Individual Fairness Under Composition

Cynthia Dwork [*1 2 3] **Christina Ilvento** [*1 4]

Abstract

Much of the literature on fairness in classification considers the case of a single classifier used once, in isolation. However, fairness is not a property of isolated classifiers alone, it is also a property of classifiers composed into systems, and systems must handle repeated use and interaction with other fair (or unfair) systems. We initiate the study of composition of individually fair classifiers proposed in [Dwork, Hardt, Pitassi, Reingold, Zemel, 2011]. We find that fairness does not behave well under composition and propose directions to remedy the situation.

1. Introduction

The increasing reach of algorithmic classification tools into our everyday lives has given rise to an explosion of research on algorithmic fairness (Pedreshi et al., 2008; Kamiran & Calders, 2009; Kamishima et al., 2011; Dwork et al., 2011; Zemel et al., 2013; Edwards & Storkey, 2015; Datta et al., 2015; Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2016; Joseph et al., 2016; Kusner et al., 2017; Nabi & Shpitser, 2017; Kilbertus et al., 2017; Hébert-Johnson et al., 2017; Hu & Chen, 2017; Kearns et al., 2017; Gillen et al., 2018; Kim et al., 2018; Liu et al., 2018). A great deal of attention has been paid to the relative strengths and weaknesses of particular fairness definitions in isolation, but the construction of fair systems from fair classifiers has been neglected: If all colleges fairly admit students, does it follow that the college admissions *system* is fair? If all advertisers on an advertising platform fairly determine to whom to show ads, is the advertising *system*, in which advertisers compete to show their ads, fair? If we use a fair classifier to select a fixed number n of students for admission to a college class, will the class-formation *system* be fair?

Our starting point is the definition of Individual Fairness from (Dwork et al., 2011). Speaking intuitively, for any given classification task T – for example, deciding whether or not to display a specific advertisement – the definition assumes the existence of a *task-specific* metric $\mathcal{D}$ determining, for any two individuals, how (dis)similar they are for the given task T.[1] The requirement is that people who are similar (as measured by $\mathcal{D}$) should be treated similarly:

Definition 1 (Individual Fairness (Dwork et al., 2011))**.** Given a universe of individuals U, and a metric $\mathcal{D}$ for a classification task T with outcome set O, a randomized classifier $C : U \times \{0,1\}^* \to O$, such that $\tilde{C} : U \to \Delta(O)$, and a distance measure $d : \Delta(O) \times \Delta(O) \to R$, C is *individually fair* if and only if for all $u, v \in U$, $\mathcal{D}(u, v) \leq d(\tilde{C}(u), \tilde{C}(v))$.[2]

Individual Fairness has a flavor similar to that of differential privacy (Dwork, 2006; Dwork et al., 2006), and indeed differentially private algorithms can sometimes be used to ensure Individual Fairness (Dwork et al., 2011). Unfortunately, in many real-life settings the fairness *goals* of system with multiple (fair or unfair) parts are idiosyncratic, and the analogy to differential privacy breaks down.

The remainder of this paper is organized as follows. To build intuition we begin with a simplified treatment of the case in which two or more advertising tasks "compete" for a user's attention (Section 2). Next, we consider functional composition and cohort selection, a setting in which decisions cannot be made independently. We enumerate a number of axes under which selection problems can vary, provide a general fairness definition that captures (most of) these, and highlight a few of our results for these variants (Section 3). A companion paper studies the analogous problems for group fairness (Dwork & Ilvento, 2018b)[3].

[1]Harvard John A Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. [2]Radcliffe Institute for Advanced Study, Cambridge, Massachusetts, 02138. [3]This work was supported in part by Microsoft Research and the Sloan Foundation. [4]This work was supported in part by the Smith Family Fellowship and Microsoft Research. Correspondence to: Cynthia Dwork <dwork@seas.harvard.edu>, Christina Ilvento <cilvento@g.harvard.edu>.

[1]We assume distances are scaled such that for all $u, v \in U$, $\mathcal{D}(u, v) \leq 1$.

