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

Alex Chohlas-Wood[1], Madison Coots[1], Henry Zhu[2],
Emma Brunskill[2], and Sharad Goel[3]

[1] *Management Science & Engineering, Stanford University*
[2] *Computer Science, Stanford University*
[3] *Kennedy School, Harvard University*

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 ignores the downstream decisions and outcomes that predictions affect, and, as a result, can induce unexpected harms. Here we present an alternative framework for fairness that directly anticipates the consequences of decisions. 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 and Lowenkamp, 2011, Latessa et al., 2010, Goel et al., 2018,

Milgram et al., 2014]; in healthcare, algorithms identify which individuals will receive limited resources, including HIV prevention counseling and kidney replacements [Wilder et al., 2021, Friedewald et al., 2013]; 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 themselves. 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 and Roth, 2018, Corbett-Davies and Goel, 2018].

To illustrate this traditional paradigm, suppose a policymaker seeks to help individuals attend appointments (e.g., medical visits or court dates) by equitably providing government-sponsored transportation to and from their appointment. To design this transportation assistance program, one might use historical data to estimate the likelihood individuals will miss their appointment, and then allocate assistance to those at highest risk. In popular approaches for designing fair algorithms, one might also exclude protected attributes (e.g., race and gender) from the feature set, and may additionally constrain the predictive model to yield similar error rates across race and gender groups.

This common approach, however, suffers from several significant shortcomings. First, 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 clients who are generally healthy. Second, standard 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 recidivate, which may thus lead to increased detention rates for women [Skeem et al., 2016]. Third, it may not yield an efficient strategy for allocating limited transportation resources to increase appearance rates, since those 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 untargeted rideshare assistance may not 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.

To address the concerns outlined above, we propose an alternative, consequentialist frame-

work to algorithmic fairness that foregrounds the results of one's decisions, rather than the predictions used to inform those choices. This consequentialist approach stands in contrast with deontological approaches, which focus primarily on the process through which decisions are made, rather than the consequences of those decisions.

In our approach, one starts by identifying the utility of different possible outcomes of a decision-making policy. For example, courts may value both efficiency and equity, preferring policies that achieve both high appearance rates as well as demographic diversity among recipients of transportation assistance. Then, given these complex preferences, we learn a decision-making policy with the largest expected utility given budget constraints. For a large and expressive family of utility functions, we show that optimal decision policies can be derived by solving a linear program (LP) that uses stakeholder preferences and historical data on decisions and outcomes. In comparison, traditional approaches to algorithmic fairness—which do not explicitly consider the consequences of decisions—typically yield sub-optimal policies, illustrating the value of our approach.

Policymakers often choose to launch new programs without any historical data on how their proposed treatments impact outcomes. We show how one can efficiently learn utility-maximizing policies in this scenario. We first demonstrate that static experimental designs (i.e., those that are nonadaptive to observed outcomes from earlier participants, such as randomized controlled trials) commonly used in these settings [Chaiyachati et al., 2018a,b] would be feasible with our framework, and bound the size of the experimental trial needed to obtain near-optimal decision policies. Second, we demonstrate that adaptive experimental designs provide some key advantages over static experimental designs. Inspired by work in multi-armed bandits (e.g. Auer et al. [2002]), we learn 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 as described above.

To illustrate and evaluate these approaches, we use client data from the Santa Clara County Public Defender Office and run a series of empirically grounded simulations. We show that using adaptive experimental designs with our framework yields better outcomes for participants during learning, and often more quickly identifies higher utility decision policies for future use, compared to static experimental approaches like randomized control trials.

2 Related Work

Our work draws on research in algorithmic fairness, fair division, 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 [Buolamwini and Gebru, 2018, Raji and Buolamwini, 2019, Blodgett and O'Connor, 2017, Caliskan et al., 2017, De-Arteaga et al., 2019, Ali et al., 2019, Datta et al., 2018, Obermeyer et al., 2019, Goodman et al., 2018, Chouldechova et al., 2018, Koenecke et al., 2020, Shroff, 2017], and concomitant development of formal criteria to characterize fairness [Barocas et al., 2017, Chouldechova and Roth, 2018, Corbett-Davies and Goel, 2018, Gupta et al., 2020]. 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 [Kilbertus et al., 2017, Wang et al., 2019, Coston et al., 2020, Kusner et al., 2017, Nabi and Shpitser, 2018, Zhang and Bareinboim, 2018, Chiappa and Isaac, 2018, Wu et al., 2019, Nyarko et al., 2021, Nilforoshan et al., 2022].

All of the above approaches conceptualize the equity of algorithmic decisions in terms of universal rules (e.g., error rate parity) rather than considering the consequences of decisions. Recent work has noted limitations to this deontological approach, which has dominated the fair machine learning literature [Cowgill and Tucker, 2019, 2020, Corbett-Davies et al., 2017, Kasy and Abebe, 2021]. In Section 4, we show that a narrow focus on predictive parity can lead to a suboptimal allocation of limited transportation resources to clients. Some recent exceptions have begun to consider algorithmic decision-making from a consequentialist perspective to varying degrees [Liu et al., 2018, Viviano and Bradic, 2020, Fang et al., 2022, Donahue and Kleinberg, 2020, Coston et al., 2020, Nilforoshan et al., 2022]. For example, Nilforoshan et al. [2022] show that common causal definitions of algorithmic fairness lead to Pareto-dominated policies.

In a related thread of research on fair division problems, groups of individuals decide how to split a limited set of resources amongst themselves [Bertsimas et al., 2011, Gal et al., 2017, Caragiannis et al., 2012, Brams et al., 1996]. The broad aim of that work—to equitably allocate a limited resource—is similar to our own, but it differs in three important respects. First, canonical fair division problems seek to arbitrate between individuals with competing preferences (e.g., as in cake-cutting style problems [Procaccia, 2013]), rather than adopting

the preferences of a social planner, as we do. Second, and relatedly, much of the fair division literature, like the algorithmic fairness literature, takes an axiomatic approach to fairness, identifying allocations that have properties posited to be desirable, such as envy-freeness [Cohler et al., 2011]. Although that perspective is useful in many applications, it does not explicitly consider the preferences of policymakers, which may be incompatible with these axiomatic constraints. Finally, work on fair division problems typically does not try to learn causal effects of allocations on downstream outcomes from data, such as the heterogeneous effect of transportation assistance on appearance rates.

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., Corbett-Davies et al. [2017], Cai et al. [2020], Rolf et al. [2020]); however, there has been little work on creating equitable learning systems that account for competing objectives. Relatedly, a large and growing body of work has shown that one can often efficiently elicit preferences for complex objectives, even in high-dimensional outcome spaces [Lin et al., 2020, Fürnkranz and Hüllermeier, 2010, Chu and Ghahramani, 2005].

One particularly challenging aspect of our setting is handling budget constraints (e.g., we may only be able to provide rideshare assistance to a limited number of clients) [Luedtke and van der Laan, 2016]. 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. Slivkins et al. [2019, Ch. 10] 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 [Badanidiyuru et al., 2014, Agrawal et al., 2016b] 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. They do not consider multi-objective optimization, and their analysis and experiments do not address continuous or large state spaces, which make their work less applicable for equitable decision making in many settings of interest.

3 Decision-Making as Optimization

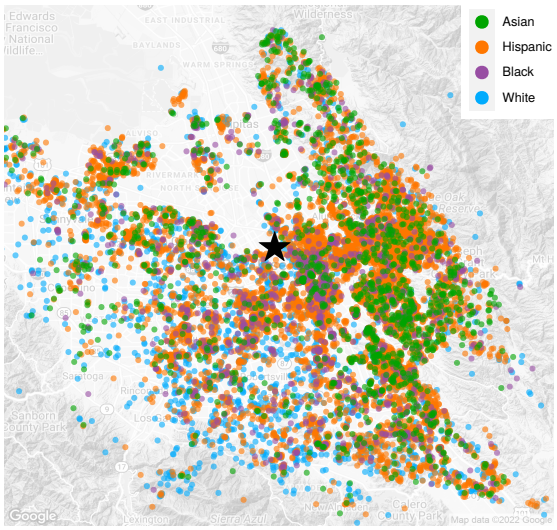
We begin this section by describing our motivating example of providing transportation to individuals with mandatory court dates. Next, we outline a general, utility-based paradigm for equitable decision-making, and by assuming complete knowledge on the distribution of potential outcomes under actions, we present a computationally efficient approach to deriving optimal policies. (Later, in Section 5, we use this result to address the more general problem of learning optimal policies via experimentation.)

3.1 Motivating Example

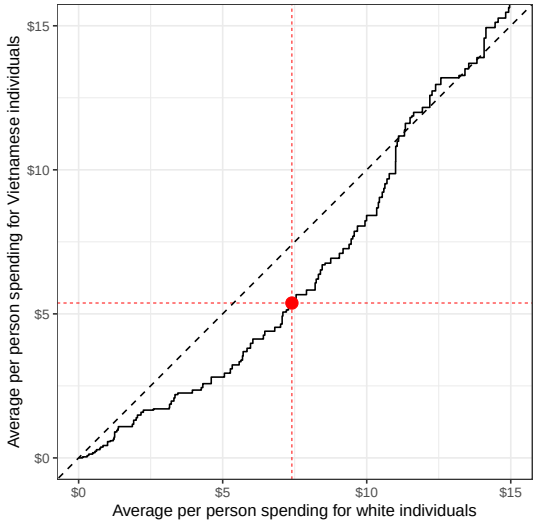
Consider the problem of allocating rideshare assistance to individuals who are required to attend mandatory court dates. The consequences of missing a court date can be severe. Often, after an individual misses a court appearance, judges will issue a "bench warrant", which can lead to the individual's arrest at their next contact with law enforcement, and possibly weeks or months of jail time. Despite these extreme consequences, some individuals struggle to attend court because of significant transportation barriers. For example, in interviews of individuals who have previously missed court, individuals stated that a combination of mobility issues, inadequate public transportation, and their lack of a personal vehicle made court attendance infeasible. Government agencies–including public defender offices—may therefore aim to improve appearance rates by offering transportation assistance to and from court for a subset of these individuals with the greatest transportation needs.[1] This type of intervention has promise for improving appearance rates by alleviating transportation burdens many clients face, as has been demonstrated in medical settings [Chaiyachati et al., 2018b, Vais et al., 2020, Fraade-Blanar et al., 2021, Saxon et al., 2019, Lyft, 2020].

