Group Fairness Under Composition

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

Abstract

We examine the composition properties of a large class of statistical group fairness definitions. One corollary of our mostly negative results is that the group intersection techniques proposed by [Kearns, Neel, Roth and Wu 2017] and [Hebert-Johnson, Kim, Reingold and Rothblum 2017] may degrade under composition. We show several cases where group fairness definitions give misleading signals under composition and conclude that additional context is needed to evaluate group fairness under composition.

1. Introduction

The increasing reach of algorithmic decision-making in daily life has yielded an explosion in 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; Lambrecht & Tucker, 2016; 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). Most of this work focuses on a classifier or learning algorithm working in isolation. We initiate the study of group fairness guarantees under composition. For example, if all advertisers in an advertising system independently satisfy Equalized Odds (Hardt et al., 2016), does the entire advertising *system* have the same guarantee? Our first result is that naïve composition of group-fair classifiers will not in general yield a group-fair system, paralleling an analogous result for *individually fair* classifiers in a companion paper (Dwork & Ilvento, 2018b).[1]

The motivations for group fairness are two-fold. Proportional representation is often desirable its own right; alternatively, the *absence* of proportional allocation of goods can signal discrimination in the allocation process, typically against historically mistreated or under-represented groups. Thus, 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 ability to pay for an item. Conditional Parity, a general framework proposed in (Ritov et al., 2017) for discussing these definitions, conveniently captures many of the popular group fairness definitions popular in the literature including Equal Odds and Equal Opportunity (Hardt et al., 2016), and Counterfactual Fairness (Kusner et al., 2017).

Definition 1 (Conditional Parity)**.** (Ritov et al., 2017) 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 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.

The definition captures the intuition that, conditioned on qualification, the setting of protected attributes should not *on average* impact the classification result. Note that this is not a guarantee about treatment at the individual level; it speaks only to group-level statistical properties in expectation. In contrast, Individual Fairness makes strict requirements on the outcomes for each pair of individuals:

Definition 2 (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 $C : U \times \{0,1\}^* \rightarrow O$, such that $\tilde{C} : U \rightarrow \Delta(O)$, and a distance measure $d : \Delta(O) \times \Delta(O) \rightarrow \mathbb{R}$, C is *individually fair*

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

if and only if for all $u, v \in U$, $\mathcal{D}(u, v) \leq d(\tilde{C}(u), \tilde{C}(v))$.[2]

A weakness of group fairness definitions, addressed by Individual Fairness, is the problem of subgroup unfairness: a classifier that satisfies Conditional Parity with respect to race and gender independently may fail to satisfy Conditional Parity with respect to the *conjunction* of race and gender. Furthermore, the protected attributes ($\mathcal{A}$) may not be sufficiently rich to describe every "socially meaningful" group one might wish to protect from discrimination. For example, preventing discrimination against women is insufficient if it allows discrimination against women who are mothers, or who dress in a particular style. To address this, two concurrent lines of work consider fairness for collections of large, possibly intersecting sets (Kearns et al., 2017; Hébert-Johnson et al., 2017). As we will see, composition exacerbates this problem uniquely for group fairness definitions (and not for Individual Fairness).

The rest of this paper is organized as follows. In Section 2, we examine composition problems in which a trade-off is required between tasks. In Section 3, we extend our understanding to settings in which the relevant outcome is derived through composition of related classifiers.

2. Task-Competitive Composition

Many systems require a trade-off between tasks. For example, a website may have only one slot for an advertisement, and two advertisers competing for that slot for each person who visits the website. In this setting each advertiser's classification task is to decide whether (and how much) to bid for the user. The single slot composition problem requires us to satisfy Conditional Parity for each task separately while choosing only one positive classification for each individual.

Definition 3 (Single Slot Composition Problem for Conditional Parity)**.** 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 Conditional Parity for a set $\mathcal{T} = \{T_1, \ldots, T_k\}$ of k tasks, with stratification sets $\mathcal{Z}_1, \ldots, \mathcal{Z}_k$ and protected attribute sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$, 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 $i \in [k]$, $\Pr[x_i | (\mathbf{a_i} = a_i, \mathbf{z_i} = z_i)] = \Pr[x_i | (\mathbf{a_i} = a'_i, \mathbf{z_i} = z_i)]$, for all $z_i \in \mathcal{Z}_i$ and for all $a_i, a'_i \in \mathcal{A}_i$, where probability is taken over the randomness of the classification procedure and the selection of elements from the universe.

