Fairness Over Utilities Via Multi-Objective Rewards

Jack Blandin [1] **Ian Kash** [1]

Abstract

Group fairness definitions make assumptions about the underlying decision-problem that restrict them to classification problems. Numerous bespoke interpretations of group fairness definitions exist as attempts to extend them to specific applications. In an effort to generalize group fairness definitions beyond classification, Blandin & Kash (2021) explore using *utility* functions to define group fairness measures. In addition to the decision-maker's utility function, they introduce a *benefit* function that represents the individual's utility from encountering a given decision-maker policy. Using this framework, we interpret fairness problems as a multi-objective optimization, where we aim to optimize for both the decision-maker's utility and the individual's benefit, as well as reduce the individual benefit difference across protected groups. We demonstrate our instantiation of this multi-objective approach in a reinforcement learning simulation.

1. Introduction

In this work, we focus on *group fairness* definitions, where an algorithm is considered fair if its results are independent of one or more protected attributes such as gender, ethnicity, or sexual-orientation. There is by now an extensive body of work on group fairness works in classification settings (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017; Galhotra et al., 2017). This narrow focus has been productive, but often conceals assumptions that do not always hold true in other contexts, such as reinforcement learning (RL) or clustering, resulting in definitions that are tightly coupled with a particular problem domain.

In this paper we first summarize the recent work of Blandin & Kash (2021) which examines four such assumptions (see Section 2). While each assumption individually has received prior scrutiny, their main contribution is a framework which demonstrates how utility functions can be used to define fairness and help resolve all four issues in a uniform way which subsumes a number of bespoke approaches (see Section 3).

Moving beyond Blandin & Kash (2021), we examine how such utility-based fairness definitions can be achieved in the context of RL. In Section 4 we describe a way to obtain fair policies via multi-objective reward functions. In Section 5 we provide an experiment showing the effectiveness of this multi-objective approach on a repeat-loan application RL environment. We discuss future work in Section 6.

2. Classification Group Fairness Issues

Blandin & Kash (2021) discuss four assumptions implicit in many group fairness definitions that cause issues when moving beyond classification settings.

Assumption 2.1. Fair predictions have fair outcomes.

Many group fairness definitions require equal predictions between protected groups (Berk et al., 2018; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2012; Hardt et al., 2016). For example, in the binary case with a *minority* group and a *majority* group, *Demographic Parity* considers a binary classifier to be fair if it predicts the positive class for individuals in the minority group and majority groups with equal probability. This implicitly assumes that a positive prediction is always a positive outcome for an individual. However, there are many problem domains where this is not true, such as in loan applications where approving an unqualified applicant for a loan may hurt the applicant since they are likely to default (Liu et al., 2018).

Assumption 2.2. Observed values of the target variable are independent of predictions.

Some fairness definitions depend on the observed value of the target variable as well as the prediction. For example, *Equal Opportunity* requires equal treatment of the qualified individuals in each group, where *qualified* refers to individuals who were observed to be in the positive class (Hardt et al., 2016). But if the prediction itself can influence the individual's qualification, such as in prison sentencing,

[1] Department of Computer Science, University of Illinois at Chicago, Chicago, United States. Correspondence to: Ian Kash <iankash@uic.edu. This material is based upon work supported by the National Science Foundation Program on Fairness in AI in collaboration with Amazon under Grant No. 1939743.>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

then the definition can be satisfied through a self-fulfilling prophecy by manipulating who is considered qualified (Ensign et al., 2018; Imai & Jiang, 2020; Barocas et al., 2017; Kasy & Abebe, 2021).

Assumption 2.3. The objective is to predict some unobserved target variable.

In classification problems, the goal is to make a single prediction of some latent qualification attribute of the individual. However, this is not true in other ML environments where the decision is not necessarily a prediction of some ground-truth value, and where there may be more than one decision per individual. In sequential decision settings such as reinforcement learning (RL), the goal is to maximize a reward rather than predict a target. Additionally, there can be multiple sequential decisions made for each individual and we may wish to measure fairness across the entire sequence. Ranking problems and clustering also have differing objectives than traditional classification, and so require alternative fairness considerations. Several works attempt application-specific remedies, such as for sequential decision processes (Jabbari et al., 2017; Bower et al., 2017; Dwork et al., 2020; Emelianov et al., 2019), for ranking (Celis et al., 2017; Singh & Joachims, 2019; Zehlike et al., 2021), and for clustering (Chierichetti et al., 2017; Bera et al., 2019; Chen et al., 2019; Abbasi et al., 2021).