A natural algorithmic approach to allocate rides is to prioritize those with the largest estimated treatment effect per dollar. In particular, suppose we have access to a rich set of covariates, X_i, for each individual i, such as their age, alleged offense, and history of appearance. Based on these covariates, we could then estimate appearance rates for each individual in the absence of assistance, $\hat{Y}_i(0)$, and appearance rates if provided with a ride, $\hat{Y}_i(1)$—for example, by using historical data and/or a randomized controlled trial. Finally, we could sort individuals by $\rho_i = [\hat{Y}_i(1) - \hat{Y}_i(0)]/c_i$, where c_i is the cost of providing a ride to the i-th individual, and offer assistance to those with the highest values of ρ_i until the

[1] As we discuss in Section 6, there are many alternative policy solutions to this issue, including discouraging judicial use of incarceration after an individual misses court.



(a) Santa Clara client locations. Each dot has been randomly perturbed to preserve privacy.



(b) Average per person spending (Vietnamese vs. white) in the absence of parity constraints.

Figure 1: The map in a shows the geographic distribution of the client base of the Santa Clara County Public Defender Office. The star on the map marks the location of the main county courthouse, where most clients are required to appear for court appointments. The plot on the right explores the consequence of following a policy that provides rides to those with the highest estimated treatment effect per dollar without parity constraints. This policy would result in higher average per person spending for white individuals than for Vietnamese individuals. The red point in b shows that, for a fixed budget, an average per person spending amount of $7.40 for white individuals would correspond to an average per person spending amount of $5.38 for Vietnamese individuals.

budget is exhausted.

This strategy aims to achieve the highest appearance rate given the available budget. However, in so doing, it implicitly prioritizes those closest to the courthouse—for whom rides are typically less expensive—which could lead to unintended consequences. For example, consider the Santa Clara County Public Defender Office (SCCPDO) in California, which represents tens of thousands of indigent clients every year. Like many American jurisdictions, Santa Clara County, which includes San Jose, is racially segregated (Figure 1a). In particular, Santa Clara's Vietnamese population, one of the county's largest ethnic minorities, and a focus of the defender's office, tends to live farther away from the courthouse than other racial groups, including white individuals.

To understand the impacts of a strategy that optimizes exclusively for appearance, we start with a dataset of 22,283 cases handled by SCCPDO between between 2017 and 2019. For each individual, we use the historical data to estimate $Y_i(0)$ based on age, gender, offense severity (misdemeanor or felony), appearance history, and the client's distance to the courthouse. For simplicity, we assume $Y_i(1) = 1$, meaning that all individuals who receive a ride attend court. Finally, we assume rides cost \$5 per mile. Under the naive optimization approach outlined above, Figure 1b shows per-capita spending for white and Vietnamese clients across different overall transportation budgets. For example, for an overall budget of \$80,000, a policy that allocates rides to those with the highest estimated treatment effect per dollar would end up spending, on average, \$7.40 for every white client, but only \$5.38 on average per Vietnamese client. Policymakers and other stakeholders may deem this disparity to be undesirable, and may thus be willing to accept lower appearance rates in return for more equal spending across groups, a trade-off that we formalize and address in the following sections.

3.2 Problem Formulation

Consider a sequential decision-making setting like the one described above, 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, history of appearance, alleged charges, and distance from court, and the set of actions might specify whether or not rideshare assistance is offered (in which case, $K = 2$). 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 only a certain amount of money can be spent on average per client, with varying known costs per context and action $c(x, a_k)$. As such, given a cap b for average per-person expenditures, we require our decision policy π to satisfy

$$\mathbb{E}_X[c(X, \pi(X))] = \sum_{x,k} \Pr(X = x) \cdot \Pr(\pi(x) = a_k) \cdot c(x, a_k) \leq b.$$

In many common scenarios, we might imagine a setup where one "control" action a_0 has no cost, i.e., $c(x, a_0) = 0$, while all other available actions are costly (i.e., $c(x, a_k) > 0$ for

$k > 0$). 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 court date if offered rideshare 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 a policymaker's (ex-post) value for the potential decisions (and corresponding outcomes) they could make for each individual. In our motivating application, we might set

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

where $a \in \{0, 1\}$ indicates whether rideshare assistance is provided, $y \in \{0, 1\}$ indicates whether a client appeared at their court date, $\mathbb{I}_{\text{frequent}}(x)$ indicates whether an individual is in frequent contact with law enforcement, and the positive constants c_1 and c_2 characterize the relative values of the terms.[2] This choice of r encodes the (hypothetical) policymaker's belief that: (1) appearing at one's court date is better than not appearing; (2) receiving rideshare assistance is better than not receiving it, regardless of the outcome; and (3) the value of both assistance and appearance is greater for those who frequently encounter law enforcement (i.e., those for whom an open bench warrant is more likely to result in jail time because they are more likely to encounter law enforcement).

Finally, given the above setup, we assume the policymaker's utility $U(\pi)$ of any decision policy π takes the form:

$$\begin{aligned} U(\pi) = \mathbb{E}_{X,Y}[r(X, \pi(X), Y(\pi(X)))] \\ - \sum_{g \in \mathcal{G}} \lambda_g \Big| \mathbb{E}_X[c(X, \pi(X)) \mid g \in s(X)] - \mathbb{E}_X[c(X, \pi(X))] \Big|, \end{aligned} \tag{2}$$

where $\mathbb{E}_X[c(X, \pi(X))]$ denotes the expected expenditure, $|\cdot|$ is an absolute value, λ_g are non-negative constants, and $s(X_i) \subseteq \mathcal{G}$ is a set of associated identities for each individual, where $\mathcal{G}$ is a finite set. In discussions of algorithmic fairness, special attention is often paid to these groups, which may consist of legally protected characteristics. For example, $s(X_i)$ might specify both an individual's race and gender.

The first term in $U(\pi)$ captures the social value directly associated with each decision. The second term captures the social value of spending parity across the population more broadly. For example, in addition to preferring transportation assistance policies that boost appearance rates, a policymaker might also prefer those for which we spend similar amounts

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

per person across neighborhoods, to ensure such investments are broadly applied across an agency's jurisdiction. Depending on the application, one could imagine replacing this term with other such penalties: one may choose to penalize a given policy if the distribution of *actions* or *successes* is unequal across groups. However, for the purposes of simplicity, we focus on solely on spending disparities in this paper.

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

$$\begin{aligned} &\pi^* \in \arg\max_{\pi} \; U(\pi) \\ &\text{subject to: } \mathbb{E}_X[c(X, \pi(X))] \leq b. \end{aligned} \tag{3}$$

We next discuss settings in which it is computationally efficient to derive these optimal policies.

3.3 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 $\mathcal{D}(Y(a_k) \mid X)$. 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. To construct the LP, first observe that any policy π corresponds to a matrix $v \in \mathbb{R}_+^{\mathcal{X}} \times \mathbb{R}_+^K$, 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 $\mathbb{E}_X[c(X, \pi(X))] \leq b$ in Eq. (3) can be expressed as a linear inequality on the decision variables:

$$\sum_{x,k} \Pr(X = x) \cdot v_{x,k} \cdot c(x, a_k) \leq b.$$

Finally, we need to express the utility $U(x)$ in linear form. First, note that:

$$U(\pi) = \sum_{x,k} \mathbb{E}_Y[r(x, a_k, Y(a_k)) \mid X = x] \cdot \Pr(X = x) \cdot v_{x,k}$$
$$- \sum_g \left| \sum_{x,k} \lambda_g \left(\frac{\mathbb{I}(g \in s(x)) \Pr(X = x)}{\Pr(g \in s(X))} \cdot c(x, a_k) - \Pr(X = x) \cdot c(x, a_k) \right) v_{x,k} \right|.$$

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

$$\alpha^T v - \sum_g \lambda_g |\beta_g^T v|, \tag{4}$$

where α and β are constant vectors. We can rewrite this optimization problem as a linear program that includes additional (slack) variables w_g:

$$\begin{aligned} \text{Maximize:} \quad & \alpha^T v - \sum_g \lambda_g w_g \\ \text{Subject to:} \quad & 0 \leq w_g, \\ & -w_g \leq \beta_g^T v \leq w_g. \end{aligned} \tag{5}$$

For completeness, we include 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:

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

Subject to:

$$v_{x,k},\ w_g \geq 0 \quad \forall\, x, k, g,$$
$$\sum_k v_{x,k} = 1 \quad \forall\, x,$$
$$\sum_{x,k} \Pr(X = x) \cdot v_{x,k} \cdot c(x, a_k) \leq b, \text{ and}$$
$$-w_g \leq \sum_{x,k} \left(\frac{\mathbb{I}(g \in s(x)) \Pr(X = x)}{\Pr(g \in s(X))} \cdot c(x, a_k) - \Pr(X = x) \cdot c(x, a_k) \right) v_{x,k} \leq w_g \quad \forall g.$$

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:

$O(|\mathcal{X}|K + |\mathcal{G}|)$ variables and constraints in our case. In practice, using open-source software running on conventional hardware, we find it takes roughly 1–2 seconds 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.[3]

In the common case of $K = 2$ treatments (e.g., with the options corresponding to whether or not one provides rideshare assistance), we show in Appendix B that optimal decision policies have a simple, interpretable form. More broadly, our optimization approach accommodates 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 an absolute value. A squared penalty could be expressed as a quadratic program (QP), but 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 below.