In this brief abstract we restrict our attention to the simple case in which there are only two tasks T and T', and there is a strict preference for T whenever an individual $w \in U$ is classified positively for both tasks. That is, "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 individualistic, tie breaking procedures.

The analysis for this problem has two cases based on whether or not the two tasks represented by the advertisers are related[3], a notion formalized in Definition 4, which captures the intuition that two tasks are *unrelated* if the protected attributes and stratification levels for the either task is not predictive of stratification level for the other task. For example if (1) a daycare service exclusively targets parents, that is, the stratification set includes only parental status, and (2) there are roughly equal fractions of men and women qualified for jobs who are parents, and (3) interest in daycare services is likewise independent of qualification for jobs, then we may say that these tasks are unrelated. In contrast, since women are more likely to be responsible for household purchases (Schiebinger, 2016), gender *is* predictive of interest in purchasing home goods, even when conditioned on qualification for jobs, so the tasks may be considered related. Thus a woman qualified for a job is more likely to be targeted for home goods advertisements than a man qualified for a job.

Definition 4 (Unrelated tasks)**.** Two tasks T, T' are considered unrelated if for all $a \in \mathcal{A}$, $a' \in \mathcal{A}'$, $z \in \mathcal{Z}$ and $z' \in \mathcal{Z}'$, $\Pr[\mathbf{z} = z | \mathbf{a}' = a', \mathbf{z}' = z'] = \Pr[\mathbf{z} = z]$ and $\Pr[\mathbf{z}' = z' | \mathbf{a} = a, \mathbf{z} = z] = \Pr[\mathbf{z}' = z']$, where probability is taken over selection of members of the universe.

2.1. Competition Between Related Tasks

If the stratification and protected attributes for one task are predictive of qualification for the other task, then the classifier that is preferred may 'claim' more than its fair share of these individuals, leaving too few remaining to satisfy Conditional Parity for the second task (see Figure 1).

Theorem 1. *(Informal)*[4] *For any two related tasks T, T' there exists a pair of classifiers C, C' which satisfy Conditional Parity in isolation but not under competitive composition.*

We omit the proof for lack of space, and instead build intuition with the following simple example. Imagine that the home goods advertiser always outbids the jobs advertiser, and only bids on women. Then the probability of a woman seeing a job advertisement, regardless of her qualification,

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

[3] The observations on related tasks are not unique to group fairness definitions, and in a companion work we prove an analogous result for Individual Fairness. The observations on unrelated tasks are unique to group fairness.

[4] We state this theorem informally to avoid the technicalities of potentially empty stratification sets and other corner cases. The formal theorem extends naturally to more complex tie-breaking functions.

Figure 1. The preferred task (shown in gray) may '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).

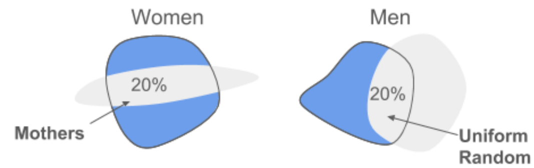

Figure 2. The preferred task (shown in gray) may 'claim' the same fraction of the qualified individuals for each of the protected groups for the second task (shown in blue), but the subgroups claimed may be socially meaningful.

is zero. Thus, unless the job advertiser refuses to advertise to *any* qualified men, Conditional Parity cannot be satisfied, as $\Pr[x_{job}|\mathbf{a} = f, \mathbf{z} = z] = 0 \neq \Pr[x_{job}|\mathbf{a} = m, \mathbf{z} = z]$, where $\mathbf{a} = m$ corresponds to male, and likewise $\mathbf{a} = f$ to female, and z is any qualification level for jobs.

This result extends to more general tie-breaking functions, which can be used to represent more complicated bidding strategies, or to represent more nuanced decision-making in settings such as scheduling or other offline decisions. This and related results indicate that each entity may learn a classifier fairly in isolation, but when the classifiers are applied in a competitive scenario the result of the composition may fail entirely to prevent discrimination.

