https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–18
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

Alex Chohlas-Wood,[a,*] Madison Coots,[b] Henry Zhu,[c] Emma Brunskill,[c] Sharad Goel[b]

[a] Department of Applied Statistics, Social Science, and Humanities, New York University, New York, New York 10003; [b] Harvard Kennedy School, Harvard University, Cambridge, Massachusetts 02138; [c] Department of Computer Science, Stanford University, Stanford, California 94305

*Corresponding author

Contact: alex.cw@nyu.edu, https://orcid.org/0000-0002-8279-6270 (AC-W); mcoots@g.harvard.edu, https://orcid.org/0009-0002-9819-8985 (MC); hyzhu@stanford.edu (HZ); ebrun@cs.stanford.edu, https://orcid.org/0000-0002-3971-7127 (EB); sgoel@hks.harvard.edu, https://orcid.org/0000-0002-6103-9318 (SG)

Received: February 3, 2022
Revised: January 31, 2023; February 12, 2024
Accepted: April 12, 2024
Published Online in Articles in Advance: December 18, 2024

https://doi.org/10.1287/mnsc.2022.00345

Abstract. In an attempt to make algorithms *fair*, the machine learning literature has largely focused on equalizing decisions, outcomes, or error rates across race or gender groups. To illustrate, consider a hypothetical government rideshare program that provides transportation assistance to low-income people with upcoming court dates. Following this literature, one might allocate rides to those with the highest estimated treatment effect per dollar while constraining spending to be equal across race groups. That approach, however, ignores the downstream consequences of such constraints and, as a result, can induce unexpected harm. For instance, if one demographic group lives farther from court, enforcing equal spending would necessarily mean fewer total rides provided and potentially more people penalized for missing court. Here we present an alternative framework for designing equitable algorithms that foregrounds the consequences of decisions. In our approach, one first elicits stakeholder preferences over the space of possible decisions and the resulting outcomes—such as preferences for balancing spending parity against court appearance rates. We then optimize over the space of decision policies, making trade-offs in a way that maximizes the elicited utility. To do so, we develop an algorithm for efficiently learning these optimal policies from data for a large family of expressive utility functions. In particular, we use a contextual bandit algorithm to explore the space of policies while solving a convex optimization problem at each step to estimate the best policy based on the available information. This consequentialist paradigm facilitates a more holistic approach to equitable decision making.

History: This paper has been accepted by Catherine Tucker, Special Issue on The Human-Algorithm Connection.

Funding: This work was supported by Stanford Impact Labs [Start-Up Impact Labs/GEJBU], the Stanford Institute for Human-Centered Artificial Intelligence, and Harvard Data Science Initiative.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.00345.

Keywords: information systems: enabling technologies (includes AI • machine learning • and data mining technologies) • artificial intelligence • decision analysis: multiple criteria • simulation: statistical analysis

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 (Latessa et al. 2010, Cadigan and Lowenkamp 2011, Milgram et al. 2014, Goel et al. 2018); in healthcare, algorithms identify which individuals receive limited resources, including HIV prevention counseling and kidney replacements (Friedewald et al. 2013, Obermeyer et al. 2019, Wilder et al. 2021); and in child services, screening decisions are based on the estimated risk of adverse outcomes (Shroff 2017, Chouldechova et al. 2018, Brown et al. 2019, De-Arteaga et al. 2020). In these applications and others, equity is a central concern. In particular, the machine learning community has proposed numerous methods to constrain predictions to

achieve formal statistical properties, such as parity in decision rates or error rates across demographic groups (Chouldechova and Roth 2020, Barocas et al. 2023, Chohlas-Wood et al. 2023a, Corbett-Davies et al. 2023).

To illustrate this traditional approach to designing equitable algorithms, consider a government agency that provides free rides for people to get to court (Brough et al. 2022). Missed court dates can lead to severe penalties, including incarceration, and so, improving court appearance rates can reduce social harm (Chohlas-Wood et al. 2023b). When designing this program, one might first use historical data to estimate the effect of a ride on increasing each person's likelihood of appearing at court, as well as the cost of providing them with a ride. Then, in an effort to distribute benefits fairly, one might allocate assistance to those with the highest estimated benefit per dollar while constraining per-person spending to be equal across demographic groups. The implicit hope in past literature is that one achieves fairness by imposing an axiomatic constraint on decisions: spending parity.