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

Proof. 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., $c(x, a_k) = b = 1$), 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. □

[3] We used the Glop linear optimization solver, as implemented in Google OR-Tools (`https://developers.google.com/optimization/`).

4 Choosing Among Potential 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., appearances in court); on the other hand, one also seeks to minimize spending disparities across groups in a population. To explore this trade-off, we return again to our motivating application of allocating rideshare assistance to public defender clients who are required to appear in court. For simplicity, here we consider a synthetic client population with two equally sized groups that have identical appearance rates in the absence of rideshare assistance, and further assume there is constant unit cost to provide a ride to each individual (i.e., $c(x, a_1) = 1$).[4] However, one group (which we refer to as the target 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.[5] We describe the data-generating process for this synthetic population in detail in Appendix C.

In Figure 2, for budget $b = 1/3$, we trace out the Pareto frontier for this example, which shows how the maximum possible number of appearances (on the vertical axis) varies under different allocations of rideshare assistance to the target population (on the horizontal axis). By Theorem 4, each point on the frontier corresponds to a threshold policy that provides assistance to clients with the largest treatment effects in each group, subject to demographic and budget constraints.

Among feasible options (i.e., points on the Pareto frontier), a policymaker ostensibly has more and less preferred outcomes. We approximate these preferences by assuming utilities follow the functional form in Eq. (2), with the family of utilities indexed by the latent parameter λ.[6] For example, imagine that a given policymaker's utility is maximized at the green point on the Pareto curve. In contrast, the point at the crest of the curve (in blue) achieves the highest number of overall appearances, but is a suboptimal policy because it underspends on the target population, at least according to the preferences of the

[4] With constant unit costs, spending parity also achieves treatment parity, and any budget between 0 and 1 represents both the average amount budgeted per capita and also the proportion of population treated.

[5] Optimizing for parity across protected demographic groups is legally impermissible in some contexts in the U.S., as we discuss more in Section 6.

[6] One might show policymakers a series of comparisons from a specific problem domain, and then select the λ that best captures their stated preferences in this domain. As such, we think of λ as a function of one's preferences over possible outcomes, rather than one's preferences being a function of λ—contrasting our consequentialist approach to a deontological one.

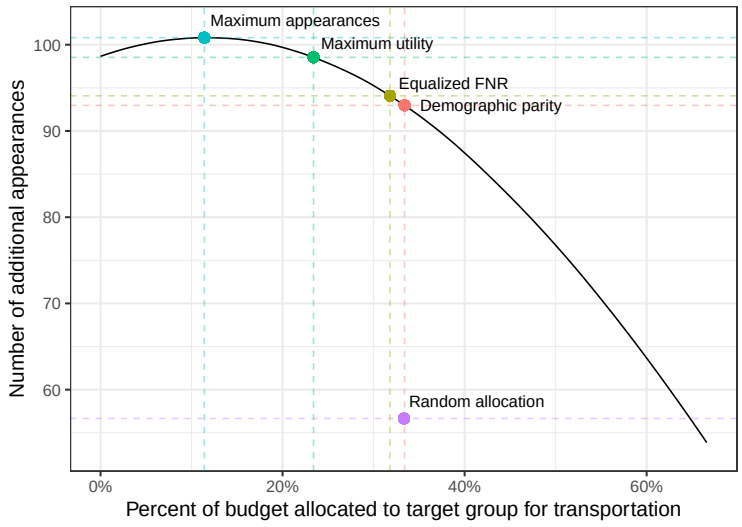


Figure 2: The Pareto frontier for a stylized population model, showing the trade-off between appearances and treatment rates in the target group. The vertical axis shows expected additional appearances relative to a policy that does not provide rideshare assistance to any clients. Under this model, with the policymaker's utility maximized at $\lambda = 0.01$, common heuristics (e.g. maximizing appearances, and demanding demographic or error-rate parity) lead to sub-optimal policies.

policymaker. Similarly, a policy that achieves perfect spending parity (i.e., the pink point) also demonstrates suboptimal outcomes relative to the policymaker's preferences, because too many appearances are lost in order to achieve spending parity.

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.

To extend this example, we also plot points on the curve corresponding to random allocation (in purple) and equal false negative rates (FNR) between groups (in dark yellow).[7] Random allocation results in demographic parity, but lies below the Pareto frontier, meaning appear-

[7]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.

ance rates are lower than if one were to more judiciously allocate rideshare 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 maximizing utility more directly—can result in a sub-optimal balance between maximizing appearances and evenly distributing transportation assistance, relative to the underlying preferences of the policymaker.

5 Learning Optimal Policies

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

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

for all values of x and k. In particular, we assumed perfect knowledge of the conditional distribution of potential outcomes, $\mathcal{D}(Y(a_k) \mid X)$. In reality, however, $f(x, k)$ must be learned from observed data. One common approach for estimating the impact of interventions is to run a randomized controlled trial (RCT) to estimate the effects of actions. In Section 5.1, we formally analyze RCT data collection strategies, and provide an upper bound on the number of samples necessary to ensure we can compute a near-optimal allocation strategy for our desired objective. In Section 5.2, we then present an alternative, contextual-bandit-based strategy that can often learn optimal policies more efficiently than an RCT by judiciously exploring the effects of actions. Finally, in Section 5.3, we demonstrate the advantages of this alternative strategy in an empirically grounded simulation study.

5.1 Sample-Size Bounds for Learning From RCTs

A natural concern for practitioners is whether balancing complicated objectives—like the competing outcomes highlighted in our utility function in Eq. (2)—requires obtaining substantially more data than in traditional, single-objective settings. Further, in most domains of practical interest, individuals are described by a set of features, and it is beneficial to know how choices about representing these individuals impact the amount of data required. To address these considerations, we provide upper bounds on the sample size of an RCT needed to construct near-optimal policies with high probability.[8] Our aim in this analysis

[8]By an RCT, we mean any data collection strategy in which the distribution of actions may depend on the context but which does not change over time. Our analysis can equivalently be viewed as a simple regret analysis under the constraint of using an RCT for pure exploration.

is not to provide tight sample complexity bounds, but rather to examine at a high level how additional fairness objectives and modeling choices affect the amount of data required from an RCT. Our results suggest that one may not need much more data to learn a fair policy—when compared with a policy that solely maximizes reward—and that modeling assumptions can substantially reduce the amount of data required.

As in Sections 3.2 and 3.3, we assume throughout this section that the state space $\mathcal{X}$ is finite, and that the costs and the distribution of X are known. In practice, information on the distribution of X can be estimated from historical data, before any interventions are attempted. Let π^* be an optimal policy solution, as defined in Eq. (3), with corresponding utility $U(\pi^*)$. We define the estimated utility function $\hat{U}(\pi)$ for a particular decision policy as

$$\begin{aligned}\hat{U}(\pi) = \mathbb{E}_{X,Y}[\hat{r}(X, \pi(X), Y(\pi(X)))] \\ - \sum_{g \in \mathcal{G}} \lambda_g \Big| \mathbb{E}_X[c(X, \pi(X)) \mid g \in s(X)] - \mathbb{E}_X[c(X, \pi(X))] \Big|,\end{aligned} \tag{6}$$

where $\hat{r}$ is the estimated reward function learned from data. Let $\hat{\pi}$ be a solution to the optimization problem in Eq. (3), where we maximize $\hat{U}(\pi)$ instead of $U(\pi)$. Further, let $r(x, k) = r(x, a_k, Y_{X=x}(a_k))$ be the (random) reward if action a_k is taken in the context x, where $Y_{X=x}(a_k)$ is the (random) potential outcome conditional on the given context. Note that the randomness in $r(x, k)$ stems entirely from the randomness in the potential outcomes $Y_{X=x}(a_k)$.

We present upper bounds on the sample size needed to learn near-optimal policies. Specifically, for fixed $\epsilon, \delta > 0$, we provide sample bounds which ensure the utility gap $U(\pi^*) - U(\hat{\pi})$ is small with high probability, i.e., $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$. We prove these bounds under three different common distributional assumptions on the reward model, described below.

1. (Tabular Rewards) We assume $r(x, k) \overset{\text{d}}{=} f(x, k) + \eta$, where $\eta \sim \sigma^2$-subGaussian and η is independent across draws of the reward function.

2. (Linear Rewards) We assume there are (known) features $\phi(x, a_k) \in \mathbb{R}^d$ of the state and action, and (unknown) parameters $\theta^* \in \mathbb{R}^d$ such that $r(x, k) \overset{\text{d}}{=} \phi(x, a_k)^T \theta^* + \eta$, where $\eta \sim \sigma^2$-subGaussian and η is independent across draws of the reward function.

3. (Logistic Rewards) We assume there are (known) features $\phi(x, a_k) \in \mathbb{R}^d$ of the state and action, and (unknown) parameters $\theta^* \in \mathbb{R}^d$ such that $\mathbb{P}(r(x, k) = 1) = \text{logit}^{-1}(\phi(x, a_k)^T \theta^*)$, where the reward is independent across draws.

Before formally stating our results, we summarize our key findings. First, in the tabular setting, we show that approximately $\sigma^2|A|/(\epsilon^2 p_{\min})$ samples are sufficient to ensure the utility gap is small with high probability, where $p_{\min} = \min_x \mathbb{P}(X = x)$ (NB: the constant δ appears in lower-order log terms). Our bound thus scales roughly as $|\mathcal{X}||\mathcal{A}|$, the product of the size of the state space and the size of the action space.

In the tabular setting, in the absence of shared structure, we must separately learn the effects of each action in each state. However, if we can use a parametric representation of the reward function, our sample bounds are substantially smaller. In particular, in the linear case we show that approximately $\sigma^2 d^2/\epsilon^2$ samples are sufficient to ensure the utility gap is small with high probability, where d is the dimension of the feature space—a significant improvement over the tabular setting when $d^2 \ll |\mathcal{X}||\mathcal{A}|$. To achieve this bound, we design the RCT to strategically select actions in each context in a way that efficiently samples the feature space. Finally, in the logistic case, we similarly establish bounds that scale as a function of d, though the exact dependence is more complex than in the linear case. Our analysis accordingly suggests that when the context and action spaces are large, estimating shared reward structures is likely to be crucial.

