. From the perspective of any individual $i \in \mathcal{X}$, when comparing to individual $j \in \mathcal{X}$, suppose we take $\pi^j(i) = \pi(i)$; then, by the fact that π satisfies IF, condition (4) is satisfied. By reflexivity of $\preceq_i$, (5) is also satisfied, so $\pi \in \Pi^{\text{PIIF}}$.

Next, consider an allocation $\pi \in \Pi^{\text{EF}}$. From the perspective of any individual $i \in \mathcal{X}$, when comparing to individual $j \in \mathcal{X}$, suppose we take $\pi^j(i) = \pi(j)$; then, condition (4) is satisfied trivially because $D(\pi(j), \pi(j)) = 0$. By the fact that π satisfies EF, we know that $\pi(j) \preceq_i \pi(i)$, so condition (5) also holds; thus $\pi \in \Pi^{\text{PIIF}}$. □

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 $\pi^j(i) = \pi(i)$, 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 $\pi^j(i) = \pi(j)$. 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 maximizes social welfare by giving everyone their most-preferred outcome. Thus, both these extreme solutions will also be feasible for PIIF. In general, PIIF will be a meaningful, 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.

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 specific cases of PIIF. To implement IF, we can set all individual's preference relation $\preceq_i$ to be the trivial reflexive relation, where for all $p \in \Delta(\mathcal{C})$ $p \preceq_i p$ and all other pairs are incomparable. To implement EF, we simply take $d(i, j) = 0$ for all $i, j \in \mathcal{X} \times \mathcal{X}$.

3.2 Optimization subject to PIIF

In the case of IF, finding *some* individually fair solution is simple (e.g., treating everyone identically). Similarly to the work of Dwork et al. [DHP+12], a central question in the study of preference-informed fairness is efficiently finding a fair allocation that minimizes a target function:

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

In this section, we answer the question of feasibility of efficient optimization in the positive – as in [DHP+12] – when $f(\cdot)$ is convex, and for two classes of more structured preference relations. We define these below.

3.2.1 Preferences admitting an expected utility representation

The first set of preferences are those that admit an *expected utility representation*. This is a standard approach in economics for modeling decision-making in the presence of uncertainty.

Definition 5 (Expected utility representation)**.** *A preference relation $\preceq$ admits an expected utility representation if and only if there exists a function $u : \mathcal{C} \to \mathbb{R}^+$, such that for any two allocations $p, q \in \Delta(C)$,*

$$p \succeq q \iff \sum_{c\in C} p_c \cdot u(c) \geq \sum_{c\in C} q_c \cdot u(c) \tag{6}$$

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 [VNM07] provide a complete characterization of this class of preference relations. For completeness and accessibility, we include a discussion of these results in Appendix A.

3.2.2 Stochastic domination

Stochastic domination gives an example of a non-total preference relation that seems of broad interest.

Definition 6 (Stochastic domination)**.** *For an individual with a utility function $u : \mathcal{C} \to \mathbb{R}^+$ and for any two allocations $p, q \in \Delta(C)$, p stochastically dominates q if*

$$p \succeq q \iff \forall 0 \leq x \leq M, \quad \mathop{\mathbf{Pr}}_{c\sim p}[u(c) \geq x] \geq \mathop{\mathbf{Pr}}_{c\sim q}[u(c) \geq x] \tag{7}$$

where $M = \max_{c\in\mathcal{C}} u(c)$.

That is, an allocation p is (weakly) preferred over q if for every possible level of "benefit" x, the probability of achieving at least x is no worse under p than it is under q. Intuitively, this preference relation captures the fact that any shift in probability mass towards more desired alternatives (i.e., higher benefit according to u) is preferred. This intuition will be made formal in the proof of Theorem 3.2. Finally, we remark that this preference notion is a special case of the statistical concept of *first-order* stochastic domination [HR69, Baw75].

3.2.3 Efficient optimization subject to PIIF constraints

We are now prepared to 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 3.2. *Let $\{\preceq_i\}_{i=1}^n$ be the set of individuals' preferences. If every $\preceq_i$ is either the stochastic domination relation or admits an expected utility representation, then the set of all $(D_{tv}, d, \preceq)$-preference-informed-individually-fair allocation forms a convex polytope in $\mathbb{R}^k$, where $k = poly\left(|\mathcal{X}|, |\mathcal{C}|\right)$.*

Proof. We specify the PIIF constraints using the following variables: for all $i \in \mathcal{X}$, let $\pi(i) \in \Delta(\mathcal{C})$ be a vector denoting the actual allocation; for every pair of individuals $(i, j) \in \mathcal{X} \times \mathcal{X}$, let $\pi^j(i) \in \Delta(\mathcal{C})$ be a vector denoting the alternative allocation for i when comparing to j. We argue that the PIIF constraints given in (4) and (5) can each be written as linear inequalities over these variables.

First, since D is taken to be the total variation distance we can translate (4) as

$$\frac{1}{2} \cdot \sum_{c \in \mathcal{C}} \left|\pi^j(i)_c - \pi(j)_c\right| \leq d(i, j).$$

This constraint can be written as $2 \cdot |\mathcal{C}| + 1$ linear inequalities (with the introduction of $|\mathcal{C}|$ additional variables representing the absolute values).

