Fairness Under Composition

Cynthia Dwork[1]
Harvard John A Paulson School of Engineering and Applied Science,
Radcliffe Institute for Advanced Study, Cambridge, MA, USA
dwork@seas.harvard.edu

Christina Ilvento[2]
Harvard John A Paulson School of Engineering and Applied Science,
Cambridge, MA, USA
cilvento@g.harvard.edu

Abstract

Algorithmic fairness, and in particular the fairness of scoring and classification algorithms, has become a topic of increasing social concern and has recently witnessed an explosion of research in theoretical computer science, machine learning, statistics, the social sciences, and law. Much of the literature considers the case of a single classifier (or scoring function) used once, in isolation. In this work, we initiate the study of the fairness properties of systems composed of algorithms that are fair in isolation; that is, we study *fairness under composition*. We identify pitfalls of naïve composition and give general constructions for fair composition, demonstrating both that classifiers that are fair in isolation do not necessarily compose into fair systems and also that seemingly unfair components may be carefully combined to construct fair systems. We focus primarily on the individual fairness setting proposed in [Dwork, Hardt, Pitassi, Reingold, Zemel, 2011], but also extend our results to a large class of group fairness definitions popular in the recent literature, exhibiting several cases in which group fairness definitions give misleading signals under composition.

2012 ACM Subject Classification Theory of computation → Computational complexity and cryptography, Theory of computation → Design and analysis of algorithms, Theory of computation → Theory and algorithms for application domains

Keywords and phrases algorithmic fairness, fairness, fairness under composition

Digital Object Identifier 10.4230/LIPIcs.ITCS.2019.33

Related Version A full version of the paper is available at `https://arxiv.org/abs/1806.06122`.

1 Introduction

As automated decision-making extends its reach ever more deeply into our lives, there is increasing concern that such decisions be fair. The rigorous theoretical study of fairness in algorithmic classification was initiated by Dwork *et al* in [4] and subsequent works investigating alternative definitions, fair representations, and impossibility results have proliferated in the machine learning, economics and theoretical computer science literatures.[3] The notions of fairness broadly divide into *individual fairness*, requiring that individuals who are similar with respect to a given classification task (as measured by a task-specific

[1] This work was supported in part by Microsoft Research and the Sloan Foundation.
[2] This work was supported in part by the Smith Family Fellowship and Microsoft Research.
[3] See also [20] [9] and [10], which predate [4] and are motivated by similar concerns.

similarity metric) have similar probability distributions on classification outcomes; and *group fairness*, which requires that different demographic groups experience the same treatment in some average sense.

In a bit more detail, a *classification task* is the problem of mapping *individuals* to *outcomes*; for example, a decision task may map individuals to outcomes in $\{0, 1\}$. A classifier is a possibly randomized algorithm solving a classification task. In this work we initiate the study of *fairness under composition*: what are the fairness properties of systems built from classifiers that are fair in isolation? Under what circumstances can we ensure fairness, and how can we do so? A running example in this work is online advertising. If a set of advertisers, say, one for tech jobs and one for a grocery delivery service, compete for the attention of users, say one for tech jobs and one for a grocery delivery service, and each chooses fairly whether to bid (or not), is it the case that the advertising system, including budget handling and tie-breaking, will also be fair?

We identify and examine several types of composition and draw conclusions about auditing systems for fairness, constructing fair systems, and definitions of fairness for systems. In the remainder of this section we summarize our results and discuss related work. A full version of this paper, containing complete proofs of all our results, is available on ArXiv.

Task-Competitive Compositions

We first consider the problem of two or more tasks competing for individuals, motivated by the online advertising setting described above. We prove that two advertisers for different tasks, each behaving fairly (when considered independently), will not necessarily produce fair outcomes when they compete. Intuitively (and as empirically observed by [17]), the attention of individuals similarly qualified for a job may effectively have different costs due to these individuals' respective desirability for other advertising tasks, like household goods purchases. That is, individuals claimed by the household goods advertiser will not see the jobs ad, regardless of their job qualification. These results are not specific to an auction setting and are robust to choice of "tie-breaking" functions that select among multiple competing tasks (advertisers). Nonetheless, we give a simple mechanism, RandomizeThenClassify, that solves the fair task-competitive classification problem using classifiers for the competing tasks, each of which is fair in isolation, in a black-box fashion and without modification. In the Appendix (Lemma 15) we give a second technique for modifying the fair classifier of the lower bidder (loser of the tie-breaking function) in order to achieve fairness.

Functional Compositions

Is the "OR" of two fair clssifiers also fair? Moe generally, when can we build fair classifiers by computing on values that were fairly obtained? Here we must understand what is the salient outcome of the computation. For example, when reasoning about whether the college admissions system is fair, the salient outcome may be whether a student is accepted to at least one college, and not whether the student is accepted to a specific college[4]. Even if each college uses a fair classifier, the question is whether the "OR" of the colleges' decisions is fair. Furthermore, an acceptance to college may not be meaningful without sufficient accompanying financial aid. Thus in practice, we must reason about the OR of ANDs of acceptance and financial aid across many colleges. We show that although in general there

[4] In this simple example, we assume that all colleges are equally desirable, but it is not difficult to extend the logic to different sets of comparable colleges.

are no guarantees on the fairness of functional compositions of fair components, there are some cases where fairness in ORs can be satisfied. Such reasoning can be used in many applications where long-term and short-term measures of fairness must be balanced. In the case of feedback loops, where prior positive outcomes can improve the chances of future positive outcomes, functional composition provides a valuable tool for determining at which point(s) fairness must be maintained and determining whether the existing set of decision procedures will adhere to these requirements.

Dependent Compositions

There are many settings in which each individual's classifications are dependent on the classifications of others. For example, if a company is interviewing a set of job candidates in a particular order, accepting a candidate near the beginning of the list precludes any subsequent candidates from even being considered. Thus, even if each candidate actually considered is considered fairly in isolation, dependence between candidates can result in highly unfair outcomes. For example, individuals who are socially connected to the company through friends or family are likely to hear about job openings first and thus be considered for a position before candidates without connections. We show that selecting a cohort of people – online or offline – requires care to prevent dependencies from undermining an independently fair selection mechanism. We address this in the offline case with two randomized constructions, PermuteThenClassify and WeightedSampling. These algorithms can be applied in the online case, even under adversarial ordering, provided the size of the universe of individuals is known; when this is not known there is no solution.

Nuances of group-based definitions

Many fairness definitions in the literature seek to provide fairness guarantees based on group-level statistical properties. For example, "Equal Opportunity" [6] requires that, conditioned on qualification, the probability of a positive outcome is independent of protected attributes such as race or gender. Group Fairness definitions have practical appeal in that they are possible to measure and enforce empirically without reference to a task-specific similarity metric.[5] We extend our results to group fairness definitions and we also show that these definitions do not always yield consistent signals under composition. In particular, we show that the intersectional subgroup concerns (which motivate [11, 7]) are exacerbated by composition. For example, an employer who uses group fairness definitions to ensure parity with respect to race and gender may fail to identify that "parents" of particular race and gender combinations are not treated fairly. Task-competitive composition exacerbates this problem, as the employers may be prohibited from even collecting parental status information, but their hiring processes may be composed with other systems which legitimately differentiate based on parental status.

Finally, we also show how naïve strategies to mitigate these issues in composition may result in learning a nominally fair solution that is clearly discriminating against a socially meaningful subgroup not officially called out as "protected," from which we conclude that understanding the behavior of fairness definitions under composition is critical for choosing which definition is meaningful in a given setting.