Although intuitively reasonable, axiomatic approaches to fairness can cause unexpected harm. For example, imagine members of one group live farther from the courthouse, making it more costly to provide them rides. Enforcing equal spending across groups would typically result in fewer rides overall and, accordingly, lower appearance rates. More generally, traditional axiomatic approaches to fairness typically do not consider the downstream consequences of constraints and thus fail to engage with the difficult trade-offs at the heart of many policy problems.

We propose an alternative, consequentialist framework to algorithmic fairness. In this framework, rather than imposing fairness axioms, one begins by eliciting stakeholder preferences over the space of potential decisions and resulting outcomes. For example, in designing our hypothetical transportation program, one would assess the degree to which stakeholders are willing to trade court appearances for reductions in spending disparities across groups. Then, using these preferences, we compute a decision-making policy with the largest expected utility while adhering to budget constraints. Given historical data on decisions and outcomes, we show that optimal decision policies can be efficiently derived for a large and expressive family of utility functions by solving a linear program (LP).

We further show how to efficiently learn optimal policies while rolling out new programs in the absence of historical data. Our approach here is inspired by the success of Thompson sampling (Chapelle and Li 2011) and optimism-under-uncertainty methods (Auer et al. 2002) in multiarmed bandits. In contrast to the standard contextual multiarmed bandit setting, we consider a multifaceted, structured objective to account for complex preferences and budget constraints inherent to many real-world applications. As such, our actions at each iteration are guided by solving an LP as above.

The rest of our paper is structured as follows. In Section 2, we review the related literature, connecting and contrasting our approach to ideas in fair machine learning, fair division, multiobjective optimization, and reinforcement learning. In Section 3, we illustrate the trade-offs inherent to many policy problems—and the concomitant benefits of a consequentialist perspective over an axiomatic approach. To do so, we draw on client data from the Santa Clara County Public Defender Office to consider the costs and benefits of a hypothetical transportation assistance program. We also describe the results of a survey that gauged stakeholders' willingness to sacrifice court appearances to reduce spending disparities across race groups. Given such preferences as well as historical data on outcomes, in Section 4, we formally state and solve the corresponding policy optimization problem. In Section 5, we theoretically derive sample complexity bounds on learning optimal policies in the absence of historical data. Finally, in Section 6, we introduce and evaluate an adaptive approach to learning optimal policies, combining contextual bandits with the optimization solution described in Section 4. We end with some concluding thoughts in Section 7.

2. Related Work

Our work draws on research in algorithmic fairness, fair division, multiobjective 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 (Blodgett and O'Connor 2017, Caliskan et al. 2017, Shroff 2017, Buolamwini and Gebru 2018, Chouldechova et al. 2018, Datta et al. 2018, Goodman et al. 2018, Ali et al. 2019, De-Arteaga et al. 2019, Obermeyer et al. 2019, Raji and Buolamwini 2019, Koenecke et al. 2020, Chohlas-Wood et al. 2023a) and associated development of formal criteria to characterize fairness (Chouldechova and Roth 2020, Gupta et al. 2020, Barocas et al. 2023, Corbett-Davies et al. 2023). 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, Kusner et al. 2017, Chiappa and Isaac 2018, Nabi and Shpitser 2018, Zhang and Bareinboim 2018, Wang et al. 2019, Wu et al. 2019, Coston et al. 2020, Nyarko et al. 2021, Nilforoshan et al. 2022).

All 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 axiomatic approach, which has otherwise dominated the fair machine learning literature (Corbett-Davies et al. 2017, Cowgill and Tucker 2020, Kasy and Abebe 2021, Grgić-Hlača et al. 2022). Some recent exceptions have begun to consider algorithmic decision making from a consequentialist perspective (Barabas et al. 2018, Liu et al. 2018, Card and Smith 2020, Coston et al. 2020, Donahue and Kleinberg 2020, Fang et al. 2022, Nilforoshan et al. 2022, Viviano and Bradic 2024). For example, Nilforoshan et al. (2022) show that common causal definitions of algorithmic fairness lead to Pareto-dominated policies. However, although these papers adopt a consequentialist approach to varying degrees, they do not consider the problem of efficiently learning optimal policies, as we do here.

