Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making

Alex Chohlas-Wood [1] **Madison Coots** [1] **Emma Brunskill** [2] **Sharad Goel** [1]

Abstract

In the dominant paradigm for designing equitable machine learning systems, one works to ensure that model predictions satisfy various fairness criteria, such as parity in error rates across race, gender, and other legally protected traits. That approach, however, typically divorces predictions from the downstream outcomes they ultimately affect, and, as a result, can induce unexpected harms. Here we present an alternative framework for fairness that directly anticipates the consequences of actions. Stakeholders first specify preferences over the possible outcomes of an algorithmically informed decision-making process. For example, lenders may prefer extending credit to those most likely to repay a loan, while also preferring similar lending rates across neighborhoods. One then searches the space of decision policies to maximize the specified utility. We develop and describe a method for efficiently learning these optimal policies from data for a large family of expressive utility functions, facilitating a more holistic approach to equitable decision-making.

1. Introduction

Statistical predictions are now used to inform high-stakes decisions in a wide variety of domains. For example, in banking, loan decisions are based in part on estimated risk of default (Leo et al., 2019); in criminal justice, judicial bail decisions are based on estimated risk of recidivism (Cadigan & Lowenkamp, 2011; Goel et al., 2018; Latessa et al., 2010; Milgram et al., 2014); and in child services, screening decisions are based on the estimated risk of adverse outcomes (Brown et al., 2019; Chouldechova et al., 2018; De-Arteaga et al., 2020; Shroff, 2017). In these applications and others, equity is a central concern.

In the machine learning community, efforts to design fair algorithms have largely focused on the predictions generated by algorithms. In particular, researchers have proposed numerous methods to constrain predictions to achieve formal statistical properties, such as parity in error rates across demographic groups (Barocas et al., 2017; Chouldechova & Roth, 2018; Corbett-Davies & Goel, 2018).

To illustrate this traditional paradigm, consider an application in which one seeks to equitably allocate a limited number of transportation vouchers to patients with upcoming medical appointments. This type of intervention has recently shown promise for improving appearance rates by alleviating the burdens many patients face in arranging transportation (Chaiyachati et al., 2018b; Vais et al., 2020), with an ultimate goal of improving health outcomes. In this example, one might use historical patient data to predict the likelihood one will miss their appointment, and then allocate transportation vouchers to those patients at highest risk. To address equity concerns, one might exclude protected attributes (e.g., race and gender) from the feature set, and may additionally constrain the predictive model to yield similar allocation rates—or, alternatively, similar error rates—across race and gender groups.

This common approach, however, suffers from several significant shortcomings. First, it may not yield an efficient strategy for allocating limited transportation resources to increase appearance rates, since those patients at highest risk of missing their appointments are not necessarily the same as those who are likely to alter their behavior in response to transportation assistance. Indeed, some prior work has shown that ride vouchers do not always improve average appearance rates (Chaiyachati et al., 2018a). In this case, as in many others like it, it is important to consider the heterogeneous causal effects of one's actions—providing transportation assistance—on downstream outcomes, like appointment appearance. Second, the strategy neglects to consider the idiosyncratic value of attending an appointment. For example, it may be more important for those with serious health conditions to attend their appointments than for patients who are generally healthy. Third, standard

[1]Department of Management Science & Engineering, Stanford University, Stanford, CA [2]Department of Computer Science, Stanford University, Stanford, CA. Correspondence to: Alex Chohlas-Wood <alexcw@stanford.edu>, Madison Coots <mcoots@stanford.edu>, Emma Brunskill <ebrun@cs.stanford.edu>, Sharad Goel <scgoel@stanford.edu>.

techniques for incorporating equity—such as blinding algorithms or equalizing error rates—may in fact harm the very groups one seeks to aid. For example, gender-blind criminal risk assessments have been shown to overestimate the risk that female defendants will recidivate; as a result, they may lead to increased detention rates for women (Skeem et al., 2016). Similarly problematic outcomes may result in our example application if one modifies or constrains the predictions in isolation.

To address the concerns outlined above, we propose an alternative, consequentialist framework to algorithmic fairness that foregrounds the results of one's decisions, rather than the predictions used to inform those choices.

In our approach, one starts by identifying the utility of different possible outcomes of a decision-making policy. For example, hospital stakeholders may indicate a preference for greater appearance rates—particularly for those with serious health conditions—and a preference for demographic diversity among recipients of the transportation assistance. These complex preferences incorporate considerations of both efficiency and equity. A large and growing body of work has shown that one can often efficiently elicit these preferences, even in high-dimensional outcome spaces (Chu & Ghahramani, 2005; Fürnkranz & Hüllermeier, 2010; Lin et al., 2020).

Next, given these preferences, we show how to learn a decision-making policy with the largest expected utility given budget constraints. Starting from a historical dataset of decisions and outcomes, we demonstrate that this optimization problem can be written as a linear program (LP) to efficiently identify optimal policies for a large and expressive family of utility functions. These utility-maximizing policies can differ considerably from those suggested by traditional approaches to algorithmic fairness, illustrating the value of our approach.

Finally, in a dynamic setting, we show that one can use our approach to efficiently learn utility-maximizing policies over time through strategic exploration. Inspired by work in multi-armed bandits, we learn these policies through optimistic exploration—where, at each step, we act according to a policy optimized under optimistic estimates of the potential outcomes under different actions. In contrast to the standard contextual multi-armed bandit setting, we consider a more complex, structured objective to account for fairness preferences and budget constraints inherent to many real-world applications. As such, our actions at each iteration are guided by solving an LP corresponding to the static case described above.

To evaluate our approach, we run a realistic simulation study. We show that our framework yields substantially better outcomes for participants during learning, and often more quickly identifies higher utility decision policies for future use, compared to standard approaches like randomized control trials that have been used in previous research (Chaiyachati et al., 2018a;b).

2. Related work

Our work draws on research in algorithmic fairness, multi-objective optimization, and contextual bandits with budgets—connections that we briefly discuss below.

Over the last several years, there has been increased attention on designing equitable machine learning systems (Ali et al., 2019; Blodgett & O'Connor, 2017; Buolamwini & Gebru, 2018; Caliskan et al., 2017; Chouldechova et al., 2018; Datta et al., 2018; De-Arteaga et al., 2019; Goodman et al., 2018; Koenecke et al., 2020; Obermeyer et al., 2019; Raji & Buolamwini, 2019; Shroff, 2017), and concomitant development of formal criteria to characterize fairness (Barocas et al., 2017; Chouldechova & Roth, 2018; Corbett-Davies & Goel, 2018). Some of the most popular definitions demand parity in predictions across salient demographic groups, including parity in mean predictions (Feldman et al., 2015) or error rates (Hardt et al., 2016). Another class of fairness definitions aims to blind algorithms to protected characteristics, including through their proxies (Chiappa & Isaac, 2018; Coston et al., 2020; Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Nyarko et al., 2021; Wang et al., 2019; Wu et al., 2019; Zhang & Bareinboim, 2018). Whereas past approaches to fairness typically focus on the predictions themselves, we foreground the consequences of actions guided by those predictions. In our running medical example, as we describe in detail in Section 4, a narrow focus on predictive parity can lead to an undesirable allocation of limited transportation resources to patients.