[5] However, defining and measuring qualification may require care.

Implications of Our Results

Our composition results have several practical implications. First, testing individual components without understanding of the whole system will be insufficient to safely draw either positive or negative conclusions about the fairness of the system. Second, composition properties are an important point of evaluation for any definitions of fairness or fairness requirements imposed by law or otherwise. Failing to take composition into account when specifying a group-based fairness definition may result in a meaningless signal under composition, or worse may lead to ingraining poor outcomes for certain subgroups while still nominally satisfying fairness requirements. Third, understanding of the salient outcomes on which to measure and enforce fairness is critical to building meaningfully fair systems. Finally, we conclude that there is significant potential for improvement in the mechanisms proposed for fair composition and many settings in which new mechanisms could be proposed.

1.1 Related Work

Fairness retained under post-processing in the single-task one-shot setting is central in [22, 19, 4]. The definition of individual fairness we build upon in this work was introduced by Dwork *et al.* in [4]. Learning with oracle access to the fairness metric is considered by [5, 13]. A number of group-based fairness definitions have been proposed, and Ritov *et al.* provide a combined discussion of the parity-based definitions in [21]. In particular, their work includes discussion of Hardt *et al.*'s Equality of Opportunity and Equal Odds definitions and Kilbertus *et al.*'s Counterfactual Fairness [6, 12]. Kleinberg *et al.* and Chouldechova independently described several impossibility results related to simultaneously satisfying multiple group fairness conditions in single classification settings [14],[2].

Two concurrent lines of work aiming to bridge the gap between individual and group consider ensuring fairness properties for large numbers of large groups and their (sufficiently large) intersections [11, 7]. While these works consider the one-shot, single-task setting, we will see that group intersection properties are of particular importance under composition. Two subsequent works in this general vein explore approximating individual fairness with the help of an oracle that knows the task-specific metric [13, 5]. Two works also consider how feedback loops can influence fair classification, and how interventions can help [8, 18].

Several empirical or observational studies document the effects of multiple task composition. For example, Lambrecht and Tucker study how intended gender-neutral advertising can result in uneven delivery due to high demand for the attention of certain demographics [17]. Datta *et al.* also document differences in advertising based on gender, although they are agnostic as to whether the cause is due to multiple task composition or discriminatory behavior on the part of the advertisers or platform [3]. Whether it is truly "fair" that, say, home goods advertisers bid more highly for the attention of women than for the attention of men, may be debatable, although there are clearly instances in which differential targeting is justified, such as wen advertising maternity clothes. This *actuarial fairness* is the industry practice, so we pose a number of examples in this framework and analyze the implications of composition.

2 Preliminary Definitions and Assumptions

2.1 General Terminology

We refer to classifiers as being "fair in isolation" or "independently fair" to indicate that with no composition, the classifier satisfies a particular fairness definition. In such cases expectation and probability are taken over the randomness of the classification procedure

and, for group fairness, selection of elements from the universe. We denote the universe of individuals relevant for a task as U, and we generally use $u, v, w \in U$ to refer to universe elements. We generally consider binary classifiers in this work, and use p_w to denote the probability of assigning the positive outcome (or simply 1) to the element w for a particular classifier. We generally write $C : U \times \{0,1\}^* \to \{0,1\}$, where $\{0,1\}^*$ represents the random bits of the classifier. This allows us to comfortably express the probability of positive classification $\mathbb{E}_r[C(u)]$ as well as the output of the classifier under particular randomness $C(u,r)$. In this notation, $p_u = \mathbb{E}_r[C(u)]$. When considering the distribution on outputs of a classifier C, we use $\tilde{C} : U \to \Delta(\{0,1\})$. When two or more classifiers or tasks are compared, we either use a subscript $_i$ to indicate the i^{th} classifier or task, or a prime ($'$) to indicate the second classifier or task. For example $\{C, C'\}$, $\{C_i | i \in [k]\}$, $\{T, T'\}$, $\{T_i | i \in [k]\}$.

2.2 Individual Fairness

Throughout this work, our primary focus is on *individual fairness*, proposed by Dwork *et al* in [4]. As noted above, a *classification task* is the problem of mapping *individuals* in a universe to *outcomes*.

▶ **Definition 1** (Individual Fairness [4]). Let $d : \Delta(O) \times \Delta(O) \to [0,1]$ denote the total variation distance on distributions over O[6]. Given a universe of individuals U, and a task-specific metric $\mathcal{D}$ for a classification task T with outcome set O, a randomized classifier $C : U \times \{0,1\}^* \to O$, such that $\tilde{C} : U \to \Delta(O)$, is *individually fair* if and only if for all $u, v \in U$, $\mathcal{D}(u,v) \geq d(\tilde{C}(u), \tilde{C}(v))$.

Note that when $|O| = 2$ we have $d(\tilde{C}(u), \tilde{C}(v)) = |\mathbb{E}_r[C(u)] - \mathbb{E}_r[C(v)]| = |p_u - p_v|$. In several proofs we will rely on the fact that it is possible to construct individually fair classifiers with particular distance properties (see Lemma 16 and corollaries in the Appendix).

2.3 Group Fairness

In principle, all our individual fairness results extend to group fairness definitions; however, there are a number of technicalities and issues unique to group fairness definitions, which we discuss in Section 6. Group fairness is often framed in terms of *protected attributes* $\mathcal{A}$, such as sex, race, or socio-economic status, while allowing for differing treatment based on a set of *qualifications* $\mathcal{Z}$, such as, in the case of advertising, the willingness to buy an item. Conditional Parity, a general framework proposed in [21] for discussing these definitions, conveniently captures many of the popular group fairness definitions popular in the literature including Equal Odds and Equal Opportunity [6], and Counterfactual Fairness [16].

▶ **Definition 2** (Conditional Parity [21]). A random variable $\mathbf{x}$ satisfies parity with respect to $\mathbf{a}$ conditioned on $\mathbf{z} = z$ if the distribution of $\mathbf{x} \mid (\mathbf{a}, \{\mathbf{z} = z\})$ is constant in $\mathbf{a}$:
$\Pr[\mathbf{x} = x \mid (\mathbf{a} = a, \mathbf{z} = z)] = \Pr[\mathbf{x} = x \mid (\mathbf{a} = a', \mathbf{z} = z)]$ for any $a, a' \in \mathcal{A}$. Similarly, $\mathbf{x}$ satisfies parity with respect to $\mathbf{a}$ conditioned on $\mathbf{z}$ (without specifying a value of $\mathbf{z}$) if it satisfies parity with respect to $\mathbf{a}$ conditioned on $\mathbf{z} = z$ for all $z \in \mathcal{Z}$. All probabilities are over the randomness of the prediction procedure and the selection of elements from the universe.

[6] [4] also considered other notions of distributional distance.

3 Multiple-Task Composition

First, we consider the problem of composition of classifiers for multiple tasks where the outcome for more than one task is decided. Multiple Task Fairness, defined next, requires fairness to be enforced independently and simultaneously for each task.

▶ **Definition 3** (Multiple Task Fairness)**.** For a set $\mathcal{T}$ of k tasks with metrics $\mathcal{D}_1, \ldots, \mathcal{D}_k$, a (possibly randomized) system $\mathcal{S} : U \times r \to \{0,1\}^k$, which assigns outputs for task i in the i^{th} coordinate of the output, satisfies multiple task fairness if for all $i \in [k]$ and all $u, v \in U$ $\mathcal{D}_i(u,v) \geq |\mathbb{E}[\mathcal{S}_i(u)] - \mathbb{E}[\mathcal{S}_i(v)]|$ where $\mathbb{E}[\mathcal{S}_i(u)]$ is the expected outcome for the i^{th} task in the system $\mathcal{S}$ and where the expectation is over the randomness of the system and all its components.