Next, we turn to the constraint given in (5). First, consider the case of the preference relations admitting an expected utility form. Let u_i be the Bernoulli utility function (Appendix A) for individual i. Then by definition, we have:

$$\pi^j(i) \preceq_i \pi(i) \iff \sum_{c \in C} \pi^j(i)_c \cdot u_i(c) \leq \sum_{c \in C} \pi(i)_c \cdot u_i(c) \tag{8}$$

Thus, the constraint given in (5) is translated to the constraint $\sum_{c \in C} \pi^j(i)_c \cdot u_i(c) \leq \sum_{c \in C} \pi(i)_c \cdot u_i(c)$, which is indeed linear in the variables $\pi^j(i)$ and $\pi(i)$, for every $i, j \in \mathcal{X}$.

Next, we consider the case of the stochastic domination preference relation. We introduce some notation as follows. Fix an individual i and their allocation, $\pi(i)$. Re-order the outcomes in $\mathcal{C}$ in decreasing order according to i's preferences:

$$u_i(c_1) \geq u_i(c_2) \geq \cdots \geq u_i(c_k)$$

Let $r \in [1, k]$. We denote i's utility from the ad they rank r as $u_{i,r}$. We use $p_r^{\pi(i),i}$ to denote the probability (under $\pi(i)$) that i receives the outcome ranked r. For brevity, when the allocation π and individual i are clear from the context, we drop the superscript notation. We use $p_1, \ldots, p_k$ for an allocation π and $p'_1, \ldots, p'_k$ for a second allocation π'. For example, p_1 is the probability under π in which i receives their favorite outcome, and $u_{i,1}$ is their utility from it. Finally, note that $\sum_{r=1}^k p_r = 1$.

By definition, we have that for every rank $r \in [1, k]$,

$$\mathop{\mathbf{Pr}}_{c \sim \pi(i)} [u_i(c) \geq u_{i,r}] \equiv \sum_{t=1}^{r} p_t \tag{9}$$

This implies the following:

$$\pi(i) \succeq_i \pi'(i) \iff \forall r \in [1, k] : \sum_{t=1}^{r} p_t \geq \sum_{t=1}^{r} p'_t \tag{10}$$

Importantly, this demonstrates that for this preference relation, the constraint given in (5) can be enforced using an additional $O(|C|)$ linear constraints, one for every $r \in [1, k]$.

□

3.2.4 Other notions of preference

Theorem 3.2 focuses on the case in which individuals' preferences satisfy one of the two forms discussed above and formalized in Definitions 5 and 6. Naturally, however, not all preference relations satisfy one of these two forms. An immediate example are preferences in which an individual deems some of the outcomes as either substitutes (they are interested in *exactly* one) or complements (they are only interested in *both*). Another appealing example could be individuals with a preference towards *diversity*. These individuals may not want *all* the probability mass put on their most preferred outcome. We leave the question of whether PIIF admits efficient optimization in the presence of these natural types of *non-convex* preferences as an interesting future research direction.

3.3 Fairness and social welfare

In this section, we quantify the extent to which IF and PIIF solutions can differ in social welfare. First, we formalize the fact that IF can constrain social welfare significantly. In particular, we show a family of instances in both the single-task and multiple-task settings (Section 4.4) where given k possible outcomes, the best social welfare under IF is a factor k worse than the best social welfare possible. Because PIIF permits the allocation that maximizes social welfare, this shows a factor-k gap between the social welfare achievable under IF and under PIIF in the worst case.

3.4 Individual fairness may restrict social welfare

We begin by focusing on the single-task setting. Intuitively, social welfare can be hurt significantly in the single-task setting, when many individuals are considered similar, but there is a diversity of preferences over outcomes. Formalizing this intuition, we can show that given k possible outcomes, the gap between the social welfare achievable under IF and under PIIF can be a factor k.

Proposition 3.3. *Suppose $|\mathcal{C}| = k$; for any $|\mathcal{X}| = nk$ for $n \in \mathbb{N}$, there exists a distribution of allocation problem instances such that*

$$\mathcal{W}^* \left(\Pi^{IF}\right) \leq \frac{1}{k} \cdot \mathcal{W}^* \left(\Pi^{PIIF}\right).$$

The construction is simple; a sketch follows. Consider a setting where every individual is considered similar; 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. Suppose that every individual $i \in \mathcal{X}$ is assigned a preferred outcome such that a $1/k$-fraction of individuals prefer each outcome c; let $u_i(c) = 1$ for their preferred outcome and $u_i(c) = 0$, otherwise. The best social welfare achievable assigns everyone to their preferred outcome; under this allocation the expected social welfare is 1. But under any identical allocation, the expected social welfare is $1/k$. It's not hard to see that for additive utilities, this gap is tight; in this setting, the best fixed allocation always achieves at least a $1/k$-fraction of the total social welfare.

3.5 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. With this in mind, it is tempting to hope that when the platform optimizes its utility over Π^{PIIF}, the social welfare will be at least as high as when it optimizes over Π^{IF}. We show that, in general, this hope is misplaced. In particular, we give a construction where the platform's optimal solution under IF (for a carefully-designed objective function) has better social welfare than the platform's optimal PIIF solution (for the same objective). In fact, the optimal PIIF solution under the platform's objective will also satisfy EF; thus, any definition that encompasses envy-freeness will also exhibit such a phenomenon.

These constructions demonstrate that PIIF on its own does not provide any guarantees about social welfare. In particular, if the platform chooses their allocation as the solution to a constrained optimization, there are objectives under which the social welfare of IF exceeds that of EF, and vice versa. An appealing aspect of the PIIF framework, however, is that it disentangles the goals of fairness and social welfare. At the end of this section, we discuss how to augment the PIIF constraints with constraints on the social welfare, in a way that provides guarantees on the resulting social welfare.