In many real-world settings, decision makers have competing priorities, linking our work to the large literature on learning to optimize in multi-objective environments (Zuluaga et al., 2013). Such inherent trade-offs have been recently considered in the fair machine learning community (e.g., Cai et al. (2020); Corbett-Davies et al. (2017); Rolf et al. (2020)); however, there has been little work on creating equitable learning systems that account for competing objectives.

One particularly challenging aspect of our setting is handling budget constraints (e.g., we may only be able to provide transportation assistance to a limited number of patients). Recent work has proposed methods for learning decision policies with fairness or safety constraints through reinforcement learning (Thomas et al., 2019) and contextual bandit algorithms (Metevier et al., 2019), given access to a batch of prior data. That work, however, neither addresses learning with budget constraints nor handles the exploration-exploitation trade-off required for online learning.

The challenge of budget constraints has been considered in a more general form of knapsack constraints in bandit settings. Chapter 10 of Slivkins et al. (2019) provides a recent review of such work, focusing on the primary literature, which has considered the (non-contextual) multi-armed bandit setting. Earlier work on contextual multi-armed bandits with knapsacks (Agrawal et al., 2016b; Badanidiyuru et al., 2014) provided regret bounds but lacked computationally efficient implementations. Agrawal et al. (2016a) later proved regret guarantees for linear contextual bandit with knapsacks. Wu et al. (2015) provide a computationally tractable, approximate linear programming method for online learning for contextual bandits with budget constraints, though they do not consider multi-objective optimization, a critical ingredient for equitable decision making.

3. Decision-making as optimization

We start by outlining a general, utility-based paradigm for equitable decision-making. Then, by assuming complete knowledge on the distribution of potential outcomes under actions, we present a computationally efficient approach to deriving optimal policies. (In Section 5, we use this result to address the more general problem of learning optimal policies from data.) Finally, we examine a common special case, and show that optimal decision policies take the particularly simple form of personalized threshold rules.

3.1. Problem formulation

Consider a sequential decision-making setting where, at each time step, one first observes a vector of covariates X_i drawn from a distribution $\mathcal{D}_X$ supported on a finite state space $\mathcal{X}$, and then must select one of K actions from the set $\mathcal{A} = \{a_1, \ldots, a_K\}$. For example, in our motivating application, X_i might encode an individual's demographics, medical status, access to transportation, and history of appearance, and the set of actions might specify whether or not transportation assistance is offered (here $K = 2$). In general, we allow randomized decision policies π, where the action $\pi(x)$ is (independently) drawn from a specified distribution on $\mathcal{A}$.

In practice, there are often constraints on the distribution of actions taken. For example, budget limitations might mean that at most 20% of people can be offered transportation assistance. As such, for constants $b_1, \ldots, b_K$, we require our decision policy π to satisfy $\Pr(\pi(X) = a_k) \leq b_k$, where the probability encompasses randomness in both the covariate vector X and the decisions themselves. The unconstrained case corresponds to setting $b_k = 1$.

Each action is associated with a potential outcome $Y_i(a_k)$, and, in particular, taking action $\pi(X_i)$ results in the (random) outcome $Y_i(\pi(X_i))$. For example, $Y_i(1)$ may indicate whether the i-th individual would attend their appointment if offered transportation assistance, and $Y_i(0)$ may indicate the outcome if assistance were not provided.

Now, suppose we have a real-valued function $r(x, a, y)$ that specifies the (ex-post) value of individual decisions and outcomes. In our motivating application, we might set

$$r(x, a, y) = (a + c_1 y) \cdot (1 + c_2 \cdot \mathbb{I}_{\text{severe}}(x)), \qquad (1)$$

where $a \in \{0, 1\}$ indicates whether transportation assistance is provided, $y \in \{0, 1\}$ indicates whether an individual appeared at their appointment, $\mathbb{I}_{\text{severe}}(x)$ indicates whether an individual has a severe medical condition, and the positive constants c_1 and c_2 characterize the relative values of the terms.[1] This choice of r encodes the belief that: (1) appearing at one's appointment is better than not appearing; (2) receiving transportation assistance is better than not receiving it, regardless of the outcome; and (3) the value of assistance and of appearance is greater for those with severe medical conditions.

In discussions of algorithmic fairness, special attention is often paid to groups defined by legally protected characteristics, such as race and gender. To facilitate these considerations, we allow individuals to be associated with a set of identities $s(X_i) \subseteq \mathcal{G}$, where $\mathcal{G}$ is a finite set. For example, $s(X_i)$ might specify both an individual's race and gender.

Finally, given the above setup, we define the utility $U(\pi)$ of any decision policy π to be:

$$\begin{aligned} U(\pi) &= \mathbb{E}[r(X, \pi(X), Y(\pi(X)))] \\ &\quad - \sum_{g \in \mathcal{G}} \lambda_g \| \mathcal{D}(\pi(X) \mid g \in s(X)) - \mathcal{D}(\pi(X)) \|_1, \end{aligned} \qquad (2)$$

where $\mathcal{D}(\pi(X) \mid g \in s(X))$ and $\mathcal{D}(\pi(X))$ denote the conditional and unconditional distributions of $\pi(X)$, represented as vectors in $\mathbb{R}^K$, $\|v\|_1 = \sum_{k=1}^{K} |v_k|$ is the L^1 norm, and λ_g are non-negative constants.

The first term in $U(\pi)$ captures the value directly associated with individual decisions. The second term captures the value of treatment parity across the population more broadly. For example, in addition to preferring transportation assistance policies that boost appearance rates, we might also prefer those for which a similar proportion of individuals receive assistance across neighborhoods.

Our goal is to find a utility-maximizing policy π^* that satisfies the budget constraints. Formally, we seek to solve the following optimization problem:

$$\begin{aligned} &\pi^* = \arg\max_{\pi} U(\pi) \\ &\text{subject to: } \Pr(\pi(X) = a_k) \leq b_k \quad \text{for } 1 \leq k \leq K. \end{aligned} \qquad (3)$$

[1] In Eq. (1), we do not multiply a by a constant, since the overall scale of r is arbitrary.

We next discuss settings in which these optimal policies can be efficiently derived.

3.2. Computing optimal decision policies

As a first step for computing optimal policies in real-world settings, we assume one knows the distribution of X and the conditional distribution of the potential outcomes $Y(a_k)$ given X—i.e., $\mathcal{D}(X)$ and $\mathcal{D}(Y(a_k)|X)$. In this case, we show the optimization problem in Eq. (3) can be expressed as a linear program (LP), yielding an efficient method for computing an optimal decision policy. In Section 5, we then integrate this result into a contextual bandit framework to learn optimal policies from data, even without prior knowledge of the potential outcomes.