Assumption 2.4. Decisions for one individual do not impact other individuals.

Each classification prediction is independent of the predictions made for other individuals. However, this does not generalize to all of ML. In clustering, for instance, the impact of one individual's cluster assignment may depend on the cluster assignments of other individuals. For example, Abbasi et al. (2021) consider redistricting as a fair clustering problem, where fairness implies that constituents from each political party are equally represented by their assigned district. In order to measure how well a constituent is represented by their district, we need to know *who else* was assigned to their district.

3. Using Utilities in Group Fairness

Several works incorporate notions of utilities when resolving fairness issues for a particular domain (Liu et al., 2018; Heidari et al., 2018; Wen et al., 2021). Building on these notions, Blandin & Kash (2021) construct utility-based group fairness definitions that help resolve the issues resulting from Assumptions 2.1-2.4, and therefore extend across ML.

Benefit Borrowing terminology from Heidari et al. (2018), they introduce a variable called *benefit*, which represents the individual's utility resulting from a prediction. They define a utility representation of Demographic Parity, for example, by requiring a decision-algorithm m to have the probability that an individual receives a beneficial outcome be independent of their group:

$$P(W_m \geq \tau \mid Z{=}0) = P(W_m \geq \tau \mid Z{=}1)\ . \qquad (1)$$

where W_m is the expected benefit received by an individual under decision-algorithm m, τ is the minimum benefit needed to be considered a beneficial outcome, and Z is the individual's protected attribute that identifies their group. By measuring fairness directly in terms of benefit, their definitions enforce fair outcomes even in domains where the predictions impact individuals differently. Furthermore, since *utility* is a more universal concept than *prediction* or *target variable*, this approach continues to make sense in domains where Assumptions 2.3 and 2.4 do not hold.

Counterfactual Outcomes We saw in the discussion of Assumption 2.2 that the standard definition of Equal Opportunity is vulnerable to self-fulfilling prophecies. In order to remedy this, they construct a more extensible Equal Opportunity definition by giving a more general interpretation of what it means to be *qualified*. They interpret qualification as an individual where there exists a decision that will yield a positive outcome for both the decision-algorithm *and the individual*. In other words, they measure qualification in terms of mutual beneficence for both the decision-algorithm and the individual. Similar to how they define a positive outcome for the individual as obtaining a benefit W_m above some threshold τ, they define a positive outcome for the decision-maker as the minimum *cost* C_m above a threshold ρ. They define a utility form of Equal Opportunity as

$$P(W_m \geq \tau \mid \Gamma{=}1, Z{=}0) = P(W_m \geq \tau \mid \Gamma{=}1, Z{=}1) \qquad (2)$$

where Γ is an indicator variable with

$$\Gamma = \begin{cases} 1 & \text{if } \exists m' \in M : W_{m'} \geq \tau \wedge C_{m'} \leq \rho \\ 0 & \text{otherwise}\ . \end{cases} \qquad (3)$$

By considering counterfactual outcomes, their Equal Opportunity definition prevents self-fulfilling prophecies and is well-defined for a broader range of ML environments.

4. Fairness via Multi-Objective Optimization

Moving beyond Blandin & Kash (2021), we now turn to the question of *obtaining* fair policies.

In a seminal paper, Liu et al. (2018) showed that adding fairness constraints to a two-step loan application decision model can negatively impact the credit scores of the disadvantaged applicants that the constraint aims to protect. Several works build on this by analyzing how the *qualification* of individuals change over time as a function of various decision-based fairness constraints (D'Amour et al., 2020;

Mouzannar et al., 2019; Zhang et al., 2020). The notion of *benefit* is a generalization of this notion of *qualification*, in that *benefit* represents the individual's utility. Although prior works study long-term qualification impact, few works offer techniques for learning policies that obtain long-term qualification equality across protected groups. Those that do either focus on single-step environments (e.g. (Martinez et al., 2020; Diana et al., 2021; Hu & Chen, 2020)) and so they do not extend to other ML environments such as RL, or they do not optimize for qualification improvement (e.g. (Wen et al., 2021; Raab & Liu, 2021; Hu & Zhang, 2022)) and so are prone to violating the *no-harm* principle (Martinez et al., 2020; Diana et al., 2021) where one or more group's qualification is lowered in order to satisfy the equality constraint.