3.1 Task-Competitive Composition

We now pose the relevant problem for multiple task fairness: competitive composition.

▶ **Definition 4** (Single Slot Composition Problem)**.** A (possibly randomized) system $\mathcal{S}$ is said to be a solution to the single slot composition problem for a set of k tasks $\mathcal{T}$ with metrics $\mathcal{D}_1, \ldots, \mathcal{D}_k$, if $\forall u \in U$, $\mathcal{S}$ assigns outputs for each task $\{x_{u,1}, \ldots, x_{u,k}\} \in \{0,1\}^k$ such that $\sum_{i \in [k]} x_{u,i} \leq 1$, and $\forall i \in [k]$, and $\forall\ u, v \in U$, $\mathcal{D}_i(u,v) \geq |\mathbb{E}[x_{u,i}] - \mathbb{E}[x_{v,i}]|$.

The single slot composition problem captures the scenario in which an advertising platform may have a single slot to show an ad but need not show any ad. Imagine that this advertising system only has two types of ads: those for jobs and those for household goods. If a person is qualified for jobs and eager and able to purchase household goods, the system must pick at most one of the ads to show. In this scenario, it may be unlikely that the advertising system would choose to show no ads, but the problem specification does not require that any positive outcome is chosen.

To solve the single-slot composition problem we must build a system which chooses at most one of the possible tasks so that fairness is preserved simultaneously for each task, across all elements in the universe. Clearly if classifiers for each task may *independently* and fairly assign outputs without interference, the system as a whole satisfies multiple task fairness. However, most systems will require trade-offs between tasks. Consider a naïve solution to the single-slot problem for ads: each advertiser chooses to bid on each person with some probability, and if both advertisers bid for the same person, the advertiser with the higher bid gets to show her ad. Formally, we define a tie-breaking function and Task-Competitive Composition:

▶ **Definition 5** (Tie-breaking Function)**.** A (possibly randomized) *tie-breaking function* $\mathbb{B} : U \times \{0,1\}^* \times \{0,1\}^k \to [k] \cup \{0\}$ takes as input an individual $w \in U$ and a $k-$bit string x_w and outputs the index of a "1" in x_w if such an index exists and 0 otherwise.

For notational convenience, in the case of two tasks T and T', we use $\mathbb{B}_w(T)$ to refer to the probability that $\mathbb{B}$ chooses task T for element w if both T and T' return positive classifications, and analogously define $\mathbb{B}_w(T')$.

▶ **Definition 6** (Task-Competitive Composition)**.** Consider a set $\mathcal{T}$ of k tasks, and a tie-breaking function as defined above. Given a set $\mathcal{C}$ of classifiers for the set of tasks, define $y_w = \{y_{w,1}, \ldots, y_{w,k}\}$ where $y_{w,i} = C_i(w)$. The task-competitive composition of the set $\mathcal{C}$ is defined as $y^*_w = \mathbb{B}(w, y_w)$ for all $w \in U$.

Definition 6 yields a system S defined by $S(w) = 0^k$ if $y_w = 0^k$ and $S(w) = e_{\mathbb{B}(w,y_w)}$ (the $\mathbb{B}(w, y_w)$ basis vector of dimension k) if $y_w \neq 0^k$. We evaluate its fairness by examining the Lipschitz requirements $|\Pr[y_u^* = i] - \Pr[y_v^* = i]| \leq \mathcal{D}_i(u, v)$ for all $u, v \in U$ and $i \in [k]$.

Task-competitive composition can reflect many scenarios other than advertising, which are discussed in greater detail in the full paper. Note that the tie-breaking function need not encode the same logic for all individuals and may be randomized. We start by introducing Lemma 7, which handles the simple case for a strict tie-breaking function for all individuals, and extend to all tie-breaking functions in Theorem 8.

▶ **Lemma 7.** *For any two tasks T and T' such that the metrics for each task ($\mathcal{D}$ and $\mathcal{D}'$ respectively) are not identical and are non-trivial*[7] *on a universe U, and if there is a strict preference for T, that is $\mathbb{B}_w(T) = 1\ \forall w \in U$, then there exists a pair of classifiers $\mathcal{C} = \{C, C'\}$ which are individually fair in isolation but when combined with task-competitive composition violate multiple task fairness.*

Proof. We construct a pair of classifiers $\mathcal{C} = \{C, C'\}$ which are individually fair in isolation for the tasks T and T', but do not satisfy multiple task fairness when combined with task-competitive composition with a strict preference for T for all $w \in U$. Task-competitive composition ensures that at most one task can be classified positively for each element, so our strategy is to construct C and C' such that the distance between a pair of individuals is stretched for the 'second' task.

By non-triviality of $\mathcal{D}$, there exist u, v such that $\mathcal{D}(u, v) \neq 0$. Fix such a pair u, v and let p_u denote the probability that C assigns 1 to u, and analogously p_v, p'_u, p'_v. We use these values as placeholders, and show how to set them to prove the lemma.

Because of the strict preference for T, the probabilities that u and v are assigned 1 for the task T' are

$$\Pr[\mathcal{S}(u)_{T'} = 1] = (1 - p_u)p'_u$$

$$\Pr[\mathcal{S}(v)_{T'} = 1] = (1 - p_v)p'_v$$

The difference between them is

$$\Pr[\mathcal{S}(u)_{T'} = 1] - \Pr[\mathcal{S}(v)_{T'} = 1] = (1 - p_u)p'_u - (1 - p_v)p'_v$$

$$= p'_u - p_u p'_u - p'_v + p_v p'_v$$

$$= p'_u - p'_v + p_v p'_v - p_u p'_u$$

Notice that if $\mathcal{D}'(u, v) = 0$, which implies that $p'_u = p'_v$, and $p_u \neq p_v$, then this quantity is non-zero, giving the desired contradiction for all fair C' and any C that assigns $p_u \neq p_v$, which can be constructed per Corollary 18.

However, if $\mathcal{D}'(u, v) \neq 0$, take C' such that $|p'_u - p'_v| = \mathcal{D}'(u, v)$ and denote the distance $|p'_u - p'_v| = m'$, and without loss of generality, assume that $p'_u > p'_v$ and $p_u < p_v$,

$$\Pr[\mathcal{S}(u)_{T'} = 1] - \Pr[\mathcal{S}(v)_{T'} = 1] = m' + p_v p'_v - p_u p'_u$$

Then to violate fairness for T', it suffices to show that $p_v p'_v > p_u p'_u$. Write $p_v = \alpha p_u$ where $\alpha > 1$,

$$\alpha p_u p'_v > p_u p'_u$$

[7] A metric $\mathcal{D}$ is said to be *non-trivial* if there exists at least one pair, $u, v \in U$ such that $\mathcal{D}(u, v) \notin \{0, 1\}$.

$$\alpha p'_v > p'_u$$

Thus it is sufficient to show that we can choose p_u, p_v such that $\alpha > \frac{p'_u}{p'_v}$. Constrained only by the requirements that $p_u < p_v$ and $|p_u - p_v| \leq \mathcal{D}(u, v)$, we may choose p_u, p_v to obtain an arbitrarily large $\alpha = \frac{p_v}{p_u}$ by Corollary 19. Thus there exist a pair of fair classifiers C, C' which when combined with strictly ordered task-competitive composition violate multiple task fairness. ◀

▶ **Theorem 8.** *For any two tasks T and T' with nontrivial metrics $\mathcal{D}$ and $\mathcal{D}'$ respectively, there exists a set $\mathcal{C}$ of classifiers which are individually fair in isolation but when combined with task-competitive composition violate multiple task fairness for any tie-breaking function.*

Proof. Consider a pair of classifiers C, C' for the two tasks. Let p_u denote the probability that C assigns 1 to u, and analogously let p_v, p'_u, p'_v denote this quantity for the other classifier and element combinations. As noted before, for convenience of notation, write $\mathbb{B}_u(T)$ to indicate the preference for each (element, outcome) pair, that is the probability that given the choice between T or the alternative outcome T', T is chosen. Note that in this system, for each element $\mathbb{B}_u(T) + \mathbb{B}_u(T') = 1$.