To construct the LP, first observe that any policy π corresponds to a vector $v \in \mathbb{R}_+^{\mathcal{X}} \times \mathbb{R}_+^K$, where $v_{x,k}$ denotes the probability x is assigned to action k. Thus, the complete space of policies Π can be written as:

$$\Pi = \left\{ v \in \mathbb{R}_+^{\mathcal{X}} \times \mathbb{R}_+^K \;\middle|\; \sum_{k=1}^{K} v_{x,k} = 1 \right\},$$

and we can accordingly view the components $v_{x,k}$ of v as decision variables in our LP.

Now, in this representation, the budget constraint $\Pr(\pi(X) = a_k) \leq b_k$ in Eq. (3) can be expressed as K linear inequalities on the decision variables:

$$\sum_{x \in \mathcal{X}} \Pr(X = x) \cdot v_{x,k} \leq b_k \quad \text{for } 1 \leq k \leq K.$$

Finally, we need to express the utility $U(x)$ in linear form, which we do in two steps. First, observe that

$$\begin{aligned}&\mathbb{E}[r(X, \pi(X), Y(\pi(X))]\\ &\quad = \sum_{x,k} \mathbb{E}[r(x, a_k, Y(a_k)) \mid X = x] \cdot \Pr(X = x) \cdot v_{x,k},\end{aligned}$$

which is linear in the decision variables.

Next, note that

$$\begin{aligned}&\|\mathcal{D}(\pi(X) \mid g \in s(X)) - \mathcal{D}(\pi(X))\|_1\\ &\quad = \sum_k |\Pr(\pi(X) = k \mid g \in s(X)) - \Pr(\pi(X) = k)|\\ &\quad = \sum_k \left| \frac{1}{\Pr(g \in s(X))} \sum_{x \,:\, s(x) \ni g} \Pr(X = x) \cdot v_{x,k} \right.\\ &\qquad\qquad \left. - \sum_x \Pr(X = x) \cdot v_{x,k} \right|.\end{aligned}$$

Due to the absolute value, the expression above is not linear in the decision variables. But we can use a standard construction to transform it into an expression that is. In general, suppose we aim to maximize an objective function of the form

$$c^T x - |d^T y|, \tag{4}$$

where c and d are constant vectors. We can rewrite this optimization problem as a linear program that includes an additional (slack) variable w:

$$\begin{aligned}\text{Maximize:} \quad & c^T x - w\\ \text{Subject to:} \quad & d^T y \leq w\\ & d^T y \geq -w\\ & w \geq 0.\end{aligned} \tag{5}$$

For completeness, we provide a proof of this equivalence in Appendix A.

Putting together the pieces above, we now write our policy optimization problem in Eq. (3) as the following linear program:

Maximize:

$$\sum_{x,k} \mathbb{E}[r(x, a_k, Y(a_k)) \mid X = x] \cdot \Pr(X = x) \cdot v_{x,k} - \sum_{g,k} \lambda_g w_{g,k}$$

Subject to:

$$\begin{aligned}& v_{x,k},\ w_{g,k} \geq 0 \quad \forall\, x, k, g\\ & \sum_k v_{x,k} = 1 \quad \forall\, x\\ & \sum_x \Pr(X = x) \cdot v_{x,k} \leq b_k \quad \forall\, k\\ & \frac{1}{\Pr(g \in s(X))} \sum_{x \,:\, s(x) \ni g} \Pr(X = x) \cdot v_{x,k} - \sum_x \Pr(X = x) \cdot v_{x,k} \leq w_{g,k} \quad \forall g, k\\ & \frac{1}{\Pr(g \in s(X))} \sum_{x \,:\, s(x) \ni g} \Pr(X = x) \cdot v_{x,k} - \sum_x \Pr(X = x) \cdot v_{x,k} \geq -w_{g,k} \quad \forall g, k.\end{aligned}$$

Our approach above is a computationally efficient method for finding optimal decision polices. In theory, linear programming is (weakly) polynomial in the size of the input: $O(|\mathcal{X}|K + |\mathcal{G}|K)$ variables and constraints in our case. In practice, using open-source software running on conventional hardware, we find it takes less than a second to solve random instances of the problem on a state space of size $|\mathcal{X}| = 1{,}000$ with $|\mathcal{G}| = 10$ groups and $K = 5$ treatment arms.[2]

[2]We used the Glop linear optimization solver, as im-

Our optimization approach is also quite flexible, and can accommodate a wide range of utility functions even beyond the specific form we present in Eq. (2). For example, we could similarly include terms in $U(\pi)$ that encode a preference for parity in the expected individual-level reward $\mathbb{E}[r(X, \pi(X), Y(\pi(X))]$ across groups. Further, rather than focusing on parity, we could set group-specific target distributions for the assignments or rewards.

We note, however, that to express our general optimization problem as a tractable LP, it is important for the group preferences encoded in $U(\pi)$ to be written in terms of L^1 norms. Although an L^2-based utility could be expressed as a quadratic program (QP), such optimization problems are typically much more computationally challenging to solve. More generally, if we were to allow arbitrary utility functions, then finding an optimal decision policy is NP-hard, as we show in Appendix B.

3.3. Personalized threshold rules

The LP described in Section 3.2 yields a solution to our general decision-making problem, with an arbitrary number of treatment arms and a potentially complex utility function. Here we show that in the common case of $K = 2$ treatments (e.g., with the options corresponding to whether or not one provides transportation assistance), optimal decision policies can be expressed in a simple, interpretable form. Moreover, for a reward function r that decomposes into aggregate and individual components—as in Eq. (1)—we can view optimal policies as personalized threshold rules.

Theorem 1. *In the setting of Section 3.1, suppose $K = 2$ and $|s(x)| = 1$ (i.e., $\mathcal{G}$ partitions $\mathcal{X}$). Let*

$$\Delta(x) = \mathbb{E}[r(x, a_1, Y(a_1)) - r(x, a_0, Y(a_0)) \mid X = x].$$

Then, for group-specific constants t_g and p_g, there exists an optimal decision policy π^* of the form

$$\Pr(\pi^*(x) = a_1) = \begin{cases} 1 & \Delta(x) > t_{s(x)} \\ p_{s(x)} & \Delta(x) = t_{s(x)} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

We include the proof of Theorem 1 in Appendix C. As an immediate corollary, we have the following result.