For all three of these settings (tabular, linear, and logistic), our sample bounds are identical whether or not we consider fairness (i.e., regardless of whether $\lambda_g > 0$ for some g or $\lambda_g = 0$ for all g in Eq. (2)). Intuitively, this is the case because the sample complexity is driven by uncertainty in the rewards, which stems from uncertainty in the potential outcomes. The fairness expression itself depends only on the allocation across subgroups, which can be computed exactly given any policy, independent of the estimated rewards. In theory, it is of course possible that tighter bounds would reveal a gap between the sample complexity of the settings with and without fairness constraints.

We now present our formal results. Proofs for this section are in Appendix D.

Theorem 1 (Tabular Rewards)**.** *Assume the reward is tabular. Suppose we collect n samples in a round-robin fashion (i.e., for each context x, select the least-sampled action a_k in that context, breaking ties arbitrarily). Then for $\epsilon > 0$, $\delta > 0$, $\lambda_g \geq 0$, and*

$$n \geq \frac{8\sigma^2|A|}{\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta} \log\left(\frac{16\sigma^2|A|}{\delta\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta}\right),$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$.

As discussed above, our sample bound in the tabular setting roughly scales linearly with the product of the size of the covariate space and the action space, which suggests that

prohibitively large sample sizes may be needed in practice. Our next two theorems show that significantly fewer samples are sufficient if the reward function is a parametric model.

Theorem 2 (Linear Rewards)**.** *Assume the reward is linear with feature representation* $\phi(x, a_k) \in \mathbb{R}^d$*. For any RCT* π *used to collect samples, let*

$$\begin{aligned}\Sigma(\pi) &= \mathbb{E}[\phi(X, \pi(X))\phi(X, \pi(X))^T] \\ &= \sum_{x,k} \mathbb{P}(X = x) \cdot \mathbb{P}(\pi(x) = a_k) \cdot \phi(x, a_k)\phi(x, a_k)^T\end{aligned}$$

be the induced covariance matrix. Also define a problem-dependent constant

$$\rho_0(\pi) = \max_{x,k} \|\Sigma(\pi)^{-1/2}\phi(x, a_k)\|/\sqrt{d}.$$

Then, we can design a data collection strategy $\tilde{\pi}$ *such that, for any* $\epsilon > 0, \delta > 0, \lambda_g \geq 0$ *and*

$$n \geq \max\{6\rho_0(\tilde{\pi})^2 d \log(3d/\delta), O\big(\sigma^2 d^2/\epsilon^2\big)\},$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$*.*

The quantity ρ_0 in the above bound is known as 'statistical leverage' [Hsu et al., 2014]. If no prior information is available, we know only that $\rho_0 \leq \|\phi\|_2/\sqrt{\lambda_{\min}(\Sigma)}$, and, in the worst case, ρ_0 may scale with the condition number of the covariance matrix. However, in many practical settings ρ_0 is not large compared to $1/\epsilon^2$, and so the upper bound scales like $\sigma^2 d^2/\epsilon^2$.

Finally, given the practical importance of binary rewards, we provide the following upper bound on sample complexity in the logistic setting.

Theorem 3 (Logistic Rewards)**.** *Assume the reward is logistic, and that assumptions D0, D1, D2, and C of Ostrovskii and Bach [2020] hold (these assumptions define problem-dependent constants* K_0, K_1, K_2, ρ*). Define* $\Sigma(\pi)$ *as in Theorem 2 and* $c = \sum_x \mathbb{P}(X = x) \max_k \|\phi(x, a_k)\|_{\Sigma(\pi)^{-1}}$*. Then, for any* $\epsilon > 0, \delta > 0, \lambda_g \geq 0$ *and*

$$n \geq O\bigg(\max\{K_2^4(d + \log\frac{1}{\delta}), \rho K_0^2 K_1^2 d^2 \log\frac{d}{\delta}, (\rho^2 c^2 K_1^2 d \log\frac{1}{\delta})/\epsilon^2\}\bigg)$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$*.*

Theorem 3 provides guarantees on the performance of the allocation strategy derived from using estimated plug-in parameters for the logistic reward model. However, the assumptions we use to establish this result are quite strong, suggesting there is significant room for similar results under more relaxed conditions. We discuss the implications of Theorem 3 and the strength of the assumptions further in Appendix D.

5.2 Adaptively Learning Fair, Optimal Policies

The prior section suggests the feasibility of solving our desired optimization problem using parameters estimated from data. However, sample size calculations for RCTs are most suitable for settings with a fixed budget for experimentation and a strong need for testing statistical hypotheses posthoc which are most easily done with independently and identically distributed data[9]. However, such estimates can be overly conservative and involve deploying non-reactive data gathering poicies. For example, in our running example of providing rideshare assistance to public defender clients, if there turns out to be a group of clients with very small need and benefit from assistance, the RCT will still allocate a proportional amount of limited resources to such individuals. 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_k 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 3.3 with three common contextual bandit approaches: ε-greedy, Thompson sampling, and upper confidence bound (UCB), as described in Algorithm 1.[10] At a high level, at each step i, our ε-greedy approach first estimates $f(x, k)$ 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 a probability set to meet our budget requirements in expectation. Our Thompson sampling approach maintains a posterior over the parameters of a model of the potential outcomes $\hat{f}(x, k)$, 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^* by solving the LP with an optimistic estimate of $f(x, k)$ (e.g., using the 97.5th percentile of the posterior of $\hat{f}(x, k)$).

Algorithm 1 Policy learning procedure.

input: Actions a_k, budget b, parity preferences λ_g, reward function r, covariate distribution $\mathbb{P}(X = x)$, 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_i^* according to Eq. (E.27)
    Find solution π_i^* of the LP in Section 3.3 with input values $\hat{f}(x, a)$, s, λ_g, b_i^*, and $\mathbb{P}(X = x)$
    if ε-greedy & BERNOULLI(ε) == 1 **then**
        Take random action A_i according to Eq. (E.28)
    else
        Take action $A_i \sim \pi_i^*(X_i)$
    end if
    Observe outcome Y_i
end for

5.3 Simulation Study

To evaluate our learning approach above, we conducted a simulation study using data on a sample of clients served by the Santa Clara County Public Defender Office. In this example, clients can receive one of three mutually exclusive treatments a_k: rideshare assistance, a

[9] Note that there has been recent interest in developing suitable inference methods for data gathered using adaptive, multi-armed bandit strategies (e.g. Hadad et al. [2021], Zhang et al. [2021])

[10] For simplicity, we assume knowledge of the covariate distribution $\mathcal{D}(X)$, 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.

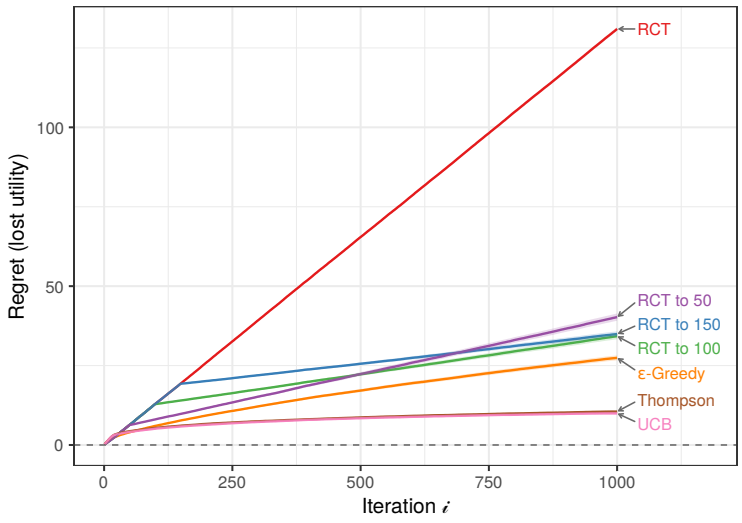


Figure 3: Mean regret, across 2,000 simulations, incurred by different learning approaches. We define regret here as the difference between the observed utility and the utility obtained by an oracle during the same experiment. Uncertainty bands represent 95% intervals for the mean. We note that the three 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 4).

transit voucher, or no transportation assistance. We fix our average per-person budget to \$5 for this simulation, and assume that round-trip rides cost \$5 for every mile between an individual's home address and the main courthouse and back. We also limit the client population to white and Vietnamese individuals to reflect the motivating example described in Section 3.1. The utility of a policy is described by Eq. (2), where we set $r(x, a, y) = y$ and $\lambda_g = 0.004$. This choice yields an oracle policy that balances between maximizing appearances and achieving parity in per-capita expenditure across groups. 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

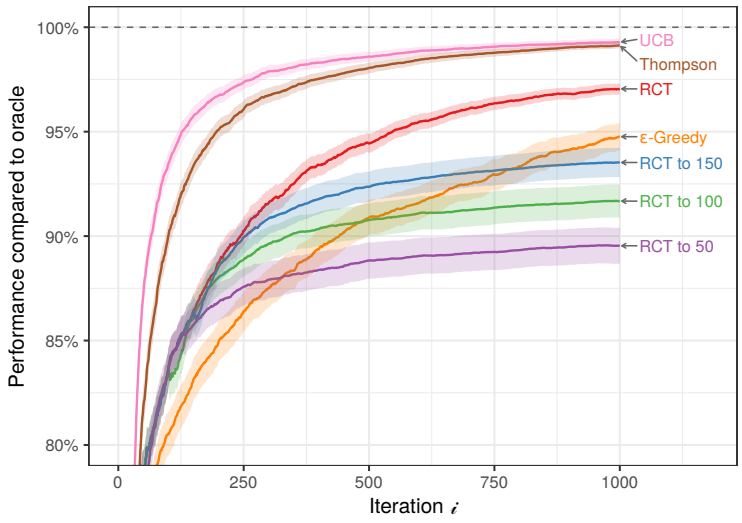


Figure 4: Mean performance, across 2,000 simulations, of optimal policies estimated with data available at each iteration i. Performance is defined as the additional utility obtained by a policy over a baseline of no treatment for all individuals, with 100% indicating this quantity for the oracle policy. Uncertainty bands represent 95% intervals around the mean. UCB and Thompson sampling generate policies that are better than a conventional RCT at any given iteration i. In contrast, the ε-greedy approach and the stop-early versions of the RCT generate policies that are slower to (or may never) reach near-oracle performance.

run an RCT on the first n individuals, and then follow the optimal policy estimated at individual n for the rest of the sample, similar to explore-first strategies. We compare all approaches against an oracle that can observe the true expected appearance probabilities.