Next, we give a construction of instances where given the constraints and a specially crafted objective f, the solution that optimizes f under PIIF has significantly worse social welfare than the corresponding IF solution; in particular, utility according to f will be directly opposed to social welfare. This shows that by expanding the solution space with PIIF to include solutions of better social welfare, we may also admit solutions with worse social welfare compared to the corresponding IF solution. Note that the optimal solution under PIIF will also satisfy EF; thus, the construction also shows a setting where the social welfare under IF exceeds that of EF.

Proposition 3.4. *Suppose $|\mathcal{C}| = k$ and $|\mathcal{X}| = n(k-1)$ for any $n \in \mathbb{N}$. Let f denote the platform's utility function; let $\pi^{IF} = \operatorname{argmax}_{\pi \in \Pi^{IF}} f(\pi)$ and $\pi^{PIIF} = \operatorname{argmax}_{\pi \in \Pi^{PIIF}} f(\pi)$. There exists a distribution of allocation problem instances such that*

$$\mathcal{W}\left(\pi^{IF}\right) \geq T \cdot \mathcal{W}\left(\pi^{PIIF}\right)$$

for any constant $T > 0$.

Proof. We give an overview of the construction in the single-task setting with outcome space $\mathcal{C}$ where $|\mathcal{C}| = k$. Consider an instance where all individuals are similar; for all $i, j \in \mathcal{X} \times \mathcal{X}$, $d(i, j) = 0$. We assume the platform wishes to optimize a function $f(\pi) = \mathbf{E}_{i \sim \mathcal{X}} f_i(\pi(i))$ defined as follows. Suppose for a special outcome $c = 0$, $f_i(c) = \frac{1}{k-1} + \varepsilon$ for some small constant $\varepsilon > 0$, for all $i \in \mathcal{X}$. Then, we partition $\mathcal{X}$ into $k - 1$ equal-sized groups of individuals $\{\mathcal{X}_c : c \in \mathcal{C} \setminus \{0\}\}$; suppose for all $c \in \mathcal{C} \setminus \{0\}$, $f_i(c) = 1$ for each individual $i \in \mathcal{X}_c$ and $f_i(c) = 0$ for $i \notin \mathcal{X}_c$. Suppose that every individual gets utility 1 from seeing $c = 0$; further, suppose individuals $i \in \mathcal{X}_c$ receive utility $1/T > 0$ for seeing c; otherwise $u(\pi(i)) = 0$.

We start by considering the allocation that optimize f under IF. Under IF, everyone must be treated identically; the allocation that optimizes f selects $c = 0$ deterministically for an expected value of $\frac{1}{k-1} + \varepsilon > \frac{1}{k-1}$. Because all individuals receive utility 1 from outcome 0, the expected social welfare is 1. Under PIIF, allocations can be chosen on a per group basis, so long as individuals from one group don't prefer to switch to another group's allocation. Consider the allocation π where for each $i \in \mathcal{X}_c$, i deterministically sees outcome c. Under this allocation $f(\pi) = 1$ but the expected social welfare has dropped to $1/T$. Still, it's not hard to see that this allocation satisfies PIIF. To see this, note that no one is allocated $c = 0$. Conditioned on this fact, everyone is seeing their most-preferred allocation; thus, the solution satisfies PIIF. □

Notably, the construction relies on the fact that the platform's revenue is completely opposed to individuals' preferences. Indeed, all the individuals preferred to see outcome $c = 0$, but this outcome generated much less revenue than the others.

Takeaways. In this section, we demonstrated that selecting an allocation under PIIF constraints may allow for better social welfare compared to IF constraints, but does not guarantee it. In fact, if the platform's objective is not in line with the social welfare of individuals, PIIF (and EF) may allow for worse social welfare than IF. Taken together, we see that PIIF decouples the issue of ensuring fair treatment from that of ensuring beneficial outcomes.

To highlight this fact, we note that PIIF can be paired with constraints on the social welfare to ensure beneficial outcomes. Even though PIIF does not guarantee improved social welfare compared to IF, we observe that social welfare is a linear function in the individuals' utilities. Thus, whenever we may be concerned that the platform's objective may be misaligned with individuals' utility, we can simply add a constraint to the optimization program that ensures the social welfare is above some baseline. In particular, an appropriate individually fair solution could act as this baseline. Under this baseline solution, we can compute the social welfare, then require that the resulting PIIF solution have at least this social welfare. Again, this program is still feasible because every IF solution is also PIIF-feasible. At the extreme, we could even require a constraint on the utility experience by each individual; under this program, any deviations from an IF solution can truly only come by improving the outcomes for individuals.

4 Fairness in the multiple-task setting

In this section we study preference-informed individual fairness in the *multiple-task* setting, formalized and studied in Dwork and Ilvento [DI19]. In this setting, we allow separate fairness metrics for each outcome (here, task). This is a natural setting for studying fairness in targeted advertising, where different ads are subject to different fairness constraints. We start with a discussion of issues of discrimination in targeted advertising in Section 4.1. We then present the definitions of IF and PIIF in the multiple-task setting. We argue that also in this setting PIIF provides effective protection against discrimination (akin to that of IF), while effectively enriching the set of possible solutions.

4.1 Issues of fairness in targeted advertising

Online advertising has become pervasive and significantly influences the exposure individuals have to the world and its opportunities [VDSKK18, NCD18]. Moreover, in recent years, it has increasingly shifted to targeted advertising – where criteria such as age, gender, ethnic affinity, interests, marital status, and political affiliation, as well as more advanced machine-learning and data-driven features such as Lookalike Audiences, can be used by advertisers to choose the target audience. As discussed in Section 1.3 the dominant rhetoric is that targeted advertising benefits everyone.