Corollary 1. *In addition to the assumptions of Theorem 1, suppose r decomposes as*

$$r(x, a, y) = c(x)u(a, y) + d(x),$$

where $c(x) > 0$. Let

$$\Delta_u(x) = \mathbb{E}[u(a_1, Y(a_1)) - u(a_0, Y(a_0)) \mid X = x].$$

Then, for group-specific constants t_g and p_g, there exists an optimal decision policy π^* of the form

$$\Pr(\pi^*(x) = a_1) = \begin{cases} 1 & \Delta(x) > t_{s(x)}/c(x) \\ p_{s(x)} & \Delta(x) = t_{s(x)}/c(x) \\ 0 & \text{otherwise.} \end{cases}$$

In Corollary 1, $u(a, y)$ can be viewed as the baseline value of actions and outcomes—irrespective of individual characteristics—with $\Delta_u(x)$ incorporating the likely outcomes for an individual under the two possible actions. For example, as in Eq. (2), u may represent the individual-agnostic value of receiving transportation assistance and of appearing at one's medical appointment, with $\Delta_u(x)$ largest for those individuals for whom the assistance would most increase their marginal likelihood of appearance. Now, if r decomposes as in the corollary, then we can view the optimal policy as providing transportation assistance to those with sufficiently large marginal gain, where the threshold for assistance $t_{s(x)}/c(x)$ accounts both for group membership (to satisfy parity considerations) and idiosyncratic utility adjustments (e.g., those associated with single caregivers).

4. Principled trade-offs

The structure of our utility function in Eq. (2) captures a common trade-off in decision problems. On one hand, one seeks to maximize the total number of realizations of a desired outcome (e.g., hospital appearances); on the other hand, however, one also seeks to maintain treatment parity across groups in a population. The weights λ_g encode the relative value of these two objectives, allowing us to make difficult decisions in a principled manner that align with our underlying preferences.

To explore this trade-off, we return again to our motivating application of allocating transportation vouchers to hospital patients. For simplicity, we consider a patient population with two equally sized groups that have identical appearance rates in the absence of transportation assistance. However, one group (which we refer to as the targeted group) has a lower average treatment effect, and so a preference for parity introduces a tension between maximizing total appearances and equitably allocating assistance across the two groups. We describe the data-generating process for this synthetic population in detail in Appendix D.

In Figure 1, for budget $b = 1/3$, we trace out the Pareto frontier for this example, which shows how the appearance rate (on the vertical axis) varies under policies that optimally allocate transportation assistance to a given fraction of the targeted patient population (on the horizontal axis). By Theorem 1, each point on the frontier corresponds to a threshold policy that provides assistance to patients with the largest treatment effects in each group, subject to the demographic and budget constraints.

plemented in Google OR-Tools (https://developers.google.com/optimization/).

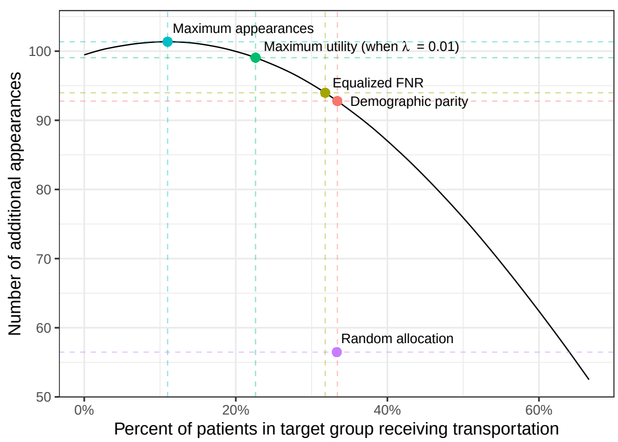

Figure 1. The Pareto frontier for a stylized population model, showing the trade-off between appearances and treatment rates in the target group. The vertical axis shows expected additional appearances relative to a policy that does not provide transportation assistance to any patients. Under this model, with the decision-maker's stated preference for $\lambda = 0.01$, common heuristics (e.g. maximizing appearances, and demanding demographic or error-rate parity) lead to sub-optimal policies.

The point at the crest of the curve (in blue) achieves the highest number of overall appearances, and corresponds to a policy without parity constraints (i.e., with $\lambda = 0$). Under this policy, a lower share of patients in the targeted group receives assistance than in the non-targeted group. If one seeks perfect demographic parity—meaning $1/3$ of patients in both groups receive assistance, corresponding to $\lambda = \infty$—then that means having fewer overall appearances, as indicated by the red point on the curve. For any non-trivial λ between 0 and ∞, the optimal policy interpolates between these two extremes, balancing appearance rates with parity considerations. For example, the green point on the plot indicates the optimal policy for $\lambda = 0.01$.

This simple example helps illustrate the value of viewing decisions from a consequentialist perspective, complementing the rule-based, deontological approach that has been the focus of much past work on algorithmic fairness. Although the extremes of maximizing appearances and requiring strict demographic parity are perhaps reasonable heuristics, they can obscure the trade-offs inherent to many policy problems.

As a final comparison, we plot points on the curve corresponding to random allocation (in purple) and equal false negative rates (FNR) between groups (in brown).[3] Random allocation results in demographic parity, but lies below the Pareto frontier, meaning appearance rates are lower than if one were to more judiciously allocate transportation assistance. The equal FNR point lies on the curve, meaning it happens to be optimal for a specific choice of λ. But a rule that simply demands error-rate parity—as opposed to selecting λ in a more principled manner—can result in a sub-optimal balance of appearances with the demographic makeup of transportation recipients, relative to the underlying preferences of the policymaker.

[3] In this case, equal FNR means that $\Pr(\pi = 0 \mid Y(0) = 0, Y(1) = 1, G = g) = \Pr(\pi = 0 \mid Y(0) = 0, Y(1) = 1)$. That is, among those who would benefit from the assistance, an equal proportion do not receive it in both groups.

5. Learning optimal policies

To solve our policy optimization problem, we have thus far assumed perfect knowledge of

$$f(x, a) = \mathbb{E}[r(x, a_k, Y(a_k)) \mid X = x],$$

for all values of x and k. In reality, however, these quantities must be estimated from observed data. One common approach is to run a randomized controlled trial (RCT) to estimate the effect of actions on individuals. But RCTs suffer from a major drawback: they favor exploration over exploitation. In our running example of providing transportation vouchers to patients, this can mean distributing limited resources to patients who do not need them.

In contrast to RCTs, contextual bandit algorithms are often designed to maximize expected utility while learning, which typically involves estimating the potential performance of each action a and using that information to accrue benefits. To efficiently learn decision policies in the real world, we now outline our procedure to integrate the LP formulation from Section 3.2 with three common contextual bandit approaches: ε-greedy, Thompson sampling, and upper confidence bound (UCB), as described in Algorithm 1.[4] Later, in Section 5.1, we demonstrate the efficacy of this procedure with experiments on synthetic data.

At a high level, at each step i, our ε-greedy approach first estimates $f(x, a)$ using the maximum likelihood estimate of a chosen parametric family, and uses this estimate to find the optimal policy π_i^* with our LP. Then, with probability $1 - \varepsilon$, we treat the i-th individual according to π_i^*; otherwise, with probability ε, we take action a_k with probability proportional to the budget $b_{k,i}^*$. Our Thompson sampling approach maintains a posterior over the parameters of a model of the potential outcomes $\hat{f}(x, a)$, samples from this posterior, uses the posterior draw in the LP formulation to compute a policy π_i^*, and then treats the i-th individual according to π_i^*. Finally, under our UCB approach, we compute π_i^*

[4] For simplicity, we assume knowledge of the covariate distribution $\Pr(X = x)$, which is often easily obtained from historical data, even in the absence of past interventions. If historical data are not available, the covariate distribution can instead be estimated from the sample of individuals observed during the decision-making process.