[2]$\Delta(O)$ is the set of probability distributions on the set O of outputs.

[3]The full version of this paper includes complete proofs and discussion of the analogous problems for group fairness and additional results (Dwork & Ilvento, 2018a).

2. Task-Competitive Composition

Many systems require a trade off between tasks. For example, a website may only have one slot in which to display an advertisement, and multiple advertisers – say, a job advertiser and an advertiser for households goods – must compete for the slot[4]. If a person is qualified for jobs *and* wants to purchase household goods, the advertisement system must pick at most one of the ads to show. In this scenario, it is unlikely that the advertising system would choose to show no ads, but our problem definition will admit this option.

Definition 2 (Multiple-Task Fairness for the Single Slot Composition Problem)**.** Let $\mathcal{U}$ denote the universe of individuals. A (possibly randomized) system $\mathcal{S}$ is said to be a solution to the single slot composition problem for a set $\mathcal{T} = \{T_1, \ldots, T_k\}$ of k tasks, with metrics $\mathcal{D}_1, \ldots, \mathcal{D}_k$, if for all $u \in \mathcal{U}$, $\mathcal{S}$ 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 for all $i \in [k]$ and all $u, v \in U$ $\mathcal{D}_i(u,v) \geq |\mathbb{E}[x_{u,i}] - \mathbb{E}[x_{v,i}]|$, where the expectation is over the randomness of the system and all its components.

In this brief abstract we restrict our attention to the simple case in which there are only two tasks T and T', they are combined in a system $\mathcal{S}$ in which "ties" are broken in favor of T. In an ad setting, for example, the advertiser corresponding to task T might consistently outbid the advertiser corresponding to task T'. The results extend to more general, and more individualized, tie breaking procedures.

Lemma 1. *Consider 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 on a universe U[5]. Then there exists a pair of classifiers $\mathcal{C} = \{C, C'\}$ which are individually fair in isolation for tasks T and T', respectively, but the system $\mathcal{S}$ which breaks all "ties" in favor of T violates multiple task fairness (Definition 2).*

Proof. (Sketch) 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 based on properties of fairness for the system. Since ties are broken in favor of T for all elements of U, $\mathcal{S}$ positively classifies u for T', denoted $\mathcal{S}(u)_{T'} = 1$, only if u is not positively classified by C and is positively classified by C'. Thus, $\Pr[\mathcal{S}(u)_{T'} = 1] = (1 - p_u)p'_u$ and similarly $\Pr[\mathcal{S}(v)_{T'} = 1] = (1 - p_v)p'_v$. Thus, $\Pr[\mathcal{S}(u)_{T'} = 1] - \Pr[\mathcal{S}(v)_{T'} = 1] = p'_u - p_u p'_u - p'_v + p_v p'_v = p'_u - p'_v + p_v p'_v - p_u p'_u$. 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 valid C' and any C that assigns $0 < p_u < p_v$.

If $\mathcal{D}'(u,v) \neq 0$, take C' such that $|p'_u - p'_v| = \mathcal{D}'(u,v)$, and without loss of generality, assume that $p'_u > p'_v > 0$ and $p_u < p_v$. then $\Pr[\mathcal{S}(u)_{T'} = 1] - \Pr[\mathcal{S}(v)_{T'} = 1] = \mathcal{D}'(u,v) + p_v p'_v - p_u p'_u$ and 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 $\alpha > 1$. Then $p_v p'_v - p_u p'_u > 0 \alpha p_u p'_v > p_u p'_u \alpha p'_v > p'_u$. 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 $\alpha = \frac{p_v}{p_u}$.[6]

□

Lemma 1 extends to the more general setting, in which $|U| > 2$ and there need not be a strict preference among tasks, no matter what method may be employed for breaking ties when an individual is qualified for more than one task, even if the method may depend on the identity of the individual.