Note that if $\mathbb{B}_w(T) = 1\ \forall w \in U$ or $\mathbb{B}_w(T') = 1\ \forall w \in U$, the setting is exactly as described in Lemma 7. Thus we need only argue for the two following cases:

1. Case $\mathbb{B}_u(T) = \mathbb{B}_v(T) \neq 1$. We can write an expression for the probability that each element is assigned to task T:

   $$\Pr[\mathcal{S}(u)_T = 1] = p_u(1 - p'_u) + p_u p'_u \mathbb{B}_u(T)$$

   $$\Pr[\mathcal{S}(v)_T = 1] = p_v(1 - p'_v) + p_v p'_v \mathbb{B}_v(T)$$

   So the difference in probabilities is

   $$\Pr[\mathcal{S}(u)_T = 1] - \Pr[\mathcal{S}(v)_T = 1] = p_u(1 - p'_u) + p_u p'_u \mathbb{B}_u(T) - p_v(1 - p'_v) - p_v p'_v \mathbb{B}_v(T)$$

   $$= p_u - p_v + p_v p'_v - p_u p'_u + p_u p'_u \mathbb{B}_u(T) - p_v p'_v \mathbb{B}_v(T)$$

   $$= p_u - p_v + (p_v p'_v - p_u p'_u)(1 - \mathbb{B}_u(T))$$

   By our assumption that $\mathbb{B}_u(T) \neq 1$, we proceed analogously to the proof of Lemma 7 choosing C' such that $p_v p'_v > p_u p'_u$ and choosing C to ensure that $p_u - p_v = \mathcal{D}(u, v)$ to achieve unfairness for T.
2. Case $\mathbb{B}_u(T) \neq \mathbb{B}_v(T)$. Assume without loss of generality that $\mathbb{B}_u(T) \neq 1$. Recall the difference in probability of assignment of 1 for the first task in terms of $\mathbb{B}$:

   $$= p_u - p_v + p_v p'_v(1 - \mathbb{B}_v(T)) - p_u p'_u(1 - \mathbb{B}_u(T))$$

   Choose C such that $p_u - p_v = \mathcal{D}(u, v)$ (or if there is no such individually fair C, choose the individually fair C which maximizes the distance between u and v). So it suffices to show that we can select C' such that $p_v p'_v(1 - \mathbb{B}_v(T)) - p_u p'_u(1 - \mathbb{B}_u(T)) > 0$. As before, write $p_u = \alpha p_v$ where $\alpha > 1$. We require:

   $$p_v p'_v(1 - \mathbb{B}_v(T)) > \alpha p_v p'_u(1 - \mathbb{B}_u(T))$$

   $$p'_v(1 - \mathbb{B}_v(T)) > \alpha p'_u(1 - \mathbb{B}_u(T))$$

Writing $\beta = (1 - \mathbb{B}_v(T))/(1 - \mathbb{B}_u(T))$ (recall that $\mathbb{B}_u(T) \neq 1$ so there is no division by zero), we require

$$p'_v\beta > \alpha p'_u$$

$$\beta/\alpha > p'_u/p'_v$$

Constrained only by $|p'_u - p'_v| \leq \mathcal{D}'(u, v)$, we can choose p'_u, p'_v to be any arbitrary positive ratio per Corollary 19, thus we can select a satisfactory C' to exceed the allowed distance.

Thus we have shown that for the cases where the tie-breaking functions are identical for u and v and when the tie-breaking functions are different, there always exists a pair of classifiers C, C' which are fair in isolation, but when combined in task-competitive compositiondo not satisfy multiple task fairness which completes the proof. ◀

The intuition for unfairness in such a strictly ordered composition is that each task inflicts its preferences on subsequent tasks, and this intuition extends to more complicated tie-breaking functions and individuals with positive distances in both tasks. Our intuition suggests that the situation in Theorem 8 is not contrived and occurs often in practice, and moreover that small relaxations will not be sufficient to alleviate this problem, as the phenomenon has been observed empirically [3, 17, 15]. We include a small simulated example in the Appendix of the full version to illustrate the potential magnitude and frequency of such fairness violations.

3.2 Simple Fair Multiple-task Composition

Fortunately, there is a general purpose mechanism for the single slot composition problem which requires no additional information in learning each classifier and no additional coordination between the classifiers.[8] The rough procedure for RandomizeThenClassify (Algorithm 1) is to fix a fair classifier for each task, fix a probability distribution over the tasks, sample a task from the distribution, and then run the fair classifier for that task. RandomizeThenClassify has several nice properties: it requires no coordination in the training of the classifiers, it preserves the ordering and relative distance of elements by each classifier, and it can be implemented by a platform or other third party, rather than requiring the explicit cooperation of all classifiers. The primary downside of RandomizeThenClassify is that it reduces allocation (the total number of positive classifications) for classifiers trained with the expectation of being run independently.

4 Functional Composition

In *Functional Composition*, the outputs of multiple classifiers are combined through logical operations to produce a single output for a single task. A significant consideration in functional composition is determining which outcomes are relevant for fairness and at which point(s) fairness should be measured. For example, (possibly different) classifiers for admitting students to different colleges are composed to determine whether the student is accepted to at least one college. In this case, the function is "OR", the classifiers are for the same task, and hence conform to the same metric, and this is the same metric one might use for defining

[8] See section Appendix Section 6.4 in the full version for another mechanism which requires coordination between the classifiers.

fairness of the system as a whole. Alternatively, the system may compose the classifier for admission with the classifier for determining financial aid. In this case the function is "AND", the classifiers are for different tasks, with different metrics, and we may use scholastic ability or some other appropriate output metric for evaluating overall fairness of the system.

4.1 Same-task Functional Composition

In this section, we consider the motivating example of college admissions. When secondary school students apply for college admission, they usually apply to more than one institution to increase their odds of admission to at least one college. Consider a universe of students U applying to college in a particular year, each with intrinsic qualification $q_u \in [0, 1]$, $\forall u \in U$. We define $\mathcal{D}(u, v) = |q_u - q_v|$ $\forall u, v \in U$. $\mathcal{C}$ is the set of colleges and assume each college $C_i \in \mathcal{C}$ admits students fairly with respect to $\mathcal{D}$. The system of schools is considered OR-fair if the indicator variable x_u, which indicates whether or not student u is admitted to at least one school, satisfies individual fairness under this same metric. More formally,

▶ **Definition 9** (OR Fairness)**.** Given a (universe, task) pair with metric $\mathcal{D}$, and a set of classifiers $\mathcal{C}$ we define the indicator

$$x_u = \begin{cases} 1 \text{ if } \sum_{C_i \in \mathcal{C}} C_i(x) \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

which indicates whether at least one positive classification occurred. Define $\tilde{x}_u = \Pr[x_u = 1] = 1 - \prod_{C_i \in \mathcal{C}} (1 - \Pr[C_i(u) = 1])$. Then the composition of the set of classifiers $\mathcal{C}$ satisfies *OR Fairness* if $\mathcal{D}(u, v) \geq d(\tilde{x}_u, \tilde{x}_v)$ for all $u, v \in U$.