Unfairness and discrimination can arise as a result of deliberate action of the advertiser, who decides to exclude people unfairly (or in ways that are considered illegal) through the targeting criteria they use, e.g., advertising a surgeon position only to men and not to women. Specifically, it has been observed that the Facebook advertising platform enabled advertisers to exclude users by "ethnic affinities" from housing ads [ATV17, AJ16, Tob18] (arguably, a violation of the Fair Housing Act of 1968), and that certain employers target job ads to users only of certain age and gender [LT18, TM18b, AST17] even when those characteristics are not relevant to qualifications for the jobs being advertised (arguably, a violation of the federal Age Discrimination in Employment Act of 1967).

The platforms' current rhetoric with regards to possible unfairness in their advertising products [TM18a, Tob19, San19] centers on assigning the responsibility for discriminatory advertising on individual advertisers and thus, claiming immunity from liability due to Section 230 of the Communications Decency Act of 1996, while emphasizing that the platform optimizes for the user. For example, Facebook's Ad Principles [Fac17b] claim "Our action system prioritizes what's most relevant to you, rather than how much money Facebook will make from any given ad" and Facebook argues that age-based targeting in employment ads benefits advertisers and users [TM18a]. Hence, the platforms' efforts in the fairness space are centered on policing and educating the advertisers on these issues: asking them to review and accept Facebook's non-discrimination policy and emphasizing "advertiser education" [Fac17a, Fac18]. The work of [SAV$^+$18] details the difficulties of ensuring individual advertisers do not run discriminatory campaigns.

However, as has begun to emerge from academic and legal work [DI19, DTD15, LT18, ASB$^+$19, Upt18, BTI19, Tob19], ensuring that advertisers run fair campaigns is not sufficient for preventing discriminatory or unfair outcomes when the advertising ecosystem is looked at as a whole. Specifically, ad allocation and delivery is a complex process that depends on many inputs, not all of which

are under the control of an individual advertiser. For example, the platforms acknowledge [Fac, Hel] that selection of the ad to show to a user among all ads targeting that user depends on the advertiser's bid, the platform's estimate of the action rate users are likely to take in response to seeing an ad, and the ads' quality and relevance to the user. It is natural to hypothesize [LT18, Upt18] that bias in the platform's estimates of action rates, ad's quality and relevance (for example, due to such estimators' being trained on biased data [O'N17] or biased exploration during the action rate estimate phase) can lead to unfair or discriminatory ad allocations.

Moreover, factors such as unequal availability and demand for certain sub-populations, differences in action rates across sub-populations, and the platform's desire to optimize its own revenue or advertiser happiness, as well as differing interpretations of what it may mean to prioritize what is most relevant for the user, can lead to allocations that are unfair. Specifically, the empirical work of [LT18] has shown that a Facebook ad intended to be gender-neutral in its delivery was shown to more men than women, because younger women are more expensive to show ads to; and [DTD15] observed that simulated female accounts received fewer instances of ads encouraging the taking of high-paying jobs than otherwise identical simulated female accounts through Google's ad platform. Most recently, the work of [ASB+19] has demonstrated that concerns about the role that market effects and the ad allocation process alone can play in creating unfair outcomes are not merely theoretical, and that Facebook delivers ads to audiences skewed by race and gender even when advertisers target large, inclusive audiences.

Finally, the work of [DI19] shows that fairness for individual campaigns does not automatically compose into fairness of the ecosystem under the individual fairness definition (Section 4.4.1), making the question of finding approaches for ensuring fairness in targeted advertising as a whole even more pressing.

4.2 Multiple-task setting and targeted advertising

In the multiple task setting, allocations consist of distributions over outcomes over a large set $\mathcal{C}$, where each $c \in \mathcal{C}$ may be viewed as a *distinct* classification task. We assume there is a separate fairness metric d_c for each task $c \in \mathcal{C}$. Assuming the existence of $k = |\mathcal{C}|$ metrics $d_1, \ldots, d_k$, [DI19] defined Individual Fairness for the multiple-task setting as enforcing the individual fairness constraint, simultaneously for every outcome.

Definition 7 (Multiple-task individual fairness)**.** *An allocation $\pi : \mathcal{X} \to \Delta(\mathcal{C})$ is said to be* $(D, \{d_1, \ldots d_k\})$-individually fair *in the multiple-task setting if for every two individuals $i, j \in \mathcal{X} \times \mathcal{X}$, the following Lipschitz condition holds:*

$$\forall c \in \mathcal{C} : \quad D(\pi(i)_c, \pi(j)_c) \leq d_c(i, j) \tag{11}$$

Throughout this section we will use targeted advertising as our running example. We formalize it through the *single-slot advertising model*, defined below. We remark that this simple model allows us to focus on fairness at the cost of abstracting away many real-world, important aspects of the online advertising world. We revisit these assumptions and address interesting future research directions in Section 4.5.

An advertising system consists of three types of parties – users, advertisers, and the platform itself:

(i) Users: $\mathcal{X} = \{1, 2, \ldots, n\}$ are the users of the platform, each of which will be displayed a *single* ad. As in Section 3, we assume that users have preferences over the possible ads they may see. For a user $i \in \mathcal{X}$, these are modeled as a reflexive and transitive binary relation $\preceq_i$ over $\Delta(\mathcal{C})$.

(ii) Advertisers: $\mathcal{C} = \{c_1, \ldots, c_k\}$ are the different advertising campaigns (or advertisers, for short), competing for the attention of the users. They express their interest in users through bidding; We use the vector $\mathbf{b}(c_t) \in \mathbb{R}^n$ to denote the bids of advertiser $t \in [k]$.