2.2. Competition Between Unrelated Tasks

Competition between unrelated tasks has more subtle group fairness implications. Recall the example of unrelated tasks of an advertiser for daycare services and an advertiser for jobs. Concretely, we take $\mathcal{A} = \mathcal{A}' = \{m, f\}$ representing gender to be the protected attribute for both advertisers, and take the stratification set for jobs to be $\mathcal{Z} = \{\text{qualified, not qualified}\}$, and for daycare to be $\mathcal{Z}' = \{\text{parent, not parent}\}$. If we make the assumptions that parental status and gender are not predictive of job qualification and that job qualification and gender are not predictive of parental status,[5] we should expect that many classifiers which independently satisfy Conditional Parity for the protected attribute gender will also satisfy Conditional Parity in the competitive setting.

Lemma 2. *Consider a system with unrelated tasks T, T', in which ties are broken in the same way for all individuals. Let C, C' be any pair of classifiers such that for all $z \in \mathcal{Z}$ and all $z' \in \mathcal{Z}'$, C treats each element with qualification $\mathbf{z} = z$ identically and C' treats each element with qualification $\mathbf{z}' = z'$ identically. Then the system will satisfy Conditional Parity under task-competitive composition.*

[5]For the purposes of this example, we assume that there are equal fractions of male and female parents who are qualified for jobs. In reality, there may be a gender imbalance in particular fields; the example can be extended to this case, but is omitted for lack of space.

This positive result of Lemma 2 reveals a weakness of Conditional Parity as a notion of fairness: returning to our example of daycare providers versus job ads, when the day care advertisers routinely outbid the jobs advertisers, the lemma says Conditional Parity is maintained, but it may be that no job ads are shown to parents! (Figure 2) This would not occur if parenthood were a protected attribute, but parenthood may not even be a feature in the job advertiser's dataset. Sensitive attributes like parental status may even be illegal to collect in employment-related training or testing datasets, but are likely legitimately collected and targeted by other potentially conflicting classification tasks.

Conditional Parity does not protect all socially meaningful subgroups from unfair treatment: even without composition, a classifier for jobs that excludes some (possibly large) similarly sized subsets of qualified men and women will still satisfy Conditional Parity, it simply discriminates equally against mothers and fathers. Indeed, problematic (large) subgroup semantics are part of the motivation for (Kearns et al., 2017; Hébert-Johnson et al., 2017). The danger of composition is that the features describing this subset may be missing from the feature set of the jobs classifier, rendering the protections proposed in (Kearns et al., 2017) and (Hébert-Johnson et al., 2017) ineffective.

In fact, if, in response to gender disparity caused by task-competitive composition, classifiers iteratively adjust their bids to try to achieve Conditional Parity, they may unintentionally *learn* themselves into a state that satisfies Conditional Parity with respect to gender, but behaves poorly for a socially meaningful subgroup. For example, let's imagine that home goods advertisers aggressively advertise to women who are new parents, as their life-time value to the advertiser ($\mathcal{Z}$) is the highest of all universe elements. 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.

3. Composition without Competition

In many settings, the outcome of a single classifier is not the relevant point at which to evaluate fairness. For example, a person applying for an auto loan may apply to several banks and, assuming comparable terms, cares only that at least one loan offer was made. Similarly, students applying to college only need to be accepted to at least one college. In such settings, the relevant outcome for determining fairness is on the 'or' of all of the classifiers. We analyze this problem in two cases to show how Conditional Parity can give conflicting signals under OR-composition.[6]

When elements with $\mathbf{z} = z$ are treated equally. Any classifier that treats all elements with identical settings of $\mathbf{z}$ identically satisfies Conditional Parity in isolation.[7] Interestingly, the 'or' of any number of such classifiers will also satisfy Conditional Parity.

Definition 5 (Conditional Parity OR-Fairness)**.** Given a stratification set $\mathcal{Z}$ and a protected attribute set $\mathcal{A}$, a set of classifiers $\mathcal{C}$ satisfies *OR-Fairness* if the indicator variable $x = 1$ if $\sum_{C_i \in \mathcal{C}} C_i(x) \geq 1$ and $x = 0$ otherwise satisfies $\Pr[x|\mathbf{a} = a, \mathbf{z} = z] = \Pr[x|\mathbf{a} = a', \mathbf{z} = z]$, for all $a, a' \in \mathcal{A}$ and for all $z \in \mathcal{Z}$.