The OR Fairness setting matches well to tasks where individuals primarily *benefit* from one positive classification for a particular task.[9] As mentioned above, examples of such tasks include gaining access to credit or a home loan, admission to university, access to qualified legal representation, access to employment, *etc.*[10] Although in some cases more than one acceptance may have positive impact, for example a person with more than one job offer may use the second offer to negotiate a better salary, the core problem is (arguably) whether or not at least one job is acquired.

Returning to the example of college admissions, even with the strong assumption that each college fairly evaluates its applicants, there are still several potential sources of unfairness in the resulting system. In particular, if students apply to different numbers of colleges or colleges with different admission rates, we would expect that their probabilities of acceptance to at least one college will be different. The more subtle scenario from the perspective of composition is when students apply to the *same* set of colleges.

Even in this restricted setting, it is still possible for a set of classifiers for the same task to violate OR fairness. The key observation is that for elements with positive distance, the difference in their expectation of acceptance by at least one classifier does not diverge linearly in the number of classifiers included in the composition. As the number of classifiers increases, the probabilities of positive classification by at least one classifier for any pair eventually converge. However, in practice, we expect students to apply to perhaps five or 10 colleges, so it is desirable to characterize when small systems are robust to such composition.

---

[9] We may conversely define NOR Fairness to take $\neg x_u$, and this setting more naturally corresponds to cases where not being classified as positive is desirable.

[10] [1] considers what boils down to AND-fairness for Equal Opportunity and presents an excellent collection of evocative example scenarios.

▶ **Theorem 10.** *For any (universe, task) pair with a non-trivial metric $\mathcal{D}$, there exists a set of individually fair classifiers $\mathcal{C}$ which do not satisfy OR Fairness, even if each element in U is classified by all $C_i \in \mathcal{C}$.*

The proof of Theorem 10 follows from a straightforward analysis of the difference in probability of at least one positive classification.[11] The good news is that there exist non-trivial conditions for sets of small numbers of classifiers where OR Fairness is satisfied:

▶ **Lemma 11.** *Fix a set $\mathcal{C}$ of fair classifiers, and let x_w for $w \in U$ be the indicator variable as in Definition 9. If $\mathbb{E}[x_w] \geq 1/2$ for all $w \in U$, then the set of classifiers $\mathcal{C} \cup \{C'\}$ satisfies OR fairness if C' satisfies individual fairness under the same metric and $\Pr[C'(w) = 1] \geq \frac{1}{2}$ for all $w \in U$.*

This lemma is useful for determining that a system is free from same-task divergence, as it is possible to reason about an "OR of ORs", and more generally an "OR" of any fair components of sufficient weight.

Functional composition can also be used to reason about settings where classification procedures for different tasks are used to determine the outcome for a single task. For example, in order to attend a particular college, a student must be admitted *and* receive sufficient financial aid to afford tuition and living expenses. Financial need and academic qualification clearly have different metrics, and in such settings, a significant challenge is to understand how the input metrics relate to the relevant output metric. Without careful reasoning about the interaction between these tasks, it is very easy to end up with systems which violate individual fairness, even if they are constructed from individually fair components. (See Section 4.2 in the full version for more details.)

5 Dependent Composition

Thus far, we have restricted our attention to the mode of operation in which classifiers act on the entire universe of individuals at once and each individual's outcome is decided independently. In practice, however, this is an unlikely scenario, as classifiers may be acting as a selection mechanism for a fixed number of elements, may operate on elements in arbitrary order, or may operate on only a subset of the universe. In this section, we consider the case in which the classification outcomes received by individuals are not independent. Slightly abusing the term "composition," these problems can be viewed as a composition of the classifications of elements of the universe. We roughly divide these topics into Cohort Selection problems, when a set of exactly n individuals must be selected from the universe, and Universe Subset problems, when only a subset of the relevant universe for the task is under the influence of the classifier we wish to analyze or construct. Within these two problems we consider several relevant settings:

Online versus offline: Advertising decisions for online ads must be made immediately upon impression and employers must render employment decisions quickly or risk losing out on potential employees or taking too long to fill a position.

Random versus adversarial ordering: The order in which individuals apply for an open job may be influenced by their social connections with existing employees, which impacts how quickly they hear about the job opening.

[11] See Appendix Section 4 in the full version for the complete proof.

Known versus unknown subset or universe size: An advertiser may know the average number of interested individuals who visit a website on a particular day, but be uncertain on any particular day of the exact number.

Constrained versus unconstrained selection: In many settings there are arbitrary constraints placed on selection of individuals for a task which are unrelated to the qualification or metric for that task. For example, to cover operating costs, a college may need at least $n/2$ of the n students in a class to be able to pay full tuition.

In dependent composition problems, it is important, when computing distances between distributions over outcomes, to pay careful attention to the source of randomness. Taking inspiration from the experiment setup found in many cryptographic definitions, we formally define two problems, Universe Subset Classification and Cohort Selection, (included in Definitions 13 and 14 in the Appendix). In particular, it is important to understand the randomness used to decide an ordering or a subset, as once an ordering or subset is fixed, reasoning about fairness is impossible, as a particular individual may be arbitrarily included or excluded.

5.1 Basic Offline Cohort Selection

First we consider the simplest version of the cohort selection problem: choosing a cohort of n individuals from the universe U when the entire universe is known and decisions are made offline. A simple solution is to choose a permutation of the elements in U uniformly at random, and then apply a fair classifier C until n are selected or selecting the last few elements from the end of the list if n have not yet been selected. With some careful bookkeeping, we show that this mechanism is individually fair for any individually fair input classifier. (See Algorithms 2 and 3 in the Appendix below; a complete analysis is included in Appendix Section 6 in the full version.)

5.2 More complicated settings

In this extended abstract, we omit a full discussion of the more complicated dependent composition scenarios, but briefly summarize several settings to build intuition.

▶ **Theorem 12.** *If the ordering of the stream is adversarial, but $|U|$ is unknown, then there exists no solution to the online cohort selection problem.*

The intuition for the proof follows from imagining that a fair classification process exists for an ordering of size n and realizing that this precludes fair classification of a list of size $n + 1$, as the classification procedure cannot distinguish between the two cases.

Constrained cohort selection

Next we consider the problem of selecting a cohort with an external requirement that some fraction of the selected set is from a particular subgroup. That is, given a universe U, and $p \in [0, 1]$, and a subset $A \subset U$, select a cohort of n elements such that at least a p fraction of the elements selected are in A. This problem captures situations in which external requirements cannot be ignored. For example, if a certain budget must be met, and only some members of the universe contribute to the budget, or if legally a certain fraction of people selected must meet some criterion (as in, demographic parity). In the full version, we characterize a broad range of settings where the constrained cohort selection problem cannot be solved fairly.