In order to ensure that qualification (i.e. benefit) is improved and that qualification equality is maintained, we propose a technique that optimizes for these values directly as second and third objectives in a multi-objective optimization approach. While our approach extends to any environment where a utility function can be applied, we focus on the RL setting. We construct a weighted sum of three distinct reward functions for decision-maker cost, qualification improvement, and qualification equality. By framing our technique as a reward function, rather than a policy intervention, we obtain two key benefits. First, any RL algorithm may learn the optimal policy since the reward adheres to the standard RL paradigm. Second, by encoding the objective in the reward, we make few assumptions about the environment, notably about the state dynamics or the level of observability. In Section 5 we provide an example of how optimizing for qualification improvement and qualification equality can produce better outcomes for individuals than a decision-constrained approach by only deviating from the cost-optimal policy when qualification improvement or equality is expected.

5. Experiment

5.1. Setup

We consider a fully observable MDP environment where a loan applicant is sampled in each timestep t and a lender makes a binary decision for the applicant. The lender is represented by a policy π which can either approve the applicant's loan ($a_t = 1$) or reject it ($a_t = 0$). The applicant at time t has a binary protected attribute $s_t^Z = z \in \{0, 1\}$ and a credit score $s_t^{Y^z} \in \{0, 1, 2\}$. An applicant's credit score defines their repayment probability. An applicant's credit score and protected attribute determine the probability of the applicant's credit score increasing or decreasing in the event of a repaid, defaulted, or rejected loan. While only the applicant from one protected group requests a loan per timestep, the decision-maker also observes the credit score of the applicant from the other group $s_t^{Y^{1-z}}$, and so is always able to observe the current credit scores of both groups.[1] We consider the applicant's credit score as their *qualification* attribute, and so we aim to improve overall applicant credit scores as well as ensure that credit scores are equal across protected groups. The lender can only give the applicant a loan if they have sufficient cash, where *cash* is a portion of the state. If the applicant is approved for a loan and repays it, the lender receives a positive profit R^C, where *profit* corresponds to the negative of the decision-maker cost C from Section 3. A repaid loan also increases the lender's available cash. If the lender approves a loan and the applicant defaults, then the lender receives a negative profit and their available cash decreases. The lender having finite cash is significant since it constrains their decisions so that loans cannot simply be granted without consequences. So the lender may need to strategically maintain a sufficient amount of cash in order to enable later loan approvals oriented around credit score improvement or equality. We consider the *demographic-variant transition* scenario studied by Zhang et al. (2020) where a disadvantaged applicant ($s_t^Z = 0$) has a lower likelihood of increasing their credit score after successfully repaying a loan than does an advantaged applicant ($s_t^Z = 1$), a scenario where Demographic Parity and Equal Opportunity constraints exacerbate qualification inequality (Zhang et al., 2020).

5.2. Measurements

We track metrics as averages over the course of an *episode* ξ which is an n-length sequence of state-action pairs $\xi = \{(s_0, a_0), (s_1, a_1), ..., (s_{n-1}, a_{n-1})\}$ that starts in an initial state and ends in a terminal state. (In our experiment we take $n = 10$.) There are several metrics of interest here. First, the lender's total profit should be tracked since this is the primary utility being optimized: $\text{TotalProfit}(\xi) = \sum_{t=0}^{n} R^C(\sigma_t)$ where $\sigma_t = (s_t, a_t, s_{t+1})$ is the state transition from t to $t+1$ after taking action a_t, from which the outcome (repaid, defaulted, rejected) can be inferred. Second, we wish to track each applicant group's average credit score $\overline{Y^z}$ $\forall z \in \{0, 1\}$ since this value serves as a proxy for the applicant financial well-being: $\overline{Y^z}(\xi) = \frac{1}{n} \sum_{t=0}^{n} s_t^{Y^z}$.