Algorithm 1 Policy Learning

input: Actions a_k, budgets b_k, parity preferences λ_g, reward function r, covariate distribution $\Pr(X = x)$, group membership function s, bandit algorithm, ℓ
initialize: Randomly treat first ℓ people
for each subsequent individual i **do**
  Set $\mathcal{D}_i := \{(X_j, A_j, Y_j)\}_{j=1}^{i-1}$, where X_j, A_j, and Y_j denote the covariates, actions, and outcomes for previously seen individuals
  Estimate $f(x, a)$ with a parametric family of functions $g(x, a; \theta)$ fit on $\mathcal{D}_i$
  if ε-greedy **then**
    $\hat{f}(x, a) := g(x, a; \hat{\theta}_i)$, where $\hat{\theta}_i$ is the MLE
  else if Thompson sampling **then**
    $\hat{f}(x, a) := g(x, a; \hat{\theta}_i^*)$, where $\hat{\theta}_i^*$ is drawn from the posterior of $\hat{\theta}_i$
  else if UCB **then**
    $\hat{f}(x, a) :=$ the α-percentile of the posterior of $g(x, a; \hat{\theta}_i)$
  end if
  Compute nominal budgets $b_{k,i}^*$ according to Eq. (7)
  Find solution π_i^* of the LP in Section 3.2 with input values $\hat{f}(x, a)$, s, λ_g, $b_{k,i}^*$, and $\Pr(X = x)$
  if ε-greedy **and** BERNOULLI$(\varepsilon) == 1$ **then**
    Take random action A_i where $\Pr(A_i = a_k) \propto b_{k,i}^*$
  else
    Take action $A_i \sim \pi_i^*(X_i)$
  end if
  Observe outcome Y_i
end for

by solving the LP with an optimistic estimate of $f(x, a)$ (e.g., using the 95th percentile of the posterior distribution of $\hat{f}(x, a)$).

Because our inferred policies π_i^* evolve over time, they are not guaranteed to adhere to the budget constraints. To account for this possibility, if we find ourselves spending more on an action than is budgeted, we gradually lower the nominal budget for that action until it meets the target budget (and vice versa for underspending). Specifically, for each treatment a_k at each iteration i, we compute a new budget $b_{k,i}^*$:

$$b_{k,i}^* = b_k \cdot \frac{\sum_{j=1}^{i-1} b_{k,j}^*}{\sum_{j=1}^{i-1} \mathbb{I}(A_j = a_k)}, \tag{7}$$

where A_j is the action taken on the j-th individual, and b_k is the target budget for the action.

5.1. Simulation study

To evaluate our learning approach above, we conducted a simulation study on a synthetic population of patients. In this example, patients can receive one of three mutually exclusive treatments a_k: a free ride, a free transit voucher, or no transportation assistance. We fix our budgets for free transit vouchers at 20% of the population, and for free rides at 5% of the population.

Our hypothetical population is comprised of two demographic groups. The utility of a policy is described by Eq. (2), where we set $r(x, a, y) = y$ and $\lambda_g = 0.02$. This choice yields an oracle policy that balances between maximizing appearances and achieving parity in the distribution of transportation assistance across groups. Both the data-generating process for this population and additional experiment parameters are described in detail in Appendix E.

We compare our contextual bandit approaches against several baselines. First, we compare to an RCT, in which treatment is randomly selected (in accordance with the budget) throughout the entire experiment. We also include variations on this approach, where we run an RCT on the first n individuals, and then follow the optimal policy estimated at individual n for the rest of the sample. We compare all approaches against an oracle that can observe the true expected appearance probabilities.[5]

We repeat this evaluation on 2,000 independent instances of the synthetic population, and compare the performance of all approaches using two different metrics. Our main two bandit approaches—Thompson sampling and UCB—not only significantly reduce regret when compared to an RCT during the training/learning process (Figure 2), but also learn policies that, if used for future populations, would equal or outperform all other approaches (Figure 3). In our setting, certain individuals respond better to transit vouchers when compared to rides, even though vouchers are more plentiful and rides are (on average) more potent. The best-performing methods quickly exploit this pattern and similar heterogeneity in the population to efficiently and equitably allocate resources, as we discuss in more detail in Appendix E.

In contrast to our two main bandit algorithms, the ε-greedy approach also manages to reduce regret, but is slower to learn a near-oracle policy. The RCT and its variations illustrate the limits of the conventional randomized approach. For example, it is possible to learn a near-oracle policy using a classic RCT, but this incurs substantial regret during the experiment. Though it is possible to reduce this regret by ending the RCT early, these alternatives do not learn a near-oracle policy.

[5] In expectation, the oracle will meet our target budgets, so we do not apply the adjustments from Eq. (7).

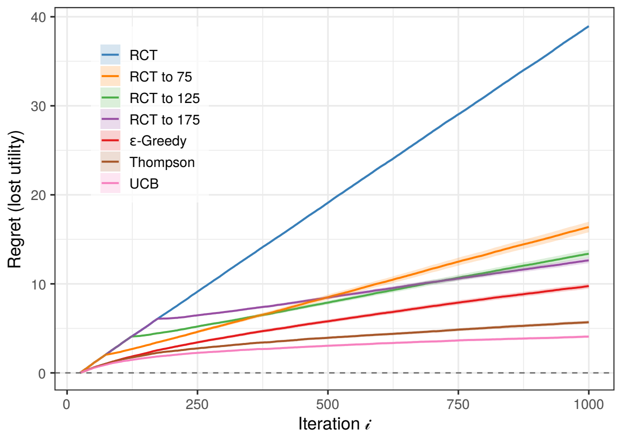

Figure 2. Mean regret, across 2,000 simulations, incurred by different learning approaches. We define regret here as the difference between the observed utility and the utility obtained by an oracle during the same experiment. Uncertainty bands represent 95% intervals for the mean. We note that the three bandit approaches—ε-greedy, Thompson sampling, and UCB—incur substantially less regret than the RCT. It is possible to reduce the regret incurred during an RCT by stopping the RCT early, and following the optimal estimated policy from that point forward. However, these stop-early RCT approaches produce worse policies than other approaches (Figure 3).