To build intuition, suppose the universe U is partitioned into sets A and B, where $n/2 = |A| = |B|/5$. Suppose further that the populations have the same distribution on ability, so that the set B is a "blown up" version of A, meaning that for each element $u \in A$ there are 5 corresponding elements $V_u = \{v_{u,1}, ..., v_{u,5}\}$ such that $\mathcal{D}(u, v_{u,i}) = 0$, $1 \leq i \leq 5$, $\forall u, u' \in A$ $V_u \cap V_{u'} = \emptyset$, and $B = \cup_{u \in A} V_u$. Let $p = \frac{1}{2}$. The constraint requires all of A to be selected; that is, each element of A has probability 1 of selection. In contrast, the average probability of selection for an element of B is $\frac{1}{5}$. Therefore, there exists $v \in B$ with selection probability at most 1/5. Letting $u \in A$ such that $v \in V_u$, we have $\mathcal{D}(u, v) = 0$ but the difference in probability of selection is at least $\frac{4}{5}$. We give a more complete characterization of the problem and impossibilities in the full version in Appendix Section 6.3 .

6 Extensions to Group Fairness

In general, the results discussed above for composition of individual fairness extend to group fairness definitions; however, there are several issues and technicalities unique to group fairness definitions which we now discuss.

Technicalities

Consider the following simple universe: for a particular $z \in \mathcal{Z}$, group B is unimodal, having only elements with medium qualification q_m, while group A is bimodal, with half of its elements having low qualification q_l and half having high qualification q_h. Choosing $p_h = 1$, $p_m = .75$, and $p_l = .5$ satisfies Conditional Parity for a single application. However, for the OR of two applications, the squares diverge ($.9375 \neq .875$), violating conditional parity (see Figure 1). Note, however, that all of the individuals with $\mathbf{z} = z$ have been drawn closer together under composition, and none have been pulled further apart. This simple observation implies that in some cases we may observe failures under composition for conditional parity, even when individual fairness is satisfied. In order to satisfy Conditional Parity under OR-composition, the classifier could sacrifice accuracy by treating all individuals with $\mathbf{z} = z$ equally. However, this necessarily discards useful information about the individuals in A to satisfy a technicality.

Subgroup Subtleties

There are many cases where failing to satisfy conditional parity under task-competitive composition is clearly a violation of our intuitive notion of group fairness. However, conditional parity is not always a reliable test for fairness at the subgroup level under composition. In general, we expect conditional parity based definitions of group fairness to detect unfairness in multiple task compositions reasonably well when there is an obvious interaction between protected groups and task qualification, as observed empirically in [17] and [3]. For example, let's return to our advertising example where home-goods advertisers have no protected set, but high-paying jobs have gender as a protected attribute. Under composition, home-goods out-bidding high-paying jobs ads for women will clearly violate the conditional parity condition for the job ads (see Figure 2).

However, suppose that, in response to gender disparity caused by task-competitive composition, classifiers iteratively adjust their bids to try to achieve Conditional Parity. This may cause them to *learn themselves* into a state that satisfies Conditional Parity with respect to gender, but behaves poorly for a socially meaningful subgroup (see Figure 3.) For example, if home goods advertisers aggressively advertise to women who are new parents

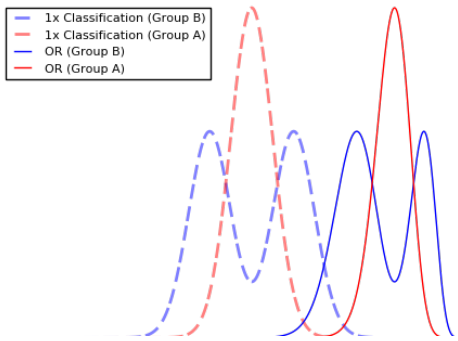


Figure 1 An illustration of the shift in groups from a single classification to the OR of two applications of the same classifier. Although the two groups originally had the same mean probability of positive classification, this breaks down under OR composition.

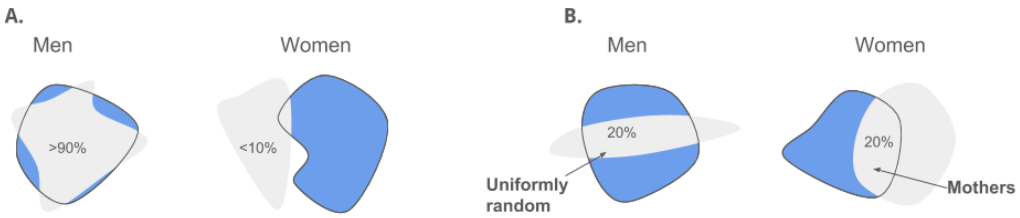
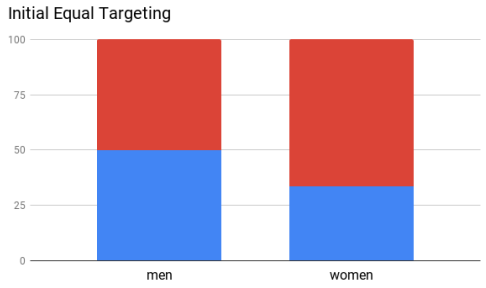


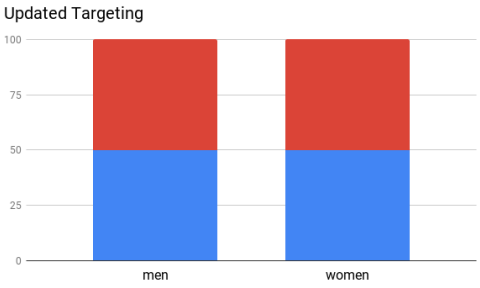
Figure 2 A. When the two tasks are related, one will 'claim' a larger fraction of one gender than another, leading to a smaller fraction of men remaining for classification in the other task (shown in blue). Conditional parity will detect this unfairness. **B.** When the tasks are unrelated, one task may 'claim' the same fraction of people in each group, but potentially select a socially meaningful subgroup, eg parents. Conditional parity will fail to detect this subgroup unfairness, unless subgroups, including any subgroups targeted by classifiers composed with, are explicitly accounted for.

(because their life-time value ($\mathcal{Z}$) to the advertiser is the highest of all universe elements), then a competing advertiser for jobs, noticing that its usual strategy of recruiting all people with skill level $\mathbf{z}' = z'$ equally is failing to reach enough women, bids more aggressively on women. By bidding more aggressively, the advertiser increases the probability of showing ads to women (for example by outbidding low-value competition), but not to women who are bid for by the home goods advertiser (a high-value competitor), resulting in a high concentration of ads for women who are *not* mothers, while still failing to reach women who *are* mothers. Furthermore, the systematic exclusion of mothers from job advertisements can, over time, be even more problematic, as it may contribute to the stalling of careers. In this case, the system discriminates against mothers without necessarily discriminating against fathers.

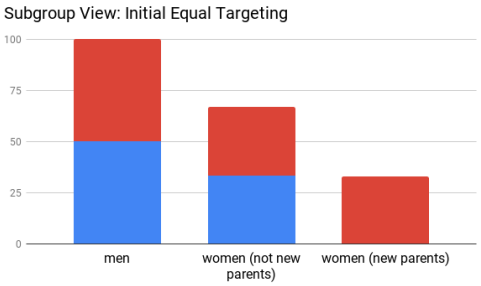
Although problematic (large) subgroup semantics are part of the motivation for [11, 7] and exclusion of subgroups is not only a composition problem, the added danger in composition is that the features describing this subset may be missing from the feature set of the jobs classifier, rendering the protections proposed in [11] and [7] ineffective. In particular, we expect that sensitive attributes like parental status are unlikely to appear (or are illegal to collect) in employment-related training or testing datasets, but may be legitimately targeted by other competing advertisers.