Proposition 3. *Let $\mathcal{C}$ be a set of classifiers with a common stratification set $\mathcal{Z}$. If $\forall C \in \mathcal{C}\ \forall z \in \mathcal{Z}$: C treats all universe elements with $\mathbf{z} = z$ identically, then $\mathcal{C}$ satisfies Conditional Parity OR-fairness.*

Although individuals with identical qualification $\mathbf{z} = z$ are treated equally, Conditional Parity does not ensure that similarly (but not identically) qualified individuals are treated similarly, and this disparity between similar qualification strata may *increase* under composition. For example, if individuals with z_1 may receive a loan with probability 0.25 and individuals with z_2 receive a loan with probability 0.15, then after two loan applications, individuals with z_1 have probability 0.44 of getting at least one loan offer whereas individuals with z_2 have 0.28. Particularly in the case of loan applications (where an extended loan search with many credit inquiries may impact an individual's credit score), similar treated similarly may be an implicit requirement for fairness.

When elements with $\mathbf{z} = z$ are not treated equally. In contrast to the case above, there is no guarantee that Conditional Parity will be satisfied under 'or' composition when individuals with the same $\mathbf{z} = z$ are not all treated equally.

There are many natural cases where we might want to treat elements with the same z differently, for example, if the randomness of the environment results in a bimodal distribution for one group, and a unimodal distribution for the other. Let us imagine that each $z \in \mathcal{Z}$ represents a range of acceptance probabilities. In each range, individuals are classified as p_h, high probability within this range, p_m, medium probability within this range, and p_l, low probability within this range. Each group may consist of a different mix of individuals mapped to p_h, p_m, and p_l.

Consider the following simple universe: for a particular $z \in \mathcal{Z}$, group a_1 has only elements with medium qualification q_m, group a_2 has half of its elements with low qualification q_l and half with high qualification q_h. Choosing $p_h = 1$, $p_m = .75$, $p_l = .5$, satisfies Conditional Parity for a single application. However, for two applications, the the squares in each group diverge ($.9375 \neq .875$):
$1 - (1 - p_m)^2 \neq \frac{1}{2}(1 - (1 - p_h)^2) + \frac{1}{2}(1 - (1 - p_l)^2)$
Thus, Conditional Parity is violated. Note, however, that many of the individuals with $\mathbf{z} = z$ have been drawn closer together under composition, and none have been pulled further apart.

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_2 to satisfy a technicality.

Although group fairness definitions make no guarantees on the treatment of individuals, the contrast between how Conditional Parity behaves under OR-composition when individuals with the same value of $\mathbf{z}$ are treated equally or not is worth considering. In some cases we may observe failures under OR-composition for Conditional Parity, even when Individual Fairness is satisfied, and failure to satisfy Individual Fairness when Conditional Parity is satisfied. This brittleness extends to other settings like selecting a cohort of exactly n elements and satisfying calibration under composition, and to other logical functions as well as constrained settings.

[6](Bower et al., 2017) considers what boils down to AND-fairness for Equal Opportunity (Hardt et al., 2016) and presents an excellent collection of evocative example scenarios.

[7]Assuming appropriate discretization, such classifiers are also individually fair. In a companion paper, we observe that OR fairness is not always satisfied under composition for Individual Fairness.

4. Conclusions and Future Work

We have shown that classifiers which appear to satisfy group fairness properties in isolation may not compose well with other fair classifiers, that the signal provided by group fairness definitions under composition is not always reliable, and that composition requires additional considerations for subgroup fairness. A promising direction for future work is the augmentation of classifiers group fairness for large, intersecting, groups (Kearns et al., 2017; Hébert-Johnson et al., 2017), as well as classifiers with Individual Fairness for large subgroups) (Kim et al., 2018), to incorporate contextual information, with the goal of improving composition.

References

Bower, Amanda, Kitchen, Sarah N., Niss, Laura, Strauss, Martin J., Vargas, Alexander, and Venkatasubramanian, Suresh. Fair pipelines. *CoRR*, abs/1707.00391, 2017. URL http://arxiv.org/abs/1707.00391.

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

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

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.

Lambrecht, Anja and Tucker, Catherine E. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. 2016.

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.

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

Schiebinger, Londa. Private Communication, 2016.

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.