(iii) The platform decides on a *solution*: an assignment of ads to users. We use $\mathbf{p}(i) \in \mathbb{R}^k$ to denote the (probabilistic) allocation that a user $i \in \mathcal{X}$ receives, where $\sum_{c \in \mathcal{C}} p(i)_c \leq 1$.

For brevity, we also use matrix form notation $P, B \in \mathbb{R}^{n \times k}$, where $P_{i,c}$ is the probability with which user i is shown ad c and $B_{i,c}$ is the bid of advertiser c on user i.

Our primary motivation for incorporating preferences in the fairness definition was the observation that individual fairness may be overly restrictive, from the individuals' perspective. In the Multiple-Task setting, and particularly in the targeted advertising model, this becomes particularly salient.

For an illustrative example, consider the single-slot advertising setup with only two advertisers $\mathcal{C} = \{A, B\}$. Advertiser A, who sells children's toys, bids only on parents, and Advertiser B, who is trying to hire individuals for a high-paying tech-job, bids similarly for equally qualified parents and non-parents. They both bid in a manner that's individually fair *separately*. This example is used in [DI19] to demonstrate failure of *naive composition*[5]. Indeed, note that if A outbids B and the platform assigns ads to the higher bidder, the resulting system will be one in which all the parents see children's toys, whereas the non-parents see the tech job ad. We focus our attention, however, to the fact that if it is the case that particular parents may be happy with their job and actually prefer to see toy ads over job ads, then this allocating A to these parents shouldn't be ruled out for the sake of fairness. More generally, the constraint in (11) may rule out the solution in which every person sees their favorite ad!

We now define preference-informed fairness in the multiple-task setting, which addresses precisely these issues.

Definition 8 (Multiple-task preference-informed individual fairness)**.** *An allocation $\pi : \mathcal{X} \to \Delta(\mathcal{C})$ 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 another allocation $\pi^j(i) \in \Delta(\mathcal{C})$ such that:*

$$\forall c \in \mathcal{C}: \quad D\left(\pi^j(i)_c, \pi(j)_c\right) \leq d_c(i, j) \tag{12}$$

$$\pi^j(i) \preceq_i \pi(i) \tag{13}$$

The idea is similar to PIIF in the single-task setting: the preference-informed extension of individual fairness 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 imagined allocation, $\pi^j(i)$. The main distinction is that now $\pi^j(i)$ has to satisfy Multiple-task IF with respect to j's current allocation.

[5]Intuitively, this is a result of the fact that users that are similar with respect to a particular ad may not be similar with respect to their general "desirability" in the system, which naturally effects the outcomes.

Revisiting the example of advertisers A and B above: unlike the constraint in (11), PIIF will, in fact, allow individuals that are similar with respect to an ad to see it with a different probability – so long that it's for a "good cause": e.g. to enable the individual who sees it with a lower probability to instead see some other ad they *prefer* with a higher probability. In particular, the solution in which every user sees their favourite ad satisfies this definition (even though it may not be obtainable in practice, due to other constraints in the system).

4.3 Efficient optimization in the multiple-task setting

We assume that individuals' preferences admit an expected utility representation, so the social welfare (in this case, *users' happiness*) is properly defined:

$$W(P) = \sum_{i \in \mathcal{X}} \sum_{c \in \mathcal{C}} u_i(c) \cdot P_{i,c} \tag{14}$$

We denote with P^* the allocation rule that, given bids B, ad-specific metrics $d_1, \ldots, d_k$, and advertiser budgets $T_1, \ldots, T_k$, maximizes social welfare subject to budget constraints and preference-informed individual fairness. Building on our results from Section 3, we have a *centralized* algorithm for efficiently computing P^*:

Theorem 4.1. *Given the advertisers' bids B and budgets $T_1, \ldots, T_k$, and assuming oracle-access to the fairness metrics $d_1, \ldots, d_k$ and individuals' utilities, P^* can be computed in time poly(n, k).*

4.4 Fairness and social welfare in the multiple-task setting

The construction given in Section 3.4 demonstrates that in the single-task setting, the gap between the best social welfare obtainable under IF and PIIF can be large. In principle, this construction can be generalized to the multiple-task setting. However, one unconvincing aspect of the construction is that every individual is identical. 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, it would seem natural that fair treatment would allow for high social welfare allocations. Still, we show that this intuition does not carry through for individual fairness: for a set of tasks $\mathcal{C}$ where $|\mathcal{C}| = k$, there are instances where the optimal social welfare under PIIF approaches a factor k larger than the best IF solution.

Proposition 4.2. *Suppose $|\mathcal{C}| = k$; for some $t > 0$, let $|\mathcal{X}| \geq t^k$. There exists a distribution of multiple-task allocation problem instances where the similarity metric for each task c is given as $d_c(i,j) = |u_i(c) - u_j(c)|$ such that as $t \to +\infty$*

$$\mathcal{W}^* \left(\Pi^{IF}\right) \leq \frac{1}{k} \cdot \mathcal{W}^* \left(\Pi^{PIIF}\right).$$

Proof. Suppose the universe of individuals $\mathcal{X} = S_0 \cup S_1 \cup \ldots S_{k-1}$ is partitioned into disjoint subpopulations ($S_i \cap S_j = \varnothing$ for $i \neq j$). The subpopulations will become progressively smaller as i increases; for each $i > 0$, we denote by p_i the size of the subpopulation S_i, where $p_i \triangleq |S_i| / |\mathcal{X}| =$

$1/t^i$ for some constant $t > 0$ and let $p_0 = 1 - \sum_{i=1}^{k-1} p_i$. We will assume individuals have additive utility functions over their allocation. For every individual $r \in S_j$, the utility for seeing outcome c_i is $u_{S_j}(c_i) = t^i$ if $j \geq i$, and $u_{S_j}(c_i) = 0$ otherwise.