Theorem 2. *For any two tasks T and T' such that the metrics for each task ($\mathcal{D}$ and $\mathcal{D}'$ respectively) are non-trivial on a universe U, there exist set of classifiers $\mathcal{C}$ which are individually fair in isolation but when combined with task-competitive composition violate multiple task fairness for any tie-breaking method.*

The proof of Theorem 2 extends, in many natural cases, to modest multiplicative and additive relaxations of the fairness constraint for T'. To build some intuition, suppose that for task T a pair of individuals have $\mathcal{D}(u,v) = \frac{1}{2}$, but for task T' they are equivalent. Then any classifier that uses the allowed distance for T will inflict this distance on the task T', resulting in an arbitrarily large multiplicative increase in distance.

2.1. Simple Fair Task-Competitive Composition

Fortunately, in some situations there is a general purpose mechanism for the single slot composition problem. Algorithm 1 only requires a set $\mathcal{C}$ of fair classifiers (for possibly distinct tasks).

Theorem 3. *For any set of k tasks $\mathcal{T}$ with metrics $\mathcal{D}_1, \ldots, \mathcal{D}_k$, the system $\mathcal{S}$ described in Algorithm 1 (*RandomizeThenClassify *) achieves multi-task fairness for the single slot composition problem given any set of classifiers $\mathcal{C}$ for the tasks which are individually fair in isolation.*

RandomizeThenClassify requires no coordination in the training of the classifiers, nor does it require any sharing of objective functions. It preserves the ordering

[4] The problem extends naturally to any number k of slots.

[5] A metric $\mathcal{D}$ over U is non-trivial if there exist $u, v \in U$ such that $1 > \mathcal{D}(u,v) \neq 0$.

[6] The full proof addresses satisfying the fairness constraints imposed by elements $w \notin \{u, v\}$.

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 $\mathcal{X} = \Delta([|\mathcal{C}|])$
$x \leftarrow 0^{|\mathcal{C}|}$
$t \sim \mathcal{X}$
if $C_t(u) = 1$ **then**
  $x_t = 1$
end if
return x

of elements by each classifier: 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 reduces allocation (the total number of positive classifications). For example, in an advertising platform it would result in showing fewer ads overall.

Remark 1. We primarily consider the case of honest designers with good intentions. However, failing to enforce multiple-task fairness allows for a significant expansion of the "catalog of evils" outlined in (Dwork et al., 2011). For example, let us assume that more women than men emphasize team work and organizational skills on their resumes. An employer seeking to hire more men than women for a technical role could aggressively advertise a second role in teamwork management (for which there is only one opening) for which many women will be qualified in order to prevent women from seeing the more desirable technical position ad.

3. Functional Composition and Set Selection

In *Functional Composition*, multiple classifiers are combined through logical functions to produce a single output for a single task. 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 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. We briefly summarize results for OR-fairness.

Definition 3 (OR-Fairness)**.** Given a (universe, task) pair with metric $\mathcal{D}$, a set of classifiers $\mathcal{C}$ satisfies *OR-Fairness* if the indicator variable defined by $x_u = 1$ if $\sum_{C_i \in \mathcal{C}} C_i(x) \geq 1$ and 0 otherwise satisfies $\mathcal{D}(u, v) \geq d(x_u, x_v)$ for all $u, v \in U$.

Clearly, students who are able to apply to more colleges (due to being able to afford the application fees, being the recipient of better college counseling, or having more time to spend on applications than their peers who need to work to support the family), have improved chances of being admitted to at least one college over those of equally qualified ($\mathcal{D}(u, v) = 0$) students who are only able to apply to fewer colleges. The key observation for OR-(un)fairness in the case of an *equal* number of classifications is that for pairs of elements with positive distance, the difference in expectation of at least one positive classification does not diverge linearly in the number of classifiers included in the composition.

Theorem 4. *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 $C_i \in \mathcal{C}$.*

For example, consider u and v for which $\mathcal{D}(u, v) = 0.10$. If two classifiers each assign 1 with probabilities $p_u = 0.25$ and $p_v = 0.15$ to u and v respectively, then the probability of positive classification by either of the two classifiers will be 0.44 for u and ≈ 0.28 for v, diverging from their original distance of $\mathcal{D}(u, v) = 0.1$. As the number of classifiers increases, the probabilities of positive classification by at least one classifier for any pair (so long as each is accepted with positive probability by sufficiently many classifiers), will eventually converge, as they both approach one. The region of divergence captures real unfairness: we don't expect students to apply to 30 or more schools, we expect them to apply to perhaps 5 or 10. For settings like loan applications (where an extended loan search with many credit inquiries may negatively impact an individual's credit score), small stretches in distance may have significant practical implications.