In a related thread of research on fair division problems, groups of individuals decide how to split a limited set of resources among themselves (Brams et al. 1996, Bertsimas et al. 2011, Caragiannis et al. 2012, Gal et al. 2017). 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 the 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 multiobjective environments (Zuluaga et al. 2013). Such inherent trade-offs have recently been 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 (Chu and Ghahramani 2005, Fürnkranz and Hüllermeier 2010, Lin et al. 2020).

One particularly challenging aspect of our setting is handling budget constraints (e.g., we may only be able to provide transportation assistance to a limited number of clients). 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. A related study (Patil et al. 2021) on online multiarmed bandits considered minimizing regret while ensuring that each arm is played a minimal number of times but did not consider context-specific decision policies and fairness in resource allocations or budget constraints, as we do here. Budget constraints have been considered in a more general form of knapsack constraints in bandit settings. Slivkins (2019, chapter 10) provides a review of such work, focusing on the primary literature, which has considered the (noncontextual) multiarmed bandit setting. Earlier work on contextual multiarmed bandits with knapsacks (Badanidiyuru et al. 2014, Agrawal et al. 2016a) provided regret bounds but lacked computationally efficient implementations. Agrawal et al. (2016b) 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 multiobjective 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. Selecting Policies in the Presence of Trade-Offs

We begin, in Section 3.1, by describing our motivating example of providing transportation to individuals with mandatory court dates. Using client data from the Santa Clara County Public Defender Office, we show that allocating benefits to maximize appearance rates induces spending disparities across race groups. Then, in Section 3.2, we continue by explicitly illustrating the inherent tension between maximizing appearance rates and equalizing spending—and arguing that popular axiomatic approaches to fairness can lead to unintended harm. Finally, in Section 3.3, we describe the results of a survey aimed at eliciting people's willingness to trade court appearances for lower spending disparities.

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 (Fishbane et al. 2020, Chohlas-Wood et al. 2023b). Despite these consequences, some individuals struggle to attend court because of significant transportation barriers (Mahoney et al. 2001, Brough et al. 2022, Allen 2024). 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. 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. 2018, Lyft 2020, Vais et al. 2020, Fraade-Blanar et al. 2021). As we discuss in Section 7, it is important to note that there are many alternative policy approaches to this issue, including discouraging judicial use of incarceration after an individual misses court.

A natural algorithmic approach for allocating 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 each individual's likelihood of appearance in the absence of assistance, $\hat{Y}_i(0)$, and their likelihood of appearance if provided with a ride, $\hat{Y}_i(1)$. These probabilities might, for example, be estimated using historical data on past outcomes, or a randomized experiment. 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 budget is exhausted.

This strategy aims to achieve the highest appearance rate given the available budget. However, in so doing, it implicitly prioritizes those who live 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 geographically segregated by race (Figure 1(a)). In particular, Santa Clara's Vietnamese population, one of the county's largest ethnic minorities, does not tend to live as close to the courthouse as other racial or ethnic groups, including White individuals.

To understand the impacts of a strategy that optimizes exclusively for appearance, we start with a data set of 65,193 court dates handled by SCCPDO between January 1, 2017, and December 18, 2023. For the sake of consistency, this population of court dates consists solely of clients' first court date after arraignment. For clients with court dates after January 1, 2021, we use the historical data from 2017–2020 to model $Y_i(0)$ with a logistic regression model based on age, race/ethnicity, offense severity (misdemeanor or felony), two-year appearance history, the day of the week and month of the court appearance, and the distance from the client's home 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 in each direction, in line with current rideshare prices.

Under the naive optimization approach outlined above, Figure 1(b) shows per-capita spending for White and Vietnamese clients across different overall transportation budgets. For example, given an annual budget of \$50,000, a policy that allocates rides to those with the highest estimated treatment effect per dollar would end up spending, on average, \$6.86 for every White client, but only \$4.54 on average per Vietnamese client. Policymakers and other stakeholders may deem this disparity to be undesirable and may thus be willing to accept lower overall appearance rates in return for more equal spending across groups.