For all $j \in [k]$, let $T_j = \cup_{i \leq j} S_i$. We take the similarity for the task c_i to be determined by the utility received by individuals; that is, for task c_i every pair of individuals $r, t \in T_i \times T_i$ are considered similar $d(r, t) = 0$ and arbitrarily dissimilar otherwise.

First, consider the allocation that assigns every individual in S_i to campaign c_i. The social welfare can then be written as:

$$\begin{aligned}\sum_{i=0}^{k-1} p_i \cdot u_{S_i}(c_i) &= \left(1 - \sum_{i=1}^{k} t^{-i}\right) + \sum_{i=1}^{k} t^{-i} \cdot t^i \\ &= k - \sum_{i=1}^{k-1} t^{-i}\end{aligned}$$

Note that this allocation satisfies PIIF; indeed, it is the social-welfare-maximizing allocation.

Then, consider the optimal IF allocation. Recall that every individual in T_i is similar according to task c_i because they all receive the same utility. Let α_i denote the probability of assigning campaign c_i to these individuals; these must be the same probabilities for all $r \in T_j$ by individual fairness. Then, we can compute the expression for the social welfare of any such assignment.

$$\sum_{i=0}^{k-1} \alpha_i \cdot \sum_{j \geq i} p_j \cdot u_{S_j}(c_i) = \sum_{i=0}^{k-1} \alpha_i t^i \cdot \sum_{j \geq i} p_j \tag{15}$$

$$= \alpha_0 \cdot \left(1 - \sum_{i=1}^{k-1} \frac{1}{t^i}\right) + \sum_{i=1}^{k-1} \alpha_i \cdot t^i \cdot \sum_{j \geq i} \frac{1}{t^j} \tag{16}$$

$$= \sum_{i=0}^{k-1} \alpha_i - \alpha_0 \cdot \sum_{i=1}^{k-1} \frac{1}{t^i} + \sum_{i=1}^{k-1} \alpha_i \cdot \sum_{j > i} \frac{1}{t^j} \tag{17}$$

$$\leq 1 + \sum_{i=2}^{k-1} \frac{1}{t^i} \tag{18}$$

The final equality follows by the fact that individuals in S_{k-1} must see all campaigns because they are similar in every campaign, so the probabilities have to sum to 1. Taking $t \to +\infty$, the ratio between these two quantities tends to $1/k$. □

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 platform.

4.4.1 RandomizeThenClassify

A corollary of our result is that the Dwork-Ilvento "RandomizeThenClassify" mechanism for composition under IF achieves worst-case optimal performance (in terms of both social welfare and utility to the platform).

If we choose to display any distribution over the set of social-welfare-maximizing (resp., utility-maximizing) campaigns, then this will achieve the social welfare (utility) of the best campaign. This cannot be worse than $1/k$ (with equality when all the campaigns give the same welfare/utility). In particular, the result shows that even if you had full information about individuals' welfare and the platform's utility, if you wish to enforce campaign-wise individual fairness, in the worst-case, you cannot achieve better social welfare (utility) than randomly picking a campaign and showing the impression to the individual if relevant.

4.5 Discussion

We conclude by highlighting two particular aspects of the targeted advertising world which our current formulation does not address, and some interesting open questions.

Offline setting, full information. In our simplified model we've made use of two assumptions: that all users are available simultaneously, and that the platform has full information of individuals' preferences. An *online* setting could naturally be more applicable. It also has the useful aspect which is that the preferences could be "discovered" during the process (see [GJKR18] for a similar approach regarding the metric itself). There are a few non-trivial aspects involved, however. First, the interpersonal nature of the PIIF constraints implies that it's impossible to reason about a single user's allocation in isolation. Second, learning individuals' preferences naturally requires some exploration, which may be at odds with ensuring fair treatment. How do these two competing objectives play out?

Prices and advertisers' incentives. We assumed that advertisers are bidding on users' attention, but we did not address the fact that their behaviour in the system could be strategic. Towards a practical implementation of the PIIF ideas, we'd want to design an *incentive-compatible* mechanism: e.g., an assignment of ads to users as well as a pricing scheme such that bidding truthfully becomes a dominant strategy for the advertisers. Therefore, a natural question is investigating conditions under which the allocation rule P^* (maximizing welfare subject to PIIF constraints) is also *implementable*, i.e., admits such a pricing scheme.

5 Preference-Informed Extensions of Other Fairness Notions

In this work, our focus was on incorporating individual preferences into the metric-based individual fairness framework of *Dwork et al.* [DHP$^+$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 subgroup, S, where the rest of the population is denoted $T = \mathcal{X} - S$. 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 [DHP+12, KNRW17, HJKRR17]. 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 individual fairness [DHP+12] also extends to group-fairness notions. In this section we demonstrate this by focusing on a particular definition, Statistical Parity (SP), in the context of binary classification, where the measure to be equalized across S and T is the rate of positive predictions.

Statistical parity for a binary classifier mapping individuals to outcomes in $\{\pm 1\}$ is defined as follows:

Definition 9 (Statistical parity)**.** *A binary classifier $h : \mathcal{X} \to \{\pm 1\}$ satisfies (exact) statistical parity with respect to S if*

$$\mathop{\mathbf{Pr}}_{i \sim S}[h(i) = 1] = \mathop{\mathbf{Pr}}_{i \sim T}[h(i) = 1] \tag{19}$$

When the outcome which is enforced to be equalized across S and T is clearly the "desirable" outcome, statistical parity guarantees equal exposure for members of the protected subgroup. Since SP ignores the ground truth (namely, the fraction of *true positives* within S and T), it is often used for corrective discrimination when data might contain biases against S.