We repeat this evaluation 2,000 times each on 1,000 randomly-selected individuals from our dataset, 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 3), but also learn policies that, if used for future populations, would outperform all other approaches (Figure 4). 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,

	Vietnamese spending disparity	
Method	With penalty ($\lambda_g = 0.004$)	No penalty ($\lambda_g = 0$)
UCB	-$2.21	-$3.36
Thompson	-$1.21	-$2.19
ϵ-Greedy	-$1.29	-$2.38

Table 1: Mean spending disparities by method for Vietnamese clients across 2,000 experiments, including both the main set of simulations (where $\lambda_g = 0.004$) and an alternative set of simulations (with identical parameters to the main set, except where $\lambda_g = 0$). Disparities are calculated by comparing average spending on Vietnamese individuals to the $5 average spending on all individuals (i.e., the target budget). Note that spending disparities are approximately $1 larger when $\lambda_g = 0$, verifying that the bandit methods we employ in our simulation learn to reduce spending disparities to maximize the policymaker's utility.

these alternatives learn poorer-performing policies.

By design, the bandit methods discussed above reduce spending disparities during the course of the simulation. We demonstrate this by comparing our main simulation to an alternate set of simulations where $\lambda_g = 0$ (Table 1). For example, with a choice of $\lambda_g = 0.004$, reflecting a mild preference for more equal spending, we observe that UCB methods spent $2.21 less on Vietnamese clients than the $5 population average (i.e., the target budget). In contrast, with a choice of $\lambda_g = 0$ (i.e., preferring policies that simply aim to maximize appearances), UCB methods spent $3.36 less on Vietnamese clients compared to the population average.

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 instance, in our running example of allocating transportation assistance to public defender clients, 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 trade-offs, and, in the process, may help 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 an empirically grounded 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. For example, challenges may arise from an unwillingness to explicitly state preferences for trade-offs involving sensitive considerations like demographic parity. 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 [Lin et al., 2020, Fürnkranz and Hüllermeier, 2010, Chu and Ghahramani, 2005].

Another challenge—particularly relevant in the dynamic setting—is accounting for delayed outcomes. In our running example, we may choose to offer rideshare assistance to a client 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 rideshare assistance to clients, one might use intermediate responses (like a client'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.

In addition to the above technical considerations, we note some practical limitations in providing transportation to public defender clients with upcoming court dates. First, in many circumstances policymakers may not be legally permitted to explicitly use race, ethnicity, or other protected attributes when deciding how to allocate limited resources. These

policymakers may instead focus on other attributes, like geography or socioeconomic status, which may be legally or socially more permissible. Second, our motivating example presupposes that resources are too limited to treat the entire population of interest. If policymakers had enough funding available to assist an entire population, it may not make sense to equalize per-capita spending across groups of interest, given that everyone would receive transportation assistance. Finally, though this study emphasizes the potential benefits of rideshare assistance for those who have mandatory court dates (e.g., one potential benefit is avoiding time in jail), a simpler and more effective policy for reducing jail time may be to discourage judges from issuing bench warrants if clients fail to appear in court. Though in isolation this policy might result in lower appearance rates, it could be accompanied by other assistance to offset this adverse outcome, including text message reminders, social services, or rideshare assistance as we describe here.

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.

7 Acknowledgements

We thank Johann Gaebler, Jonathan Lee, Hamed Nilforoshan, Julian Nyarko, and Ariel Procaccia for helpful comments. We also thank colleagues at the Santa Clara County Public Defender Office for their assistance, including Molly O'Neal, Sarah McCarthy, Terrence Charles, and Sven Bouapha. 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

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

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual

bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18. PMLR, 2016b.

Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095*, 2019.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 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.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS tutorial*, 1:2, 2017.

Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011. ISSN 0030364X, 15265463. URL `http://www.jstor.org/stable/23013103`.

Su Lin Blodgett and Brendan O'Connor. 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.

Steven J Brams, Steven John Brams, and Alan D Taylor. *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press, 1996.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 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*, pages 1–12, 2019.

Emma Brunskill and Lihong Li. The online discovery problem and its application to lifelong reinforcement learning. *CoRR*, abs/1506.03379, 2015. URL `http://arxiv.org/abs/1506.03379`.

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

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

William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 22–28, 2020.

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

Ioannis Caragiannis, Christos Kaklamanis, Panagiotis Kanellopoulos, and Maria Kyropoulou. The efficiency of fair division. *Theory of Computing Systems*, 50(4):589–610, 2012.

Krisda H Chaiyachati, Rebecca A Hubbard, Alyssa Yeager, Brian Mugo, Stephanie Lopez, Elizabeth Asch, Catherine Shi, Judy A Shea, Roy Rosin, and David Grande. Association of rideshare-based transportation services and missed primary care appointments: a clinical trial. *JAMA internal medicine*, 178(3):383–389, 2018a.

Krisda H Chaiyachati, Rebecca A Hubbard, Alyssa Yeager, Brian Mugo, Judy A Shea, Roy Rosin, and David Grande. 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.

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

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

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

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

Yuga J Cohler, John K Lai, David C Parkes, and Ariel D Procaccia. Optimal envy-free cake cutting. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

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

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593. Association for Computing Machinery, 2020.

Bo Cowgill and Catherine E Tucker. Economics, fairness and algorithmic bias. *In preparation for: Journal of Economic Perspectives*, 2019.

Bo Cowgill and Catherine E Tucker. Algorithmic fairness and economics. *Columbia Business School Research Paper*, 2020. doi: 10.2139/ssrn.3361280.

Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency*, pages 20–34, 2018.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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*, pages 120–128. ACM, 2019.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 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*, pages 1–12, 2020.

Kate Donahue and Jon Kleinberg. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 658–668, 2020.

Ethan X Fang, Zhaoran Wang, and Lan Wang. Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, pages 1–14, 2022.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Laura Fraade-Blanar, Tina Koo, and Christopher M. Whaley. Going to the doctor: Rideshare as nonemergency medical transportation. Technical report, RAND Corporation, Santa Monica, CA, 2021.

John J Friedewald, Ciara J Samana, Bertram L Kasiske, Ajay K Israni, Darren Stewart, Wida Cherikh, and Richard N Formica. The kidney allocation system. *Surgical Clinics*, 93(6):1395–1406, 2013.

Johannes Fürnkranz and Eyke Hüllermeier. Preference learning and ranking by pairwise comparison. In *Preference learning*, pages 65–82. Springer, 2010.

Ya'akov Gal, Moshe Mash, Ariel D Procaccia, and Yair Zick. Which is the fairest (rent division) of them all? *Journal of the ACM (JACM)*, 64(6):1–22, 2017.

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

Sharad Goel, Ravi Shroff, Jennifer L Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. *Equity, and Jurisprudence of Criminal Risk Assessment (December 26, 2018)*, 2018.

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

Swati Gupta, Akhil Jalan, Gireeja Ranade, Helen Yang, and Simon Zhuang. Too many fairness metrics: Is there a solution? *Available at SSRN 3554829*, 2020. doi: 10.2139/ssrn.3554829.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression, 2014.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445919. URL https://doi.org/10.1145/3442188.3445919.

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

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689, 2020.

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

Edward J Latessa, Richard Lemke, Matthew Makarios, and Paula Smith. The creation and validation of the ohio risk assessment system (oras). *Fed. Probation*, 74:16, 2010.

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

Zhiyuan (Jerry) Lin, Adam Obeng, and Eytan Bakshy. Preference learning for real-world multi-objective decision making. In *International Conference on Machine Learning*, 2020.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Alexander R Luedtke and Mark J van der Laan. Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303, 2016.

Lyft. Modernizing medical transportation with rideshare. Technical report, FierceHealthcare, 2020.

Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip Thomas. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

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

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

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 16848–16887. PMLR, 17–23 Jul 2022.

Julian Nyarko, Sharad Goel, and Roseanna Sommers. Breaking taboos in fair machine learning: An experimental study. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance, 2020.

Ariel D Procaccia. Cake cutting: Not just child's play. *Communications of the ACM*, 56 (7):78–87, 2013.

Inioluwa Deborah Raji and Joy Buolamwini. 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.

Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Bjorkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, pages 8158–8168. PMLR, 2020.

Leslie Saxon, Rebecca Ebert, and Mona Sobhani. Health impacts of unlimited access to networked transportation in older adults. *Journal of mHealth*, 2019.

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

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

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

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Simone Vais, Justin Siu, Sheela Maru, Jodi Abbott, Ingrid St Hill, Confidence Achilike, Wan-Ju Wu, Tejumola M Adegoke, and Courtney Steer-Massaro. Rides for refugees: A transportation assistance pilot for women's health. *Journal of immigrant and minority health*, 22(1):74–81, 2020.

Lieven Vandenberghe, Stephen Boyd, and Shao-Po Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–35, 04 1998. URL `https://www.proquest.com/scholarly-journals/determinant-maximization-with-linear-matrix/docview/923756606/se-2?accountid=14026`. Copyright - Copyright] © 1998 Society for Industrial and Applied Mathematics; Last updated - 2021-09-11.

Davide Viviano and Jelena Bradic. Fair policy targeting. *arXiv preprint arXiv:2005.12395*, 2020.