OR-composition improves if the constituent probabilities are at least $\frac{1}{2}$ (Lemma 5). This observation allows us to certify that a system is free of divergence when the function can be decomposed into an OR of subcircuits in which each individual has probability at least $\frac{1}{2}$ of positive classification. Let $\mathcal{C}^w \subseteq \mathcal{C}$ denote the set of classifiers which act on w, and let x_w be the indicator bit for the OR of the classifications $x_w = 1$ if $\sum_{C_i \in \mathcal{C}^w} C_i(w) \geq 1$ and 0 otherwise.

Lemma 5. *If a set of classifiers $\mathcal{C}$ satisfies OR-fairness where $\mathbb{E}[x_w] \geq 1/2$ 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 $w \in U$.*

We now turn to *Set Selection* problems. A classification system may need to cope with limits: a university may only

take n students in the freshman class; for the college to remain solvent, at least half the students must be able to pay full tuition; an advertiser has a finite budget. In addition, classifiers may operate on elements in arbitrary order, or may operate on only a subset of the universe. There are several axes to consider:

Known versus unknown subset or universe size: it is rare that a single classifier dictates the outcomes for a precisely defined universe or subset, and instead they generally act on a subset or universe of unknown size. The subset size may not be known in advance if it is generated randomly, or if the classifier simply doesn't have access to hidden subset selection processes. For example, an advertiser may know the average number of individuals who visit a website on a particular day, but be uncertain on any particular day of the exact number, and of the fraction interested in the products or services they wish to advertise.

Online versus offline: in many real-world settings, immediate classification response is critical. For example, advertising decisions for online ads must be made immediately upon impression; employers must render employment decisions quickly or risk losing out on potential employees or taking too long to fill a position.

Random permutations versus adversarial ordering: when operating in the online setting, the ordering of individuals may be adversarial or random. In practice, we expect that ordering will most likely not be a random permutation on the universe. For example, the order in which individuals apply for a job opening may be influenced by their social interactions with existing employees, which influences how quickly they hear about the job opening.

Constrained versus unconstrained selection: in many settings there are arbitrary constraints on the 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 accepted in a class to be able to pay full tuition.

We now formally define the *Cohort Selection* problem. *Universe Subset Selection* is defined analagously but without the constraint that the selected set have cardinality n.

Definition 4 (Cohort Selection Problem)**.** Given a universe U, an integer n and a task with metric $\mathcal{D}$, select a set of n individuals such that the probability of selection is 1-Lipschitz with respect to $\mathcal{D}$, where the probability of selection is taken over all randomness in the system. As above, 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 $\mathcal{S}_n : \Pi(2^U) \times \{0,1\}^* \to U^n$, we define the following experiment.

$\mathsf{Expt}(\mathcal{S}_n, \mathcal{X}, \mathcal{Y}, u)$**:** Step 1: Choose $r \sim \{0,1\}^*$. Step 2: Choose $V \sim \mathcal{Y}$. Step 3: Choose $\pi \sim \mathcal{X}(V)$. Step 4: Run $\mathcal{S}_n$ on π with randomness r, and output 1 if u is selected (positively classified).

The system is individually fair and a solution to the Cohort Selection Problem if for all $u, v \in U$, $\mathcal{S}_n$ outputs a set of n distinct elements of U and
$|\mathbb{E}[\mathsf{Expt}(\mathcal{S}_n, \mathcal{X}, \mathcal{Y}, u)] - \mathbb{E}[\mathsf{Expt}(\mathcal{S}_n, \mathcal{X}, \mathcal{Y}, v)]| \leq \mathcal{D}(u, v)$.