When individuals have diverse preferences over the outcome space, enforcing statistical parity may 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 a small fraction of S, denoted S', who prefer the outcome -1. In this case, the statistical parity constraints "block" the solution which is, from the individuals' perspective, optimal.

Building on this intuition, and using a similar approach to the one used in Section 3, we extend the set of fair classifiers. Assuming every individual $i \in X$ has a preference relation over $\{\pm 1\}$ (or even distributions over $\{\pm 1\}$), the definition of *preference informed statistical parity* allows deviations from SP, as long as they are aligned with the individuals' preferences.

Definition 10 (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:*

$$\forall j \in T, h'(j) = h(j) \tag{20}$$

$$\forall i \in S, h(i) \succeq_i h'(i) \tag{21}$$

$$\mathop{\mathbf{Pr}}_{i \sim S}\left[h'(i) = 1\right] = \mathop{\mathbf{Pr}}_{i \sim T}\left[h'(i) = 1\right] \tag{22}$$

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 preference-informed statistical parity:

1. It only enriches the set of solutions that satisfy SP. This is because any classifier that satisfies Definition (9) also satisfies Definition (10), by taking the alternative to be itself.

2. The classifier that is optimal for the individuals (each individual is assigned their favourite 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 constraints in Equations (20)-(22).

3. It 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 S. Rightfully, it does not satisfy preference-informed statistical parity. The alternative SP-classifier, that gives everyone $+1$, is preferred by some of the members of S (those not in S').

Acknowledgments. *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.*

References

[AJ16] Julia Angwin and Terry Parris Jr. Facebook lets advertisers exclude users by race. *ProPublica*, Oct 28, 2016. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/.

[ASB+19] 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. *arXiv preprint*, 2019.

[AST17] Julia Angwin, Noam Scheiber, and Ariana Tobin. Dozens of companies are using facebook to exclude older workers from job ads. *ProPublica*, Dec 20, 2017. https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men.

[ATV17] Julia Angwin, Ariana Tobin, and Madeleine Varner. Facebook (still) letting housing advertisers exclude users by race. *ProPublica*, Nov. 21, 2017. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

[Baw75] Vijay S Bawa. Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, 2(1):95–121, 1975.

[BDNP18] Maria-Florina Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. *arXiv preprint arXiv:1809.08700*, 2018.

[BTI19] 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. https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html.

[Das18] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, Oct 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecrui

[DHP+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.

[DI19] 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.

[DTD15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[Fac] Facebook. Facebook: About our ad auction. https://www.facebook.com/business/help/430291176997542.

[Fac17a] Facebook. Improving enforcement and promoting diversity: Updates to ads policies and tools, February 8, 2017. https://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ad

[Fac17b] Facebook. Our advertising principles, Nov 27, 2017. https://newsroom.fb.com/news/2017/11/our-advertising-principles.

[Fac18] Facebook. Keeping advertising safe and civil, August 21, 2018. https://www.facebook.com/business/news/keeping-advertising-safe-and-civil.

[Fol67] Duncan K Foley. *Resource allocation and the public sector*. PhD thesis, Yale University, 1967.

[GJKR18] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*, 2018.

[Hei12] Russell Heimlich. Internet Users Don't like Targeted Ads. *PEW Research Center*, Mar 13, 2012. http://www.pewresearch.org/fact-tank/2012/03/13/internet-users-dont-like-targeted-ads/l.

[Hel] Google AdSense Help. About the ad auction. https://support.google.com/adsense/answer/160525?hl=en.

[HJKRR17] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

[HR69] Josef Hadar and William R Russell. Rules for ordering uncertain prospects. *The American economic review*, 59(1):25–34, 1969.

[KNRW17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[LT18] Anja Lambrecht and Catherine E Tucker. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *SSRN https://ssrn.com/abstract=2852260*, 2018.

[NCD18] Anthony Nadler, Matthew Crain, and Joan Donovan. Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech. *Data and Society*, Oct 2018. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf.

[O'N17] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[PRMT17] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. Exploring user perceptions of discrimination in online targeted advertising. In *USENIX Security*, 2017.

[RW98] Jack Robertson and William Webb. *Cake-cutting algorithms: Be fair if you can*. AK Peters/CRC Press, 1998.

[San19] Sheryl Sandberg. Doing more to protect against discrimination in housing, employment and credit advertising. *Facebook Newsroom*, Mar 19, 2019. https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/.

[SAV+18] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, volume 81, pages 1–15, 2018.

[Su99] Francis Edward Su. Rental harmony: Sperner's lemma in fair division. *The American mathematical monthly*, 106(10):930–942, 1999.

[Swe13] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[TH19] Joseph Turow and Chris Jay Hoofnagle. Mark Zuckerberg's Delusion of Consumer Consent. *The New York Times*, Jan 29, 2019. Opinion in the New York Times https://www.nytimes.com/2019/01/29/opinion/zuckerberg-facebook-ads.html.

[TM18a] Ariana Tobin and Jeremy B. Merrill. Besieged Facebook Says New Ad Limits Aren't Response to Lawsuits. *ProPublica*, Aug 23, 2018. https://www.propublica.org/article/facebook-says-new-ad-limits-arent-response-to-lawsuits.