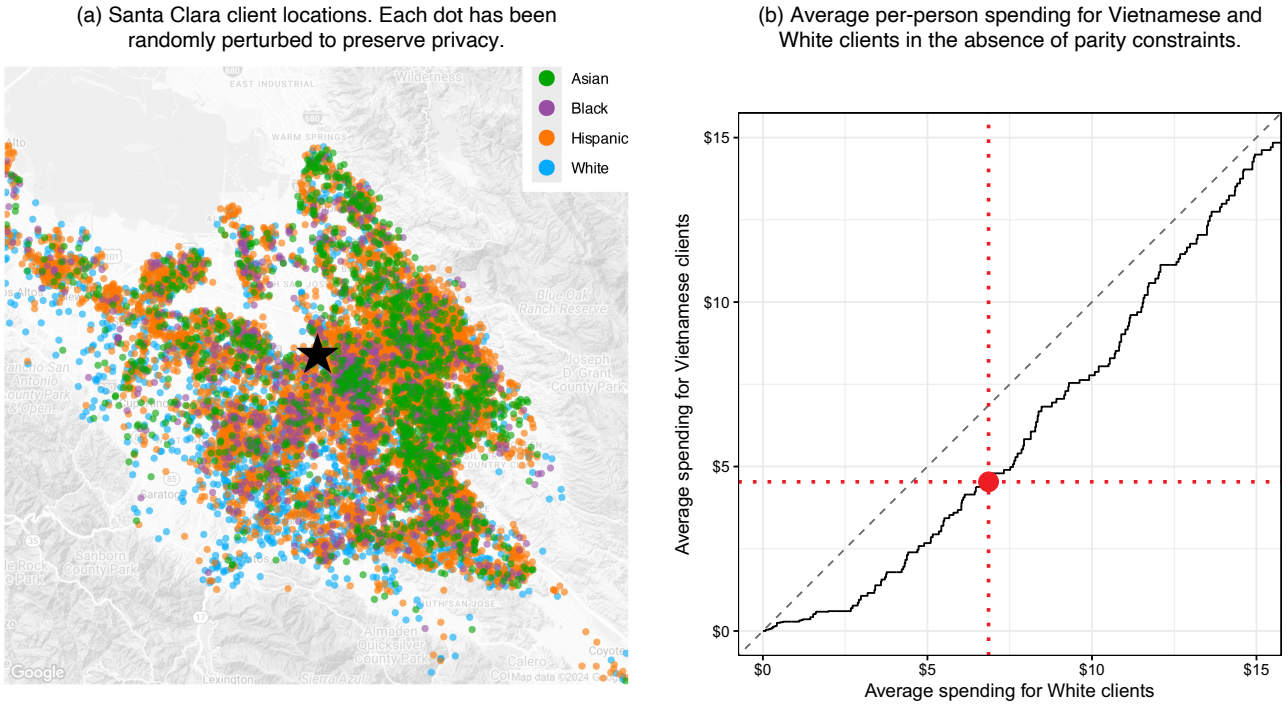
3.2. Exploring Inherent Trade-Offs

To further explore the trade-off between appearance rates and spending parity, we now consider a synthetic client population with 5,000 Black and 5,000 White clients. For simplicity, we assume that each client has a 75% chance of appearing at court in the absence of rideshare assistance and is guaranteed to appear if provided a ride. Further, we set a fixed annual budget of \$5 per person, or \$50,000 total. Finally, we assume that Black clients live farther from court on average. Consequently, the average expected treatment effect per dollar is lower for Black clients than for White clients. This pattern induces a tension between maximizing total appearances and equalizing spending across the two groups.[1] We describe the data-generating process for this synthetic population in detail in Online Appendix EC.3.

In Figure 2, 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 Black clients (on the horizontal axis). 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.