In Definition 4 we specify $\mathcal{S}$ independently of $\mathcal{X}$ and $\mathcal{Y}$ as these two distributions are not likely to be under the control of $\mathcal{S}$, and in practice may not even be known to $\mathcal{S}$. For example, an employer may create a resume screening system without knowledge of the subset of eligible candidates who will apply within a week of posting a new job. However, we still want the employer to fairly hire regardless of the ordering of the applicants.

Some variants of the cohort selection problem are amenable to solution. For example, in the full paper we give two algorithms, PermuteThenClassify and WeightedSampling, for the case in which the universe is known and decisions are made offline. Each algorithm is constructed from an arbitrary fair classifier C; the former is very simple, while the latter gives higher quality results when the underlying classifier C is overly selective. At the opposite extreme, in the online case, when the ordering of the stream is adversarial and $|U|$ is unknown, there is no fair solution.

In the remainder of this section we briefly touch on our results for the constrained cohort selection problem, which 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 (students able to pay full tuition) contribute to the budget, or if legally a certain fraction of people selected must meet some criterion (ie, demographic parity).

Definition 5 (The Constrained Cohort Selection Problem)**.** Given a universe U, $p \in [0,1]$, a subset $A \subset U$, and a metric for the task $\mathcal{D}$, solve the cohort selection problem with the added requirement that at least a p fraction of the members of the selected cohort are in A.

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 scholastic 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 $v \in V_u$,

$\mathcal{D}(u, v) = 0$ but the difference in probability of selection is at least $\frac{4}{5}$.

Cohort can be harder than Subset: Consider the problem of assigning students to public schools. Some fraction of students have been diverted to private schools, and fairly assigned to science or humanities campuses, so the public school system can only influence the outcomes of the students not attending private schools. If this system is able to change the focus of each public campus, it may be able to assign the remaining students so that Individual Fairness holds $\forall u, v \in U$. However, if the public school system only has a limited number of seats at science campuses and cannot change the focus of any campus, then the problem becomes an instance of constrained cohort selection, and sometimes cannot be solved.

4. Conclusions and Future Work

We have initiated the study of composition of fair classifiers and the first explicit study of the case in which classification decisions cannot be made independently, even by a fair classifier. Promising directions for future work include the search for higher revenue solutions for task-competition than RandomizeThenClassify and the construction of augmented classifiers that are *not necessarily* individually fair but that yield fairness when combined in a system.

References

Chouldechova, Alexandra. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

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

Dwork, Cynthia. Differential privacy. *Proceedings, Part II, chapter Differential Privacy. Springer*, 2006.

Dwork, Cynthia and Ilvento, Christina. Composition of fair systems. *arXiv preprint arXiv:1806.06122*, 2018a.

Dwork, Cynthia and Ilvento, Christina. Group fairness under composition. FATML, 2018b.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard S. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. URL http://arxiv.org/abs/1104.3913.

Edwards, Harrison and Storkey, Amos. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

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

Hardt, Moritz, Price, Eric, Srebro, Nati, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

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

Hu, Lily and Chen, Yiling. Fairness at equilibrium in the labor market. *CoRR*, abs/1707.01590, 2017. URL http://arxiv.org/abs/1707.01590.

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

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

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

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

Kilbertus, Niki, Rojas-Carulla, Mateo, Parascandolo, Giambattista, Hardt, Moritz, Janzing, Dominik, and Schölkopf, Bernhard. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Kim, Michael P, Reingold, Omer, and Rothblum, Guy N. Fairness through computationally-bounded awareness. *arXiv preprint arXiv:1803.03239*, 2018.

Kleinberg, Jon M., Mullainathan, Sendhil, and Raghavan, Manish. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL http://arxiv.org/abs/1609.05807.

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

Liu, Lydia T, Dean, Sarah, Rolf, Esther, Simchowitz, Max, and Hardt, Moritz. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

Nabi, Razieh and Shpitser, Ilya. Fair inference on outcomes. *arXiv preprint arXiv:1705.10378*, 2017.

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

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