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

Bryan Wilder, Laura Onasch-Vera, Graham Diguiseppi, Robin Petering, Chyna Hill, Amulya Yadav, Eric Rice, and Milind Tambe. Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14948–14956, May 2021. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17754`.

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

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. *arXiv preprint arXiv:1910.12586*, 2019.

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

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Statistical inference with m-estimators on bandit data. *Neural Information Processing Systems*, 2021.

Marcela Zuluaga, Guillaume Sergent, Andreas Krause, and Markus Püschel. Active learning for multi-objective optimization. In *International Conference on Machine Learning*, pages 462–470. PMLR, 2013.

Appendices

A Absolute Value in an LP Objective

If (v^*, w^*) is a solution to the LP in Eq. (5), then we claim v^* 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 $w_g = |\beta_g^T v|$ satisfies the LP constraints, $\text{OPT}_{\text{abs}} \leq \text{OPT}_{\text{LP}}$.

Conversely, because the LP objective function decreases in w_g, if $\beta_g^T v^* \geq 0$, then $w_g^* = \beta_g^T v^*$ (since $\beta_g^T v^* \leq w_g$, and the other two constraints are immediately satisfied in this case). On the other hand, if $\beta_g^T v^* \leq 0$, then $w_g^* = -\beta_g^T v^*$ (since $\beta_g^T v^* \geq -w_g$). Thus, in either case, $w_g^* = |\beta_g^T v^*|$, which implies that $\text{OPT}_{\text{abs}} = \text{OPT}_{\text{LP}} = \alpha^T v^* - \lambda_g \sum_g |\beta_g^T v^*|$.

B Group-Specific Threshold Rules

The LP described in Section 3.3 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 rideshare 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 group-specific threshold rules.

Theorem 4. *In the setting of Section 3.2, suppose $K = 2$, $c(x, a_0) = 0$, $c(x, a_1) > 0$, and $|s(x)| = 1$ (i.e., $\mathcal{G}$ partitions $\mathcal{X}$). Further assume $\Delta(x) > 0$, where*

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

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

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

X	$\Pr(X=x)$	$\mathbb{E}[Y(0)\|X=x]$	$\mathbb{E}[Y(1)\|X=x]$	$c(x,a_1)$	$\Delta(x)/c(x,a_1)$	$\mathbb{E}[Y(2)\|X=x]$	$c(x,a_2)$	$\Delta(x)/c(x,a_2)$
x_1	0.1	0.1	0.6	\$10	0.05	0.3	\$1	0.2
x_2	0.9	0.1	0.2	\$10	0.01	0.12	\$1	0.02

Table B.1: Setup for counterexample.

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

$$\begin{aligned} U(\pi) = &\sum_x \mathbb{E}_Y[r(x, a_0, Y(a_0)) \mid X = x] \cdot \Pr(X = x) \\ &+ \sum_x \Delta(x) \cdot \Pr(\pi(x) = a_1) \cdot \Pr(X = x) \\ &- \sum_{g \in \mathcal{G}} \lambda_g \delta_g(\pi), \end{aligned}$$

where

$$\delta_g(\pi) = |\mathbb{E}_X[c(X, \pi(X)) \mid g \in s(X)] - \mathbb{E}_X[c(X, \pi(X))]|.$$

Now, for any policy π, we construct a threshold policy $\tilde{\pi}$ of the form in Eq. (B.1) by assigning to action a_1 those x in each group g having the largest values of $\Delta(x)/c(x, a_1)$ such that

$$\mathbb{E}_X[c(X, \tilde{\pi}(X) \mid g \in s(X)] = \mathbb{E}_X[c(X, \pi(X)) \mid g \in s(X)].$$

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

$$\sum_x \Delta(x) \Pr(\tilde{\pi}(x) = a_1) \Pr(X = x) \geq \sum_x \Delta(x) \Pr(\pi(x) = a_1) \Pr(X = x).$$

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

In the theorem above, we assume $K = 2$, which yields a simple threshold solution for the optimal policy. In general, with $K > 2$, the structure of the optimal policy can be more complicated. As a counterexample to the above, suppose $K = 3$, with a no-cost baseline action a_0, and two costly actions, a_1 and a_2. We further imagine a population with a single group (i.e., $|\mathcal{G}| = 1$) with individuals in two contexts characterized by attributes x_1 and x_2, and a utility $U(\pi) = \mathbb{E}_{X,Y}[Y(\pi(X))]$. Table B.1 lists the responsiveness of each type of individual to each of the three possible actions, the costs of the two costly actions, and the relative benefit per dollar of the two costly actions over the free baseline action. In this setup, we set $b = \$1$, i.e., we can spend, on average, one dollar per person.

For both types of individuals in this example, action a_2 has the highest relative benefit per dollar over the baseline action a_0 (highlighted in Table B.1 in gray). As such, one intuitive

strategy π is to treat every individual with a_2, exhausting our budget, yielding

$$U(\pi) = (0.3 \cdot 0.1) + (0.12 \cdot 0.9) = 0.138.$$

However, in this case, a better strategy π^* is to assign action a_1 to all individuals of type x_1, and to assign a_0 to all individuals of x_2, exhausting our budget and yielding

$$U(\pi) = (0.6 \cdot 0.1) + (0.1 \cdot 0.9) = 0.15.$$

In this setup, even though a_2 has the highest relative benefit *per dollar*, it is low cost in absolute terms, meaning that we have to treat individuals of both types to exhaust our budget, including individuals of type x_2 who see little benefit to the treatment. As a result, it is better to exhaust one's budget on action a_1 with individuals of type x_1, which sees a higher return on average than treating the entire population with a_2. This type of scenario demonstrates the need for more complicated solutions to identifying an optimal policy, justifying the use of a linear program (or other similar approaches).

C Simulation Details for Section 4

We consider a client 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 rideshare 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.

D Proofs for Section 5.1: Sample Bounds for RCTs

In this section we use the shorthand notation $p_x = \mathbb{P}(X = x)$, $\pi_{xk} = \mathbb{P}(\pi(x) = a_k)$, and $r_{xk} = f(x, k)$.

First we present a lemma which allows us to bound the utility error by the reward estimation errors which we will use for the proofs of the theorems.

Lemma 1. *We can bound the utility error by*

$$U(\pi^*) - U(\hat{\pi}) \leq 2 \sum_x p_x \max_k |r_{xk} - \hat{r}_{xk}| \leq 2 \max_{xk} |r_{xk} - \hat{r}_{xk}|.$$

Proof. Proof. Since $\hat{\pi}$ maximizes the estimated utility $\hat{U}(\pi)$ subject to relevant constraints it follows that $\hat{U}(\hat{\pi}) \geq \hat{U}(\pi^*)$ (since $\hat{\pi}, \pi^*$ by definition both satisfy the constraints). Thus

$$U(\pi^*) - U(\hat{\pi}) = U(\pi^*) - \hat{U}(\hat{\pi}) + \hat{U}(\hat{\pi}) - U(\hat{\pi}) \leq U(\pi^*) - \hat{U}(\pi^*) + \hat{U}(\hat{\pi}) - U(\hat{\pi}). \quad \text{(D.2)}$$

Since the fairness part of the utility function depends only on the policy and not the rewards it cancels out in Equation D.2 leaving

$$U(\pi^*) - \hat{U}(\pi^*) + \hat{U}(\hat{\pi}) - U(\hat{\pi}) = \sum_x p_x \sum_k \pi^*_{xk}(r_{xk} - \hat{r}_{xk}) + \sum_x p_x \sum_k \hat{\pi}_{xk}(\hat{r}_{xk} - r_{xk}) \quad \text{(D.3)}$$

$$\leq 2 \sum_x p_x \max_k |r_{xk} - \hat{r}_{xk}| \quad \text{(D.4)}$$

$$\leq 2 \max_{xk} |r_{xk} - \hat{r}_{xk}|. \quad \text{(D.5)}$$

□

Now we leverage the following lemma to prove our sample bounds for RCT's under our three reward settings.

Theorem 5 (Restatement of Theorem 1)**.** *Assume the reward is tabular and the costs are known. Let $p_{min} = \min_x \mathbb{P}(X = x)$. Suppose we collect n samples in a round-robin fashion (i.e., for each context x, select the least-sampled action a_k in that context, breaking ties arbitrarily). Then for $\epsilon > 0$, $\delta > 0$, $\lambda_g \geq 0$, and*

$$n \geq \frac{8\sigma^2|A|}{\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta} \log\left(\frac{16\sigma^2|A|}{\delta\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta}\right),$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$.

Proof. Proof. Using a sample mean estimator of reward we have that

$$\hat{r}_{xk} = \frac{1}{n_{xk}} \sum_{i=1}^{n_{xk}} R_{xk,i} \sim \text{subGaussian}(\frac{\sigma^2}{n_{xk}^2}). \tag{D.6}$$

Using Hoeffding's concentration inequality and a union bound we get that if $n_{xk} \geq \frac{8\sigma^2}{\epsilon^2} \log \frac{4|X||A|}{\delta}$, $\forall x, k$ then $\mathbb{P}(\max_{xk} |r_{xk} - \hat{r}_{xk}| < \frac{\epsilon}{2}) > 1 - \delta/2$.

Now by Proposition 1 of Brunskill and Li [2015], if $n \geq \log(2/(p_{\min}\delta))/p_{\min}$ then with probability at least $1-\delta/2$ we will observe each context at least once. Thus if we choose our actions in a round-robin fashion for each context then after $n \geq |A| \log(2|A|/(p_{\min}\delta))/p_{\min}$ then with probability at least $1-\delta/2$ we will observe each context-action pair at least once. Thus, repeating this n_{xk} times, by a union bound, if we observe at least

$$n \geq \frac{|A|n_{xk}}{p_{\min}} \log \frac{2|A|n_{xk}}{\delta p_{\min}} = \frac{8\sigma^2|A|}{\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta} \log\left(\frac{16\sigma^2|A|}{\delta\epsilon^2 p_{\min}} \log \frac{4|X||A|}{\delta}\right) \tag{D.7}$$

samples then with probability at least $1 - \delta/2$ we will observe at least n_{xk} samples for each context-action pair.