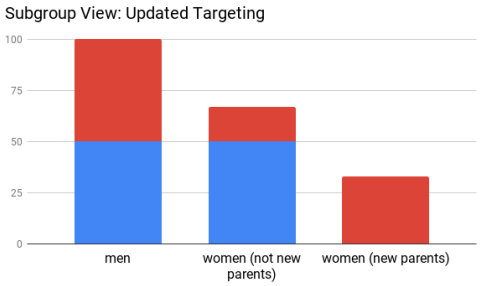
(a) Initial equal targeting of qualified men and women results in violation of conditional parity, as there are unequal rates of ads shown (blue).



(b) By increasing the targeting of women, the jobs advertiser "fixes" conditional parity at the coarse group level.



(c) At the subgroup level, it's clear that the lack of conditional parity is due to "losing" all of the new parent women to the home-goods advertiser.



(d) New targeting strategy increases ads shown to non new-parent women, but continues to exclude new parent women.

Figure 3 Home-goods advertisers aggressively target mothers, out-bidding the jobs advertiser. When the jobs advertiser bids more aggressively on "women" (b) the overall rate of ads shown to "women" increases, but mothers may still be excluded (d), so Pr[ad |qualified, woman] > Pr[ad | qualified, mother].

References

1. Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair Pipelines. *CoRR*, abs/1707.00391, 2017. `arXiv:1707.00391`.
2. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint*, 2017. `arXiv:1703.00056`.
3. Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.
4. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.
5. Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online Learning with an Unknown Fairness Metric. *arXiv preprint*, 2018. `arXiv:1802.06936`.
6. Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.
7. Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (Computationally-Identifiable) Masses. *arXiv preprint*, 2017. `arXiv:1711.08513`.

8 Lily Hu and Yiling Chen. Fairness at Equilibrium in the Labor Market. *CoRR*, abs/1707.01590, 2017. arXiv:1707.01590.

9 Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

10 Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 643–650. IEEE, 2011.

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

12 Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. *arXiv preprint*, 2017. arXiv:1706.02744.

13 Michael P Kim, Omer Reingold, and Guy N Rothblum. Fairness Through Computationally-Bounded Awareness. *arXiv preprint*, 2018. arXiv:1803.03239.

14 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR*, abs/1609.05807, 2016. arXiv:1609.05807.

15 Peter Kuhn and Kailing Shen. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics*, 128(1):287–336, 2012.

16 Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. *arXiv preprint*, 2017. arXiv:1703.06856.

17 Anja Lambrecht and Catherine E Tucker. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads, 2016.

18 Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. *arXiv preprint*, 2018. arXiv:1803.04383.

19 David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. *arXiv preprint*, 2018. arXiv:1802.06309.

20 Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

21 Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint*, 2017. arXiv:1706.08519.

22 Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

A Appendix

A.1 Algorithm for Task-Competitive Composition

RandomizeThenClassify, Algorithm 1 has several nice properties. First, it requires no coordination in the training of the classifiers. In particular, it does not require any sharing of objective functions. Second, it preserves the ordering of elements by each classifier. That is, if $\Pr[C_i(u) = 1] > \Pr[C_i(v) = 1]$ then $\Pr[\mathsf{RandomizeThenClassify}(u)_i = 1] > \Pr[\mathsf{RandomizeThenClassify}(v)_i = 1]$. Finally, it can be implemented by a platform or other third party, rather than requiring the explicit cooperation of all classifiers. The primary downside of RandomizeThenClassify is that it drastically reduces allocation (the total number of positive classifications) for classifiers trained with the expectation of being run independently.

Algorithm 1 RandomizeThenClassify.

Input: universe element $u \in U$, set of fair classifiers $\mathcal{C}$ (possibly for distinct tasks) operating on U, probability distribution over tasks $\mathcal{X} \in \Delta(\mathcal{C})$
$x \leftarrow 0^{|\mathcal{C}|}$
$C_t \sim \mathcal{X}$
if $C_t(u) = 1$ **then**
  $x_t = 1$
end if
return x

Algorithm 2 PermuteThenClassify.

Input: $n \leftarrow$ the number of elements to select
$C \leftarrow$ a classifier $C : U \times \{0,1\}^* \rightarrow \{0,1\}$
$\pi \sim S_{|U|}$ a random permutation from the symmetric group on $|U|$
$L \leftarrow \pi(U)$ An ordered set of elements
$M \leftarrow \emptyset$
while $|M| < n$: **do**
  $u \leftarrow pop(L)$
  if $C(u) = 1$ **then**
    $M \leftarrow M \cup \{u\}$
  end if
  if $n - |M| \geq |L|$ **then**
    // the end condition
    $M \leftarrow M \cup \{u\}$
  end if
end while
return M

A.2 Algorithms for Cohort Selection

PermuteThenClassify, Algorithm 2, works through a list initialized to a random permutation $\pi(U)$, classifying elements one at a time and independently until either (1) n elements have been selected or (2) the number of remaining elements in the list equals the number of remaining spots to be filled. Case (2) is referred to as the "end condition". Elements in the "end condition" are selected with probability 1.

WeightedSampling, Algorithm 3, chooses sets of elements with probability proportional to their weight under a fair classifier. This prevents the arbitrary behavior of the end condition in case the classifier is poorly tuned for the specific number of desired elements.

A.3 Universe Subset Problems

▶ **Definition 13** (Universe Subset Classification Problem)**.** Given a universe U, let $\mathcal{Y}$ be a distribution over subsets of U. Let $\mathcal{X} = \{\mathcal{X}(V)\}_{V \subseteq U}$ be a family of distributions, one for each subset of U, where $\mathcal{X}(V)$ is a distribution on permutations of the elements of V. Let $\Pi(2^U)$ denote the set of permutations on subsets of U. Formally, for a system $\mathcal{S} : \Pi(2^U) \times \{0,1\}^* \rightarrow U^*$, we define **Experiment**$(\mathcal{S}, \mathcal{X}, \mathcal{Y}, u)$ as follows:

1. Choose $r \sim \{0,1\}^*$
2. Choose $V \sim \mathcal{Y}$

Algorithm 3 WeightedSampling.

Input: $n \leftarrow$ the number of elements to select
$C \leftarrow$ a classifier $C : U \times r \to \{0, 1\}$
$L \leftarrow$ the set of all subsets of U of size n
for $l \in L$ **do**
  $w(l) \leftarrow \sum_{u \in l} \mathbb{E}[C(u)]$ // set the weight of each set
  Define $\mathcal{X} \in \Delta(L)$ such that $\forall l \in L$, the weight of l under $\mathcal{X}$ is $\frac{w(l)}{\sum_{l' \in L} w(l')}$
  $M \sim \mathcal{X}$ // Sample a set of size n according to $\mathcal{X}$
end for
return M

3. Choose $\pi \sim \mathcal{X}(V)$
4. Run $\mathcal{S}$ on π with randomness r, and output 1 if u is selected (positively classified).

The system $\mathcal{S}$ is individually fair and a solution to the Universe Subset Classification Problem for a particular $(\mathcal{X}, \mathcal{Y})$ pair if for all $u, v \in U$,

$$|\mathbb{E}[\mathsf{Experiment}(\mathcal{S}, \mathcal{X}, \mathcal{Y}, u)] - \mathbb{E}[\mathsf{Experiment}(\mathcal{S}, \mathcal{X}, \mathcal{Y}, v)]| \leq \mathcal{D}(u, v)$$

Note that for any distinct individuals $u, v \in U$, in any given run of the experiment V may contain u, v, neither or both.