We also wish to ensure that the two applicant groups are treated fairly, and so we track two utility fairness measures. First, we track the absolute difference in average group credit score: $\text{CreditDiff}(\xi) = |\overline{Y^0} - \overline{Y^1}|$, which is a non-threshold adaptation of Equation (1) and where $W_m = \frac{1}{n} \sum_{t=0}^{n} s_t^{Y^z}$. Second, we track the absolute group difference in the percentage of qualified ap-

[1] Our model is a variant of Zhang et al. (2020) with more credit scores but full observability. It can be interpreted as one individual from each group repeatedly applying for loans or a stylized model of population-level decisions and credit-score evolution.

plicants, where *qualified* refers to applicants who have a credit score above the threshold α that makes them more likely to repay a loan than to default: α-CredDiff$(\xi) = |P(s_t^{Y^0} \geq \alpha) - P(s_t^{Y^1} \geq \alpha)|$ with α defined such that $P(\texttt{Repay} \mid s_t^{Y^z} \geq \alpha) > P(\texttt{Default} \mid s_t^{Y^z} \geq \alpha)$. This is an instantiation of Equations (2)-(3) with $\tau = \rho = \alpha$ and $\Gamma = 1$ if $s_t^{Y^z} \geq \alpha$.

5.3. Multi-Objective Optimization Algorithms

We consider a combination of three objectives

$$R(\lambda^C, \lambda^Q, \lambda^F, \sigma_t) = \lambda^C R^C(\sigma_t) + \lambda^Q R^Q(\sigma_t) + \lambda^F R^F(\sigma_t) \ . \tag{4}$$

where $R^Q(\sigma_t) = \text{norm}(s_{t+1}^{Y^z})$ is the credit score of the sampled applicant at time $t+1$, normalized to lie in [0,1]. $R^F(\sigma_t)$ is computed as the (normalized) absolute difference in the credit score samples of each group at time $t+1$:

$$R^F(\sigma_t) = \text{norm}(-|s_{t+1}^{Y^0} - s_{t+1}^{Y^1}|) \ . \tag{5}$$

Thus Equation (4) is a weighted sum of reward contributions for how much the transition contributes towards the lender profit, population credit score, and credit score difference across groups. The weights λ^C, λ^Q, and λ^F allow tradeoffs among the three objectives. We denote Equation (4) parameterized by $\lambda^C = \text{i}, \lambda^Q = \text{j}, \lambda^F = \text{k}$ as $R^{\text{i,j,k}}$. The optimal policy is

$$\pi^*_{i,j,k} = \underset{\pi \in \Pi}{\text{argmax}} \ \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t R^{\text{i,j,k}}(s_t, \pi(s_t), s_{t+1})\big] \ . \tag{6}$$

where Π is the space of policies and γ is the discount rate, which we take to be .9.

5.4. Benchmark Algorithms

To benchmark our multi-objective approaches, we compare two baseline algorithms as well. First, we consider a reward function that only considers decision-maker cost:

$$R(\sigma_t) = R^C(\sigma_t) \equiv R^{1,0,0}(\sigma_t) \ . \tag{7}$$

Next we compare a policy that optimizes the same cost-optimized reward but also has a constraint that requires the group difference in loan approval rate for applicants with an α-qualified credit score be less than an allowable margin ϵ:

$$\begin{aligned} \pi^*_{\text{C},1,0,0} = &\underset{\pi \in \Pi}{\text{argmax}} \ \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t R^C(\sigma_t)\big] \\ \text{s.t.} \quad &\big|P(a_t = 1 \mid s_t^{Y^0} \geq \alpha, s_t^z = 0) \\ &- P(a_t = 1 \mid s_t^{Y^1} \geq \alpha, s_t^z = 1)\big| < \epsilon \end{aligned} \tag{8}$$

where ϵ is a practitioner-supplied parameter set at .05 in this case. This approach is an instantiation of an Equal Opportunity constraint (Hardt et al., 2016), and we implement the algorithmic approach proposed by Wen et al. (2021). The MDP is simple enough that we can compute optimal policies for each objective via linear programming.

Algo	TotalProfit	$\overline{Y^0}$	$\overline{Y^1}$	CredDiff	α-CredDiff
$\pi^*_{1,0,0}$	**.937**	.648	.745	.097	7.7 %
$\pi^*_{\text{C},1,0,0}$	.860	$\underline{.531}$	$\underline{.713}$	$\underline{.182}$	$\underline{12.1\ \%}$
$\pi^*_{1,1,0}$	.625	.883	.782	.101	**5.1 %**
$\pi^*_{1,1,1}$	$\underline{.278}$	**.906**	**.839**	**.067**	8.9 %

Table 1. Results of the repeat-loan RL environment where bolded/underlined values indicate the column best/worst values.