Finally, by Lemma 1 we see that if $\max_{xk} |r_{xk} - \hat{r}_{xk}| < \frac{\epsilon}{2}$ then $U(\pi^*) - U(\hat{\pi}) < \epsilon$. Thus if Equation D.7 holds then

$$\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > \mathbb{P}(\max_{xk} |r_{xk} - \hat{r}_{xk}| < \frac{\epsilon}{2}) > 1 - \delta. \tag{D.8}$$

□

Now we prove Theorem 2. First, in the following theorem, we prove a result for sample complexity under an arbitrary RCT data collection strategy π. In the following lemma, we then show that we can design this RCT π to achieve a bound that scales roughly like d^2/ϵ^2.

Theorem 6 (Restatement of Theorem 2)**.** *Assume the reward is linear. For any RCT* π *used to collect samples, let*

$$\begin{aligned}\Sigma(\pi) &= \mathbb{E}[\phi(X, \pi(X))\phi(X, \pi(X))^T] \\ &= \sum_{x,k} \mathbb{P}(X = x) \cdot \mathbb{P}(\pi(x) = a_k) \cdot \phi(x, a_k)\phi(x, a_k)^T\end{aligned}$$

be the induced covariance matrix. Define a problem-dependent constant

$$\rho_0 = \max_{x,k} \|\Sigma(\pi)^{-1/2}\phi(x, a_k)\|/\sqrt{d}.$$

Then, we can design a data collection strategy such that, for any $\epsilon > 0, \delta > 0, \lambda_g \geq 0$ *and*

$$n \geq \max\{6\rho_0^2 d \log(3d/\delta), O\big(\sigma^2 d^2/\epsilon^2\big)\},$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$.

Proof. Proof. Let $\hat{\theta}$ be the linear regression estimator and $\hat{r}_{xk} = \langle \phi(x,k), \hat{\theta} \rangle$. Then by Theorem 1 of Hsu et al. [2014] we have that for $n \geq 6\rho_0^2 d \log \frac{3d}{\delta}$,

$$\|\hat{\theta} - \theta^*\|^2_{\Sigma(\pi)} \leq \frac{\sigma^2(d + 2\sqrt{d \log \frac{3}{\delta}} + 2\log \frac{3}{\delta})}{n} + o(1/n) \tag{D.9}$$

with probability at least $1 - \delta$. Now by Lemma 1 and Cauchy-Schwarz, we have that

$$U(\pi^*) - U(\hat{\pi}) \leq 2 \sum_x p_x \max_k |r_{xk} - \hat{r}_{xk}| \leq 2\|\hat{\theta} - \theta^*\|_{\Sigma(\pi)} \sum_x p_x \max_k \|\phi(x,k)\|_{\Sigma(\pi)^{-1}}. \tag{D.10}$$

Let $c(\pi) = \sum_x p_x \max_k \|\phi(x,k)\|_{\Sigma(\pi)^{-1}}$ be a data-dependent constant. From Lemma 2, there exists a RCT assignment stategy $\tilde{\pi}$ such that $c(\tilde{\pi}) \leq \sqrt{d}$.

Combining this with Equation D.10 and Equation D.9, we obtain that if we collect at least

$$n \geq \max\left\{6\rho_0^2 d \log \frac{3d}{\delta}, O\left(\frac{\sigma^2 d^2}{\epsilon^2}\right)\right\}$$

under data collection strategy $\tilde{\pi}$ then $U(\pi^*) - U(\hat{\pi}) \leq \epsilon$ with probability at least $1 - \delta$.

□

Lemma 2 (Modified Kiefer-Wolfowitz Theorem)**.** *Let* $\Pi = \{\pi \in \mathbb{R}^{|X|*K} | \sum_k \pi_{xk} = 1, \ \forall x\}$ *be the set of context-conditioned policy distributions. Then for any context distribution and feature space, we can design a contextual data collection strategy* $\tilde{\pi} \in \Pi$ *such that*

$$c(\tilde{\pi}) = \sum_x p_x \max_k \|\phi(x,k)\|_{\Sigma(\tilde{\pi})^{-1}} \leq \sqrt{d}.$$

Proof. Proof. We adapt the Kiefer-Wolfowitz Theorem concerning G-optimal experimental designs for our setting where we do not have full control over the sampling distribution, but rather can only control the policy distribution (not the context distribution).

For our proof, define $g(\pi) = \sum_x p_x \max_k \|\phi(x,k)\|^2_{\Sigma(\pi)^{-1}}$. Our goal will be to show that $\min_{\pi \in \Pi} g(\pi) = g(\tilde{\pi}) = d$ and by convexity,

$$c(\tilde{\pi})^2 = (\sum_x p_x \max_k \|\phi(x,k)\|_{\Sigma(\tilde{\pi})^{-1}})^2 \leq \sum_x p_x \max_k \|\phi(x,k)\|^2_{\Sigma(\tilde{\pi})^{-1}} = g(\tilde{\pi}) \leq d. \tag{D.11}$$

To show this we will first optimize $f(\pi) = \log\det\Sigma(\pi)$ and then show that $f(\pi)$ and $g(\pi)$ have the same optimizer $\tilde{\pi}$ and that $f(\tilde{\pi}) = g(\tilde{\pi}) = d$. Note that

$$\frac{\partial}{\partial \pi_{xk}} f(\pi) = \frac{1}{\det\Sigma(\pi)} \frac{\partial}{\partial \pi_{xk}} \det\Sigma(\pi) \tag{D.12}$$

$$= \mathrm{trace}\left(\frac{\mathrm{adj}(V(\pi))}{\det\Sigma(\pi)} p_x \phi(x,k)\phi(x,k)^T \right) \tag{D.13}$$

$$= \mathrm{trace}\big(\Sigma(\pi)^{-1} p_x \phi(x,k)\phi(x,k)^T\big) \tag{D.14}$$

$$= p_x \|\phi(x,k)\|^2_{\Sigma(\pi)^{-1}}. \tag{D.15}$$

Since f is concave, by first order optimality conditions, for any $\pi \in \Pi$ and $\tilde{\pi} = \arg\max_{\pi\in\Pi} f(\pi)$,

$$0 \geq \langle \nabla f(\tilde{\pi}), \pi - \tilde{\pi} \rangle = \sum_x p_x \sum_k \pi_{xk} \|\phi(x,k)\|^2_{\Sigma(\tilde{\pi})^{-1}} - \sum_x p_x \sum_k \tilde{\pi}_{xk} \|\phi(x,k)\|^2_{\Sigma(\tilde{\pi})^{-1}} \tag{D.16}$$

$$= \sum_x p_x \sum_k \pi_{xk} \|\phi(x,k)\|^2_{\Sigma(\tilde{\pi})^{-1}} - d \tag{D.17}$$

since for any π,

$$\sum_x p_x \sum_k \pi_{xk} \|\phi(x,k)\|^2_{\Sigma(\pi)^{-1}} = \mathrm{trace}(\sum_x p_x \sum_k \pi_{xk} \phi(x,k)\phi(x,k)^T \Sigma(\pi)^{-1}) = \mathrm{trace}(I_d) = d. \tag{D.18}$$

Thus letting $\pi_{xk} = \mathbb{1}\big\{k = \arg\max_{k'} \|\phi(x,k')\|_{\Sigma(\tilde{\pi})^{-1}}\big\}$ we have that

$$g(\tilde{\pi}) = \sum_x p_x \max_k \|\phi(x,k)\|^2_{\Sigma(\tilde{\pi})^{-1}} \leq d. \tag{D.19}$$

But it also follows that for any π,

$$g(\pi) = \sum_x p_x \max_k \|\phi(x,k)\|^2_{\Sigma(\pi)^{-1}} \geq \sum_x p_x \sum_k \pi_{xk} \|\phi(x,k)\|^2_{\Sigma(\pi)^{-1}} = d. \tag{D.20}$$

Therefore $\tilde{\pi}$ minimizes $g(\pi)$ and $g(\tilde{\pi}) = d$.

□

We note that if we know the context distribution we can efficiently solve for π^* since by the arguments of Lemma 2 we can solve the equivalent optimization problem

$$\begin{aligned} \max_{\pi} \quad & \log\det\Sigma(\pi) \\ \text{s.t.} \quad & 0 \preceq \pi \preceq 1 \end{aligned} \tag{D.21}$$

where $\Sigma(\pi) = \sum_x p_x \sum_k \pi_{xk} \phi(x,k)\phi(x,k)^T$. This is an example of a determinant maximizing problem subject to linear matrix constraints, which can be solved efficiently by interior point methods (Vandenberghe et al. [1998]).

Theorem 7 (Restatement of Theorem 3)**.** *Assume the reward is logistic, the costs are known, and that the assumptions D0, D1, D2, and C of Ostrovskii and Bach [2020] hold (these assumptions define problem-dependent constants $K0, K1, K2, \rho$). Also define Σ and c as in Theorem 2. Then, for any $\epsilon > 0, \delta > 0, \lambda_g \geq 0$ and*

$$n \geq O\bigg(\max\{K_2^4(d + \log\frac{1}{\delta}), \rho K_0^2 K_1^2 d^2 \log\frac{d}{\delta}, (\rho^2 c^2 K_1^2 d \log\frac{1}{\delta})/\epsilon^2\}\bigg)$$

we have $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) < \epsilon) > 1 - \delta$.