▶ **Definition 14** (Cohort Selection Problem)**.** The Cohort Selection Problem is identical to the Universe Subset Classification Problem, except the system is limited to choosing exactly n individuals.

▶ **Lemma 15.** *Given an instance of the universe subset classification problem (Definition 13) where $\mathcal{Y}$ assigns positive weight to all elements $w \in U$, the following procedure applied to any individually fair classifier C which solely controls outcomes for a particular task will result in fair classification under the input distribution $\mathcal{Y}$.*
Procedure: *for each $w \in U$, let q_w denote the probability that w appears in V. Let $q_{min} = \min_w q_w$. For each element $w \in V$, with probability q_{min}/q_w classify w normally, otherwise output the default for no classification.*

Proof. Let $u = \text{argmin}_w(q_w)$. Then u will be classified positively with probability $p_u q_{min}$ where probability is taken over $\mathcal{Y}$ and C. All other elements $v \in V$ will be classified positively with probability $q_v(q_{min}/q_v)p_v = p_v q_{min}$. As positive classification by C is the only way to get a positive outcome for the task, reasoning about $|p_v - p_u|$ is sufficient to ensure fairness. Therefore, if $|p_v - p_u| \leq \mathcal{D}(u, v)$, then the distance under this procedure is also $\leq \mathcal{D}(u, v)$. ◀

A.4 Construction of Fair Classifiers

▶ **Lemma 16.** *Let V be a (possibly empty) subset of U. If there exists a classifier $C : V \times \{0, 1\}^* \to \{0, 1\}$ such that $\mathcal{D}(u, v) \geq d(\tilde{C}(u), \tilde{C}(v))$ for all $u, v \in V$, then for any $x \in U \backslash V$ there exists classifier $C' : V \cup \{x\} \times \{0, 1\}^* \to \{0, 1\}$ such that $\mathcal{D}(u, v) \geq d(\tilde{C}(u), \tilde{C}(v))$ for all $u, v \in U$, which has identical behavior to C on V.*

Proof. For $V = \emptyset$, any value p_x suffices to fairly classify x. For $|V| = 1$, choosing any p_x such that $|p_v - p_x| \leq \mathcal{D}(v, x)$ for $v \in V$ suffices.

Algorithm 4 FairAddition$(\mathcal{D}, V, p_t, C, x)$.

Input: metric $\mathcal{D}$ for universe U, a subset $V \subset U$, target probability p_t, an individually fair classifier $C : V \times \{0,1\}^* \to \{0,1\}$, a target element $x \in U \backslash V$ to be added to C.
Initialize $L \leftarrow V$
$\hat{p}_x \leftarrow p_t$
for $l \in L$ **do**
  $dist \leftarrow \mathcal{D}(l, x)$
  if $dist < p_l - \hat{p}_x$ **then**
    $\hat{p}_x \leftarrow p_l - dist$
  else if $dist < \hat{p}_x - p_l$ **then**
    $\hat{p}_x \leftarrow p_l + dist$
  end if
end for
return $\hat{p}_x$

For $|V| \geq 2$, apply the procedure outlined in Algorithm 4 taking p_t to be the probability of positive classification of x's nearest neighbor in V under C. As usual, we take p_w to be the probability that C positively classifies element w.

Notice that Algorithm 4 only modifies $\hat{p}_x$, and that $\hat{p}_x$ is only changed if a distance constraint is violated. Thus it is sufficient to confirm that on each modification to $\hat{p}_x$, no distance constraints between x and elements in the opposite direction of the move are violated.

Without loss of generality, assume that $\hat{p}_x$ is decreased to move within an acceptable distance of u, that is $\hat{p}_x \geq p_u$. It is sufficient to show that for all v such that $p_v > \hat{p}_x$ that no distances are violated. Consider any such v. By construction $\hat{p}_x - p_u = \mathcal{D}(u, x)$, and $p_v - p_u \leq \mathcal{D}(u, v)$. From triangle inequality, we also have that $\mathcal{D}(u, v) \leq \mathcal{D}(u, x) + \mathcal{D}(x, v)$. Substituting, and using that $p_v \geq \hat{p}_x \geq p_u$:

$$\mathcal{D}(u, v) \leq \mathcal{D}(u, x) + \mathcal{D}(x, v)$$

$$\mathcal{D}(u, v) - \mathcal{D}(u, x) \leq \mathcal{D}(x, v)$$

$$\mathcal{D}(u, v) - (\hat{p}_x - p_u) \leq \mathcal{D}(x, v)$$

$$(p_v - p_u) - (\hat{p}_x - p_u) \leq \mathcal{D}(u, v) - (\hat{p}_x - p_u) \leq \mathcal{D}(x, v)$$

$$p_v - \hat{p}_x \leq \mathcal{D}(x, v)$$

Thus the fairness constraint for x and v is satisfied, and C' is an individually fair classifier for $V \cup \{x\}$. ◀

Lemma 16 allows us to build up a fair classifier in time $O(|U|^2)$ from scratch, or to add to an existing fair classifier for a subset. We state several useful corollaries:

▶ **Corollary 17.** *Given a subset $V \subset U$ and a classifier $C : V \times \{0,1\}^* \to \{0,1\}$ such that $\mathcal{D}(u, v) \geq d(\tilde{C}(u), \tilde{C}(v))$ for all $u, v \in V$, there exists an individually fair classifier $C' : U \times \{0,1\}^* \to \{0,1\}$ which is individually fair for all elements $u, v \in U$ and has identical behavior to C on V.*

Corollary 17 follows immediately from applying Algorithm 4 to each element of $U \backslash V$ in arbitrary order.

▶ **Corollary 18.** *Given a metric $\mathcal{D}$, for any pair $u, v \in U$, there exists an individually fair classifier $C : U \times \{0,1\}^* \to \{0,1\}$ such that $d(\tilde{C}(u), \tilde{C}(v)) = \mathcal{D}(u,v)$.*

Corollary 18 follows simply from starting from the classifier which is fair only for a particular pair and places them at their maximum distance under $\mathcal{D}$ and then repeatedly applying Algorithm 4 to the remaining elements of U. From a distance preservation perspective, this is important; if there is a particular 'axis' within the metric where distance preservation is most important, then maximizing the distance between the extremes of that axis can be very helpful for preserving the most relevant distances.

▶ **Corollary 19.** *Given a metric $\mathcal{D}$ and $\alpha \in \mathbb{R}^+$, for any pair $u, v \in U$, there exists an individually fair classifier $C : U \times \{0,1\}^* \to \{0,1\}$ such that $p_u/p_v = \alpha$, where $p_u = \mathbb{E}[C(u)]$ and likewise $p_v = \mathbb{E}[C(v)]$.*

Corollary 19 follows from choosing $p_u/p_v = \alpha$ without regard for the difference between p_u and p_v, and then adjusting. Take $\beta|p_v - p_u| = \mathcal{D}(u,v)$, and choose $\hat{p}_u = \beta p_u$ and $\hat{p}_v = \beta p_v$ so that $|\beta p_v - \beta p_u| = \beta|p_v - p_u| \leq \mathcal{D}(u,v)$, but the ratio $\frac{\beta p_u}{\beta p_v} = \frac{p_u}{p_v} = \alpha$ remains unchanged.