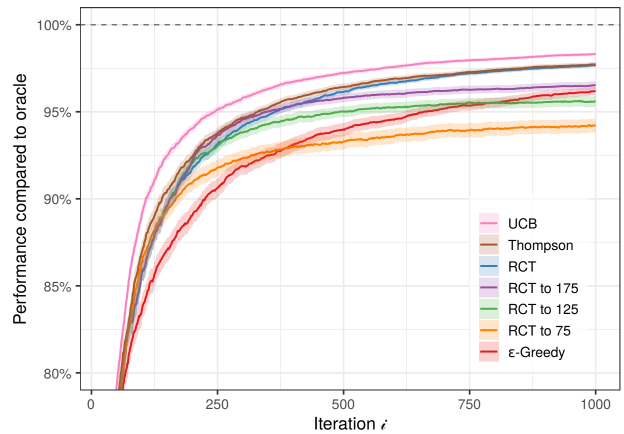

Figure 3. Mean performance, across 2,000 simulations, of optimal policies estimated with data available at each iteration i. Performance is defined as the additional utility obtained by a policy over a baseline of no treatment, with 100% indicating this quantity for the oracle policy. Uncertainty bands represent 95% intervals around the mean. In contrast with Figure 2—which shows the expected loss in utility while learning—this plot shows the expected loss in utility if one were to follow the optimal estimated policy after iteration i on subsequent individuals. Thompson sampling and UCB generate policies that are as good as (or better than) a conventional RCT. In contrast, the ε-greedy approach and the stop-early versions of the RCT generate policies that are slower to (or may never) reach the performance of the best approaches.

6. Discussion

We have outlined a consequentialist framework for equitable algorithmic decision-making. Our approach foregrounds the role of an expressive utility function that captures preferences for both individual- and group-level outcomes. In this conceptualization, we explicitly consider the inherent trade-offs between competing objectives in many real-world problems. For example, in our running example of allocating transportation assistance to patients, there is tension between maximizing appearance rates and ensuring an equitable distribution of benefits. Popular rule-based approaches to algorithmic fairness—such as requiring equal false negative rates across groups—implicitly balance these competing objectives in ways that may be at odds with the actual preferences of stakeholders. Our approach, in contrast, requires one to confront the consequences of difficult choices, and, in the process, helps one improve those decisions.

For a rich class of utility functions, we showed that one can efficiently learn optimal decision policies by coupling ideas from the contextual bandit and optimization literatures. For example, with our UCB-based algorithm, we do so by repeatedly solving a linear program under optimistic estimates of the potential outcomes of actions. In a simulation study, we showed that this strategy can outperform common alternatives, including learning through randomized controlled trials or acting greedily based on the available information.

In this work, we have assumed access to a well-specified utility function that reflects stakeholder preferences. In practice, inferring this utility is a complex task in its own right. There are, however, several established techniques to elicit multi-faceted preferences less directly. One family of approaches selects pairs of similar realistic scenarios, asks stakeholders to pick their preferred outcome, and infers their preferences from these choices (Chu & Ghahramani, 2005; Fürnkranz & Hüllermeier, 2010; Lin et al., 2020).

Another challenge—particularly relevant in the dynamic setting—is accounting for delayed outcomes. In our running example, we may choose to offer transportation assistance to a patient days or weeks before their appointment date. As a result, there may be large gaps between when an action is taken and when we observe its outcome. One way to address this issue is through the use of *proxies* or *surrogates*, in which intermediate outcomes are used as a temporary stand-in for the eventual outcome of interest (Athey et al., 2019). For example, with transportation assistance to patients, one might use intermediate responses (like a patient's confirmation to attend their appointment) as a proxy for appearance. Another strategy is to reduce the budget for costly actions, effectively limiting the resources spent while waiting to observe outcomes.

Algorithms impact individuals both through the decisions they guide and the outcomes they engender. Looking forward, we hope our work helps to elucidate the subtle interplay between actions and consequences, and, in turn, furthers the design and deployment of equitable algorithms.

Acknowledgements

We thank Johann Gaebler, Jonathan Lee, Hamed Nilforoshan, and Julian Nyarko for helpful comments. This work was supported in part by grants from the Stanford Impact Labs and the Stanford Institute for Human-Centered Artificial Intelligence. Code to replicate our analysis is available online at: https://github.com/stanford-policylab/learning-to-be-fair.

References

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 599–600, 2016a.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016b.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095*, 2019.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134. PMLR, 2014.

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

Blodgett, S. L. and O'Connor, B. Racial disparity in natural language processing: A case study of social media African-American English. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, KDD*, 2017.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Cadigan, T. P. and Lowenkamp, C. T. Implementing risk assessment in the federal pretrial services system. *Fed. Probation*, 75:30, 2011.

Cai, W., Gaebler, J., Garg, N., and Goel, S. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 22–28, 2020.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Chaiyachati, K. H., Hubbard, R. A., Yeager, A., Mugo, B., Lopez, S., Asch, E., Shi, C., Shea, J. A., Rosin, R., and Grande, D. Association of rideshare-based transportation services and missed primary care appointments: a clinical trial. *JAMA internal medicine*, 178(3):383–389, 2018a.

Chaiyachati, K. H., Hubbard, R. A., Yeager, A., Mugo, B., Shea, J. A., Rosin, R., and Grande, D. Rideshare-based medical transportation for medicaid patients and primary care show rates: a difference-in-difference analysis of a pilot program. *Journal of general internal medicine*, 33 (6):863–868, 2018b.

Chiappa, S. and Isaac, W. S. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pp. 3–20. Springer, 2018.

Chouldechova, A. and Roth, A. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148. PMLR, 2018.

Chu, W. and Ghahramani, Z. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pp. 137–144, 2005.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

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.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

Datta, A., Datta, A., Makagon, J., Mulligan, D. K., and Tschantz, M. C. Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency*, pp. 20–34, 2018.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128. ACM, 2019.

De-Arteaga, M., Fogliato, R., and Chouldechova, A. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Fürnkranz, J. and Hüllermeier, E. Preference learning and ranking by pairwise comparison. In *Preference learning*, pp. 65–82. Springer, 2010.

Gelman, A. and Su, Y.-S. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2020. URL https://CRAN.R-project.org/package=arm. R package version 1.11-1.

Goel, S., Shroff, R., Skeem, J. L., and Slobogin, C. The accuracy, equity, and jurisprudence of criminal risk assessment. *Equity, and Jurisprudence of Criminal Risk Assessment (December 26, 2018)*, 2018.

Goodman, S. N., Goel, S., and Cullen, M. R. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Latessa, E. J., Lemke, R., Makarios, M., and Smith, P. The creation and validation of the ohio risk assessment system (oras). *Fed. Probation*, 74:16, 2010.

Leo, M., Sharma, S., and Maddulety, K. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

Lin, Z. J., Obeng, A., and Bakshy, E. Preference learning for real-world multi-objective decision making. 2020.

Metevier, B., Giguere, S., Brockman, S., Kobren, A., Brun, Y., Brunskill, E., and Thomas, P. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

Milgram, A., Holsinger, A. M., Vannostrand, M., and Alsdorf, M. W. Pretrial risk assessment: Improving public safety and fairness in pretrial decision making. *Fed. Sent'g Rep.*, 27:216, 2014.

Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nyarko, J., Goel, S., and Sommers, R. Breaking taboos in fair machine learning: An experimental study. 2021.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, volume 1, 2019.

Rolf, E., Simchowitz, M., Dean, S., Liu, L. T., Bjorkegren, D., Hardt, M., and Blumenstock, J. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, pp. 8158–8168. PMLR, 2020.

Shroff, R. Predictive analytics for city agencies: Lessons from children's services. *Big data*, 5(3):189–196, 2017.

Skeem, J., Monahan, J., and Lowenkamp, C. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5):580, 2016.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Thomas, P. S., da Silva, B. C., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Vais, S., Siu, J., Maru, S., Abbott, J., Hill, I. S., Achilike, C., Wu, W.-J., Adegoke, T. M., and Steer-Massaro, C. Rides for refugees: A transportation assistance pilot for women's health. *Journal of immigrant and minority health*, 22(1):74–81, 2020.

Wang, Y., Sridhar, D., and Blei, D. M. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

Wu, H., Srikant, R., Liu, X., and Jiang, C. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 28:433–441, 2015.

Wu, Y., Zhang, L., Wu, X., and Tong, H. Pc-fairness: A unified framework for measuring causality-based fairness. *arXiv preprint arXiv:1910.12586*, 2019.

Zhang, J. and Bareinboim, E. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zuluaga, M., Sergent, G., Krause, A., and Püschel, M. Active learning for multi-objective optimization. In *International Conference on Machine Learning*, pp. 462–470. PMLR, 2013.

Appendices

A. Absolute value in an LP objective

If (x^*, y^*, w^*) is a solution to the LP in Eq. (5), then we claim (x^*, y^*) is a solution to the original optimization problem in Eq. (4). Let OPT_{abs} and OPT_{LP} denote the optima of Eqs. (4) and (5) above. Now, since $w = |d^T y|$ satisfies the LP constraints, $\text{OPT}_{\text{abs}} \leq \text{OPT}_{\text{LP}}$. Conversely, because the LP objective function decreases in w, if $d^T y^* \geq 0$, then $w^* = d^T y^*$ (since $d^T y^* \leq w$, and the other two constraints are immediately satisfied in this case). On the other hand, if $d^T y^* \leq 0$, then $w^* = -d^T y^*$ (since $d^T y^* \geq -w$). Thus, in either case, $w^* = |d^T y^*|$, which implies that $\text{OPT}_{\text{abs}} = \text{OPT}_{\text{LP}} = c^T x^* - |d^T y^*|$.

B. Hardness of Policy Optimization

Proposition 1. *If we allow arbitrary utility functions U in Eq. (3), then finding an optimal policy is NP-hard.*

Proof. We reduce to the NP-hard subset sum problem. Given integers $x_1, \ldots, x_n$, consider the policy optimization problem for $K = 2$ actions and no budget constraints (i.e., $b_k = 1$), with utility

$$U(\pi) = \begin{cases} 1 & A(\pi) \neq \emptyset \,\wedge\, \sum_{i \in A(\pi)} x_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $A(\pi) = \{i : \Pr(\pi(x_i) = a_1) = 1\}$. Then $\max_\pi U(\pi) = 1$ if and only if there exists a non-trivial subset of the integers $\{x_1, \ldots, x_n\}$ that sums to zero, establishing the claim. □

C. Proof of Theorem 1

Proof. We start by rewriting the utility $U(\pi)$ as

$$\begin{aligned} U(\pi) = &\sum_x \mathbb{E}[r(x, a_0, Y(a_0)) \mid X = x] \cdot \Pr(X = x) \\ &+ \sum_x \Delta(x) \cdot \pi(x) \cdot \Pr(X = x) \\ &- 2 \sum_{g \in \mathcal{G}} \lambda_g \delta_g(\pi), \end{aligned}$$

where

$$\delta_g(\pi) = |\Pr(\pi(X) = a_1 \mid s(X) = g) - \Pr(\pi(X) = a_1)|.$$

Now, for any policy π, we construct a threshold policy $\tilde{\pi}$ of the form in Eq. (6) by assigning to action a_1 those x in each group g having the largest values of $\Delta(x)$ such that

$$\Pr(\tilde{\pi}(X) = a_1 \mid s(X) = g) = \Pr(\pi(X) = a_1 \mid s(X) = g).$$

By construction, $\delta_g(\tilde{\pi}) = \delta_g(\pi)$, and

$$\sum_x \Delta(x) \tilde{\pi}(x) \Pr(X = x) \geq \sum_x \Delta(x) \pi(x) \Pr(X = x).$$

Consequently, $U(\tilde{\pi}) \geq U(\pi)$, establishing the result. □

D. Experiment details for Section 4

We consider a patient population with one observable covariate $X_i \sim \text{Unif}(0, 1)$, and two equally sized groups $G_i \sim \text{Bernoulli}(0.5)$ that have identical appearance rates in the absence of transportation assistance, but which, on average, respond differently to the assistance. Specifically, for actions $a \in \{0, 1\}$, potential outcomes in this stylized model are generated according to:

$$Y_i(a) = \mathbb{I}(U_i \leq \text{logit}^{-1}((1 + a)X_i + (1 - G_i)X_i a - 1)),$$

where $\mathbb{I}(\cdot) \in \{0, 1\}$ indicates whether its argument is true, and $U_i \sim \text{Unif}(0, 1)$ is a latent, individual-level covariate that ensures $Y_i(0) \leq Y_i(1)$.

For this example, we use the following utility function:

$$\begin{aligned} U(\pi) &= \mathbb{E}[Y(\pi)] - \lambda \sum_{g \in \{0,1\}} \|\mathcal{D}(\pi \mid G = g) - \mathcal{D}(\pi)\|_1 \\ &= \mathbb{E}[Y(\pi)] - 4\lambda |\Pr(\pi = 1 \mid G = 1) - \Pr(\pi = 1)| \\ &= \mathbb{E}[Y(\pi)] - 4\lambda |\Pr(\pi = 1 \mid G = 1) - b|, \end{aligned}$$

where $b = \Pr(\pi = 1)$ is our budget.