Proof. Proof. By Theorem 3.1 of Ostrovskii and Bach [2020] (in the well-specified case), for $n \geq O\big(\max\{K_2^4(d + \log\frac{1}{\delta}), \rho K_0^2 K_1^2 d^2 \log\frac{ed}{\delta}\}\big)$ with probability at least $1 - \delta$,

$$\|\hat{\theta}_n - \theta^*\|_H^2 \leq \frac{K_1^2 d \log\frac{e}{\delta}}{n} \tag{D.22}$$

where $H = \nabla^2 L_\pi(\theta^*)$ is the Hessian of the cross-entropy loss evaluated at the true parameter. By assumption C of Ostrovskii and Bach [2020], we assume that that the covariance matrix $\Sigma = \mathrm{Cov}_\pi[\phi(X, A)]$ is bounded above by H by a data-dependent factor ρ, i.e. that $\rho H - \Sigma$ is positive semi-definite. Thus by Lemma 1,

$$U(\pi^*) - U(\hat{\pi}) \leq 2\sum_x p_x \max_k |r_{xk} - \hat{r}_{xk}| \tag{D.23}$$

$$\leq 2\|\hat{\theta}_n - \theta^*\|_H \sum_x p_x \max_k \|\phi(x, k)\|_{H^{-1}} \tag{D.24}$$

$$\leq 2\sqrt{\frac{K_1^2 d \log\frac{e}{\delta}}{n}} \sum_x p_x \max_k \rho\|\phi(x, k)\|_{\Sigma^{-1}} \tag{D.25}$$

$$\leq 2c\rho\sqrt{\frac{K_1^2 d \log\frac{e}{\delta}}{n}}. \tag{D.26}$$

Thus it follows that if $n \geq O\Big(\max\{K_2^4(d + \log\frac{1}{\delta}), \rho K_0^2 K_1^2 d^2 \log\frac{ed}{\delta}, \frac{\rho^2 c^2 K_1^2 d}{\epsilon^2} \log\frac{e}{\delta}\}\Big)$ then $\mathbb{P}(U(\pi^*) - U(\hat{\pi}) \leq \epsilon) \geq 1 - \delta$.

□

We note that the assumptions D1 and D2 are quite restrictive (as explained in Remark 2.2 of Ostrovskii and Bach [2020]). The corresponding constants $K1, K2$ can depend on the magnitude true parameter θ^* and the data collection policy π. The authors also note that bounding these constants can be non-trivial, even when the context distribution is known. This makes designing a data collection strategy π that minimizes the higher order terms of the derived upper bounds much more difficult than in the linear setting, since this includes $K1, \rho, c$ which all depend on π.

E Experiment Details for Section 5.3

We define a subpopulation of clients for our simulated experiment from case data at the Santa Clara Public Defender Office according to the following process. First, we restrict our population to clients with recorded court dates between January 1st, 2010 and November 15th, 2021. Next, we limit our population to individuals who have stated that their race is white, or that their ethnicity is Vietnamese, or those who have stated that Vietnamese is their preferred language. We limit to these demographic groups to reflect the motivating example from Section 3.1. Finally, for consistency across case types and differences between court proceedings, we select only the first post-arraignment appearance for all individuals.

Next, we calculate a feature set x for each case describing: (1) whether the client identifies as Vietnamese; (2) whether the case is a felony; (3) whether the client identifies as male; (4) the client's age; (5) the natural log of the distance, in miles, between the client's home address and the courthouse, minus the natural log of the maximum allowed distance of 20 miles (so that all distance attributes are negative, with values of higher magnitude being closer to the courthouse); (6) the number of known failures to appear in the past two years; and (7) the inverse number of required court appearances in the past two years. We further restrict the population to only cases which have complete information on all the above attributes. The above process results in 12,646 example cases for use in our simulation.

With this information, we model the likelihood a client will appear in court with a logistic regression trained on the above population using the stated feature set. Specifically, we have:

$$\Pr(Y(0) = 1) \sim \text{logit}^{-1}(\beta x)$$

Once constructed, we modify the model to prepare for our simulation. First, we set β_0 to zero to set a mean population appearance rate of 50%. Next, we set β_1 to 1, which sets Vietnamese clients as more likely to appear in court, on average. This change—in addition to the fact that Vietnamese individuals tend to live farther away from court—magnifies the likelihood of a spending disparity between Vietnamese and white individuals in our simulation, given that treatment effects would tend to be lower, and costs higher, for Vietnamese clients. As a result, a preference for spending parity is at tension with simply allocating assistance to those with the highest treatment effect per dollar.

The predicted appearance probabilities from the model described above serve as the base for our simulation's structural equation model. To begin, we define three potential outcomes for each individual, corresponding to appearance in the absence of assistance ($k = 0$, i.e., that

predicted by the above model), appearance if provided with rideshare assistance ($k = 1$), and appearance if provided a transit voucher ($k = 2$). We do so in terms of the following structural equation:

$$\begin{aligned} f_Y(k, x, u) \sim \mathbb{1}(u \leq \text{logit}^{-1}(\text{logit}(\Pr(Y(0) = 1)) + \\ \gamma_1 \cdot \mathbb{1}(k = 1) + \\ \gamma_2 \cdot \mathbb{1}(k = 2) \cdot x_{\text{dist}})) \end{aligned}$$

where we set γ_1 to 4 and γ_2 to -0.75. Finally, for a latent variable $U_Y \sim \textsc{Unif}(0, 1)$, we define the potential outcomes:

$$Y(a) = f_Y(a, X, U_Y).$$

This structure ensures that $Y(0) \leq Y(1)$ and that $Y(0) \leq Y(2)$, meaning that receiving any form of assistance is always better than no assistance. Further, the type of assistance—transit voucher or rideshare assistance—that is best for each individual varies across the population.

As described in the main text, the utility U is defined by Eq. (2), where we set $r(x, a, y) = y$ and $\lambda_g = 0.0004$. In other words,

$$\begin{aligned} U(\pi) = \mathbb{E}[Y(\pi(X))] \\ - \sum_{g \in \mathcal{G}} \lambda_g \Big| \mathbb{E}_X[c(X, \pi(X)) \mid g \in s(X)] - \mathbb{E}_X[c(X, \pi(X))] \Big|. \end{aligned}$$

The first term in U is the expected number of clients 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 balanced allocation between simple appearance maximization and perfect demographic parity. 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 and Su, 2020]. For UCB, we used the 97.5th percentile estimate of the posterior of $g(x, a, \hat{\theta}_i)$.

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 randomly selected warm-up group of 4 people, with at least one male and at least one Vietnamese client. During this period, the first two clients were assigned to actions $k = 1$ and $k = 2$, respectively. The other two individuals were assigned to control, i.e., $k = 0$. The treatments during this warm-up period are not included as expenditures against the overall budget b.

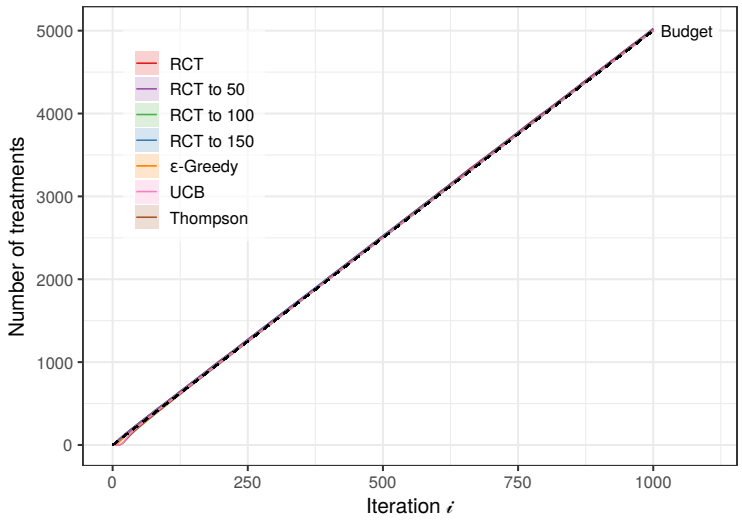


Figure E.1: Mean spending by method across 2,000 simulations. The budget is illustrated with a dashed line.

For our simulation, we set a per-person budget of \$5. We also set round-trip rideshare costs at \$10 per mile, and daily transit voucher costs at \$7.50, reflecting typical prices observed in Santa Clara county. 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 iteration i, we compute a new budget b_i^*:

$$b_i^* = b \cdot \frac{b \cdot (i-1)}{\sum_{j=1}^{i-1} c(x_j, A_j)}, \tag{E.27}$$

where A_j is the action taken on the j-th individual, and b is the target budget. In Figure E.1, we show that all approaches included in our simulation spend the allowed budget.

For our RCT and ϵ-greedy approaches, care must be taken to define "random selection" when handling both varying costs and an overall per-person budget. For example, an RCT that selects all available treatments with equal probability could overshoot the budget if rides cost \$100 on average and the per-person budget is \$5. To avoid this outcome, we first calculate the expected cost of all actions—including both costly and no-cost actions—when following random allocation, and then calculate the proportion p of individuals to whom

we can afford to allocate randomly:

$$p = \frac{b}{\tilde{c}} \text{ where } \tilde{c} = \frac{\sum_k \mathbb{E}_X[c(x, a_k)]}{k}. \tag{E.28}$$

Once calculated, we randomly select p of the population to receive a random allocation, and treat the remainder of the population $1 - p$ with the no-cost treatment, which ensures we meet our budget in expectation.

Our optimization procedure (our linear program) 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 $n = 1{,}000$ clients, and approximate the full population by a discrete distribution over this observed sample, with each client 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 client in the initial sample $\mathcal{C}$. Consequently, when making decisions for a new individual with covariates x, we act according to the learned policy for the most similar client in $\mathcal{C}$—among those having the same group membership $s(x)$ as the new client—where similarity is defined in terms of the estimated reward $\hat{f}(x, a_k)$ normalized by the cost of that treatment $c(x, a_k)$. Specifically, for a new client, we define its nearest neighbor $\mathrm{NN}(x)$ to be:

$$\mathrm{NN}(x) = \underset{\substack{x' \in \mathcal{C} \\ s(x') = s(x)}}{\arg\min} \left\| \frac{\hat{f}(x', \cdot)}{c(x', .)} - \frac{\hat{f}(x, \cdot)}{c(x, .)} \right\|_2.$$

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