[TM18b] Ariana Tobin and Jeremy B. Merrill. Facebook is letting job advertisers target only men. *ProPublica*, Sept 18, 2018. https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men.

[Tob18] Ariana Tobin. Facebook promises to bar advertisers from targeting ads by race or ethnicity. again. *ProPublica*, July 25, 2018. https://www.propublica.org/article/facebook-promises-to-bar-advertisers-from-targeting-ads-by-race-or-e

[Tob19] Ariana Tobin. HUD sues Facebook over housing discrimination and says the company's algorithms have made the problem worse. *ProPublica*, Mar 28, 2019. https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms.

[Upt18] Upturn. Upturn amicus brief in Onuoha v. Facebook. Nov 16, 2018. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf.

[Var74] Hal Varian. Efficiency, equity and envy. *Journal of Economic Theory*, 9:63–91, 1974.

[VDSKK18] Jennifer Valentino-DeVries, Natasha Singer, Michael H. Keller, and Aaron Krolik. Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret. *The New York Times*, Dec 10, 2018. https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html.

[VNM07] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton University Press, 2007.

[Zuc19] 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.

[ZVR+17] 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.

A Expected Utility Theory: A Brief Overview

In this section, we give a brief overview of expected utility theory [VNM07]. This has been the standard approach in economics for modeling decision-making under uncertainty. This section mostly consists of the technical language requires to present the main result, the vNM Expected utility theorem of [VNM07], which characterizes the precise conditions under which preference admit an *expected utility form.*

A.1 Preferences over certain outcomes

Before we model individuals' preferences over $\Delta(\mathcal{C})$ (representing uncertain outcomes, or *lotteries*), let us first consider preferences over the set of alternatives $\mathcal{C}$ (representing certain outcomes). Ideally, we'd want to represent these preferences in the form of a utility function, $u(x)$, that assigns a numerical value to each $c \in \mathcal{C}$, such that the rank ordering of the alternatives is preserved. This function is often referred to as a (Bernoulli) utility (payoff) function.

Definition 11 (Bernoulli utility function)**.** *$u : \mathcal{C} \to R$ is a utility function representing a preference relation $\succeq$ if for every $c_1, c_2 \in \mathcal{C}$, $c_1 \succeq c_2 \iff u(c_1) \geq u(c_2)$.*

A preference relation is said to be *rational* if it satisfies the two following axioms.

Axiom 1 (Completeness): A preference ordering $\succeq$ over $\mathcal{C}$ is complete iff for any two outcomes $c_1, c_2 \in \mathcal{C}$, either $c_1 \succ c_2$, $c_2 \succ c_1$, or $c_1 \sim c_2$. That is, any two alternatives in $\mathcal{C}$ are comparable: either one is strictly preferred to the other, or the individual is indifferent between them.

Axiom 2 (Transitivity): A preference ordering $\succeq$ over $\mathcal{C}$ is transitive if for any three outcomes $c_1, c_2, c_3 \in \mathcal{C}$, if $c_1 \succeq c_2$ and $c_2 \succeq c_3$, then $c_1 \succeq c_3$.

Rational preferences provide an exact characterization of when a preference relation over certain outcomes can be represented by a utility function.

Theorem A.1. *If $\mathcal{C}$ is finite, then a preference relation $\succeq$ over $\mathcal{C}$ admits a utility function representation $u : C \to R$ iff it is rational (satisfies Axioms 1,2).*

A.2 Preferences over uncertain outcomes

The expected utility form extends the above result to preferences over uncertain outcomes, $\Delta(\mathcal{C})$.

Definition 12 (Expected utility form)**.** *A utility function $U : \Delta(C) \to \mathbb{R}$ has an expected utility form if there exists a (Bernoulli) utility (payoff) function $u : C \to \mathbb{R}$ that assigns real numbers to outcomes, such that $\forall \pi \in \Delta(C), U(\pi) = \sum_{c \in C} \pi(c) \cdot u(c)$* [6].

Note a preference relation over $\Delta(C)$ that has this form effectively ranks allocations by their expected utility over alternatives. Importantly for our purpose, is that $U(\cdot)$ is linear in the probabilities.

[6] Typically, U is referred to as an expected utility function and u as a von Neumann-Morgenstern utility function

We introduce two additional assumptions.

Axiom 3 (Independence over lotteries): Let $\alpha \in (0,1)$ and p, q, r three lotteries. Then, $p \succeq q$ iff $\alpha p + (1-\alpha)r \succeq \alpha q + (1-\alpha)r$.

Axiom 4 (Continuity): Let c_1, c_2, c_3 be three alternatives such that $c_1 > c_2 > c_3$. Then there exists a unique $\alpha \in (0,1)$ such that you are indifferent between the lottery $\alpha c_1 + (1-\alpha)c_3$ and c_2 with certainty.

We denote $\succeq^{EUT}$ the set of all preferences over $\Delta(\mathcal{C})$ that satisfy Axioms (1) - (4). The seminal result of Von Neumann and Morgenstern [VNM07] proves that this is a complete characterization of the set of preferences that admit an expected utility form.

Theorem A.2 (vNM Expected utility theorem)**.** *A preference relation $\succeq$ on $\Delta(\mathcal{C})$ is in $\succeq^{EUT}$ (satisfies Axioms 1-4) iff there exists a function that assigns a real number to each outcome, $u : C \to \mathbb{R}$, such that for any two allocations $\pi, \pi' \in \Delta(C)$, $\pi \succeq \pi' \iff \sum_{c \in C} \pi(c) \cdot u(c) \geq \sum_{c \in C} \pi'(c) \cdot u(c)$.*