5.5. Results

Table 1 shows the Section 5.2 metrics averaged across 5,000 episodes. We see that the constrained policy $\pi^*_{\text{C},1,0,0}$ violates the *no-harm* principle since it produces worse credit scores for both groups and worse credit score differences than the profit-optimal policy. The results also support Zhang et al.'s (2020) claim that when the probability of credit scores increasing upon repayment is different between groups, Equal Opportunity constraints exacerbate qualification (credit score) inequality. This is counterintuitive since the whole point of the added constraint is to benefit the applicants and to ensure that the two groups are treated fairly, but it ends up hurting both groups as well as increasing the disparity between them. The problem is that the constraint only requires equal *actions* be taken for some subset of the applicant population ($s_t^{Y^z} \geq \alpha$ in this case), but does not require any equality over *outcomes*. Because the two groups have differing outcome transition dynamics, even for the same action (i.e. Assumption 2.1 is violated), equal actions do not imply equal outcomes. On the other hand, the multi-objective policies that optimize for the credit score outcomes directly ($\pi^*_{1,1,0}$ and $\pi^*_{1,1,1}$) ensure that any deviation from the cost-optimal policy must correspond to an improvement in credit scores or group equality. This is reflected in Table 1 where the two multi-objective policies collectively obtain the best performance on every credit-focused measure. This demonstrates the promise of our multi-objective approach in avoiding the issues Zhang et al. (2020) observed with prior approaches.

6. Future Work

We conclude with several avenues for future work. First, we can explore the space of preference parameters that produces the best combination of objective functions for various environments, as well as explore if there are other objective functions that are more effective. For example, it may be better to optimize for Y^0 and Y^1 directly as the second and third objective functions, instead of their weighted average and their absolute difference. Additionally, we can explore *no-preference* multi-objective techniques which do not require preference weights as input. Lastly, we can analyze the necessary conditions needed for utility-based fairness definitions to be satisfied when optimizing for the various reward functions.

References

Abbasi, M., Bhaskara, A., and Venkatasubramanian, S. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 504–514, 2021.

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. Fair algorithms for clustering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Blandin, J. and Kash, I. Fairness through counterfactual utilities. *arXiv preprint arXiv:2108.05315*, 2021.

Bower, A., Kitchen, S. N., Niss, L., Strauss, M. J., Vargas, A., and Venkatasubramanian, S. Fair pipelines. *arXiv preprint arXiv:1707.00391*, 2017.

Celis, L. E., Straszak, D., and Vishnoi, N. K. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

Chen, X., Fain, B., Lyu, L., and Munagala, K. Proportionally fair clustering. In *International Conference on Machine Learning*, pp. 1032–1041. PMLR, 2019.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/978fce5bcc4eccc88ad48ce3914124a2-Paper.pdf.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 66–76, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Dwork, C., Ilvento, C., and Jagadeesan, M. Individual fairness in pipelines. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Emelianov, V., Arvanitakis, G., Gast, N., Gummadi, K., and Loiseau, P. The price of local fairness in multistage selection. In *IJCAI-2019-Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5836–5842. International Joint Conferences on Artificial Intelligence Organization, 2019.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pp. 160–171. PMLR, 2018.

Galhotra, S., Brun, Y., and Meliou, A. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498–510, 2017.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems*, 31, 2018.

Hu, L. and Chen, Y. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 535–545, 2020.

Hu, Y. and Zhang, L. Achieving long-term fairness in sequential decision making. *arXiv preprint arXiv:2204.01819*, 2022.

Imai, K. and Jiang, Z. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*, 2020.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International Conference on Machine Learning*, pp. 1617–1626. PMLR, 2017.

Kasy, M. and Abebe, R. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 576–586, 2021.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR, 2020.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368, 2019.

Raab, R. and Liu, Y. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34, 2021.

Singh, A. and Joachims, T. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056*, 2019.

Wen, M., Bastani, O., and Topcu, U. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021.

Zehlike, M., Yang, K., and Stoyanovich, J. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.