E. Experiment details for Section 5.1

We create a synthetic population of patients according to the following structural equation model. We imagine individuals have three observable covariates, X_a, X_d, X_m, which can be viewed as describing one's age, proximity to transit, and income, respectively. In addition, each individual is equally likely to belong to one of two groups, with group membership denoted by G. Specifically, we have

$$\begin{aligned} X_a, X_d, X_m &\sim \textsc{Unif}(0, 1) \\ G &\sim \textsc{Bernoulli}(0.5) \end{aligned}$$

Next, we define three potential outcomes for each individual, corresponding to appearance in the absence of assistance ($a = 0$), appearance if provided a transit voucher ($a = 1$), and appearance if provided a free ride ($a = 2$). We do so in terms of the following structural equation:

$$\begin{aligned} & f_Y(a, x_a, x_d, x_m, g, u) \\ & = g \cdot \mathbb{1}(a=0) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a)) \\ & + g \cdot \mathbb{1}(a=1) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a + 2x_d)) \\ & + g \cdot \mathbb{1}(a=2) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a + 4x_m)) \\ & + (1-g) \cdot \mathbb{1}(a=0) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a)) \\ & + (1-g) \cdot \mathbb{1}(a=1) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a + x_d)) \\ & + (1-g) \cdot \mathbb{1}(a=2) \cdot \mathbb{1}(u \le \text{logit}^{-1}(-x_a + 2x_m)). \end{aligned} \tag{8}$$

Finally, for a latent variable $U \sim \text{UNIF}(0, 1)$, we define the potential outcomes:

$$Y(a) = f_Y(a, X_a, X_d, X_m, G, U).$$

This structure ensures that $Y(0) \le Y(1)$ and that $Y(0) \le Y(2)$, meaning that receiving any form of assistance is always better than no assistance. Further, the type of assistance—transit voucher or free ride—that is best for each individual varies across the population. Finally, although the base rate of appearance in the absence of assistance is the same across groups, one group ($g = 1$) is systematically more responsive to the assistance. As a result, a preference for parity is at tension with simply allocating assistance to those who would benefit the most.

As described in the main text, the utility U is defined by Eq. (2), where we set $r(x, a, y) = y$ and $\lambda_g = 0.02$. In other words,

$$\begin{aligned} U(\pi) &= \mathbb{E}[Y(\pi(X))] \\ &\quad - \sum_{g \in \{0,1\}} \lambda_g \|\mathcal{D}(\pi(X) \mid G = g) - \mathcal{D}(\pi(X))\|_1. \end{aligned}$$

The first term in U is the expected number of patients that would show up under the policy π, and the second term captures our parity preferences. The constant λ_g was chosen so that the oracle policy exhibited a balance between perfect demographic parity and simple appearance maximization.

When estimating $f(x, a)$ during policy learning, we use a logistic regression with the same functional form as the data-generating process above. We started each of our experiments with a small warm-up phase of 25 people. During this period, the first three patients seen from each group were assigned to $a = 0$, $a = 1$, and $a = 2$, respectively. All other individuals were assigned to one of the three actions uniformly at random. The treatments during this warm-up period are not included in the budget adjustment calculation for $b^*_{k,i}$.

Our learning framework formally relies on having a discrete covariate space, but our synthetic population has continuous covariates. To address this mismatch, we transfer our continuous setting to the discrete setting in two steps. First, at the start of our experiments, we draw one random sample $\mathcal{C}$ of $n = 1{,}000$ patients, and approximate the full population by a discrete distribution over this observed sample, with each patient assigned probability $1/n$. Now, the policies we construct (i.e., those produced by our LP) are technically defined only for individuals having covariates matching those of a patient in the initial sample $\mathcal{C}$. Consequently, when mak-

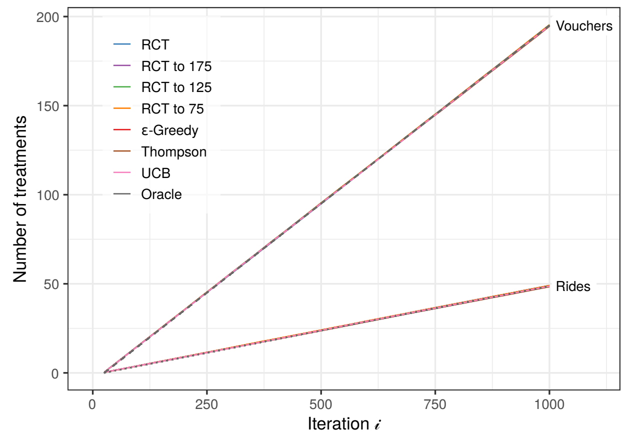

Figure E1. Mean number of vouchers and rides provided across 2,000 simulations. The voucher and ride budgets are illustrated with a dashed and dotted gray line, respectively.

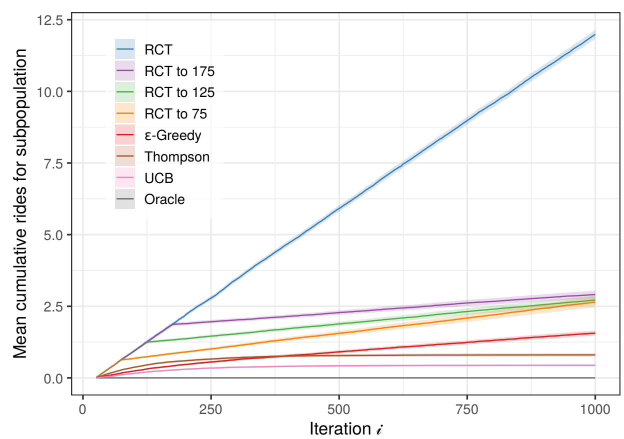

Figure E2. Mean cumulative rides, across 2,000 simulations, among individuals where $X_d > 2 \cdot X_m$ as each simulation progresses. As a direct consequence of our population model given by Eq. (8), the oracle never treats these individuals with rides. Our two main bandit methods, UCB and Thompson sampling, learn to mimic the oracle faster than other policies by quickly refraining from treating this population with rides. The structure of this chart mimics the regret plot in Figure 2.

ing decisions for a new individual, we act according to the learned policy for the most similar patient in $\mathcal{C}$, where similarity is defined in terms of estimated potential outcomes. Specifically, for a patient with covariates x, we define its nearest neighbor $\mathrm{NN}(x)$ to be:

$$\mathrm{NN}(x) = \arg\min_{x' \in \mathcal{C}} \|\hat{f}(x', \cdot) - \hat{f}(x, \cdot)\|_2.$$

Then, for any policy π defined on $\mathcal{C}$, we extend it to a policy $\tilde{\pi}$ on the full population by setting $\tilde{\pi}(x) = \pi(\mathrm{NN}(x))$.

For the ε-greedy model, we set $\varepsilon = 0.1$. For both UCB and Thompson sampling, we use the default weakly informative priors provided by the `sim` function in `arm` (Gelman & Su, 2020). For UCB, we used the 95th percentile estimate of the posterior of $g(x, a, \hat{\theta}_i)$.

Eq. (7) describes a method for constraining allocations to achieve the desired budgets in expectation. In Figure E1, we show that this approach works well in practice.

Our two bandit methods strategically exploit structural features of our population model, as we illustrate with a specific example. The form of our population model in Eq. (8) allows for some treatments to be strictly better for specific subpopulations. For example, individuals with $X_d > 2X_m$ will always respond to the (plentiful) transit voucher $a = 1$ more potently than to the (limited) ride assistance $a = 2$. For these individuals, our bandit algorithms learn it is disadvantageous to provide these individuals with a ride. In contrast, other approaches (the RCT in particular) continue to expend limited rides on these individuals (Figure E2).