Along the Pareto frontier, a policymaker ostensibly has more and less preferred outcomes. For example, imagine that a given policymaker's utility is maximized at the blue point on the curve. In contrast, the point at the crest of the curve (in green) achieves the highest number of overall appearances but is a suboptimal policy because it underspends on Black clients, at least according to the stakeholder's preferences. Similarly, a policy that achieves perfect spending parity (i.e., the purple point) also yields suboptimal outcomes relative to the policymaker's preferences because too many appearances are

Figure 1. (Color online) Empirical Illustration of Potential Spending Disparities in Santa Clara County



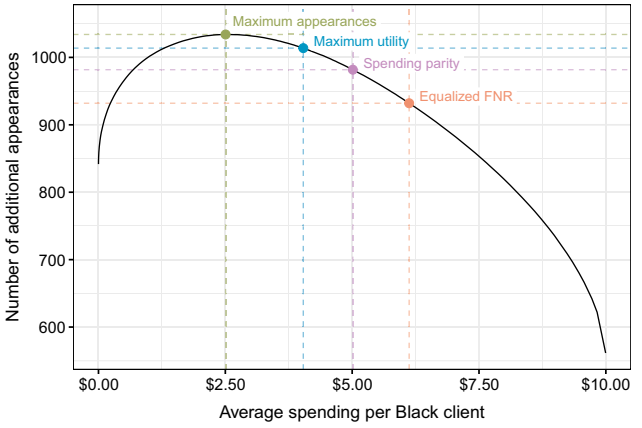
Notes. The map in (a) shows the geographic distribution of the client base of the Santa Clara County Public Defender Office. The star on the map marks the location of the main county courthouse, where most clients are required to appear for court appointments. The plot in (b) explores the consequence 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 shows that a hypothetical annual ride budget of $50,000 would result in an average per-person spending amount of $6.86 for White individuals and an average per-person spending amount of $4.54 for Vietnamese individuals.

lost in order to achieve spending parity. We also plot the point on the curve corresponding to equal false-negative rates (FNR) between groups (in pink).[2] A constraint that demands error-rate parity—as opposed to maximizing utility more directly—can again result in a suboptimal balance between maximizing appearances and evenly distributing transportation assistance relative to the underlying preferences of the policymaker. In contrast to the axiomatic approach common to past work, this simple example helps illustrate the value of viewing decisions from a consequentialist perspective.

3.3. Eliciting Preferences

We now empirically examine preferences for allocating transportation assistance in our hypothetical scenario above. To do so, we designed and administered a poll to a diverse sample of 297 Americans. We ran our survey on the Prolific platform, selecting the platform's "U.S. representative sample" option to recruit respondents, where respondents' self-identified sex, age, and ethnicity are comparable to a random population of U.S. adults, as determined by the U.S. Census. Survey respondents learned about our running example of providing clients with free rides to court and then read a short description of the hypothetical jurisdiction described above. (This prompt is included in full in Online Appendix EC.4). We then asked respondents to select their preferred trade-off among five possible options drawn from

Figure 2. (Color online) The Pareto Frontier for a Stylized Population Model Showing the Trade-Off Between Appearances and Spending per Black Client



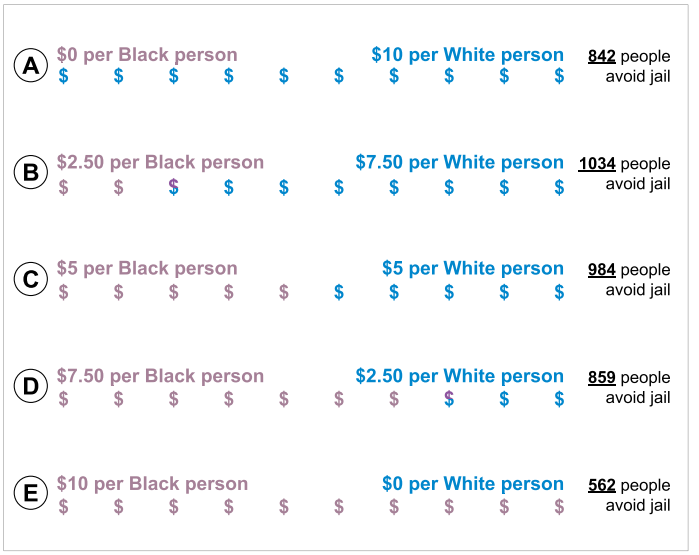
Notes. The vertical axis shows expected additional appearances relative to a policy that does not provide rideshare assistance to any clients. Under this model, common heuristics (e.g., maximizing appearances and demanding demographic or error-rate parity) lead to suboptimal policies.

the Pareto frontier in Figure 2. To aid in their decision, participants were shown the graphic depicted in Figure 3(a). Participants were randomly shown either an ascending or descending version of this graphic to mitigate anchoring to the first options shown.

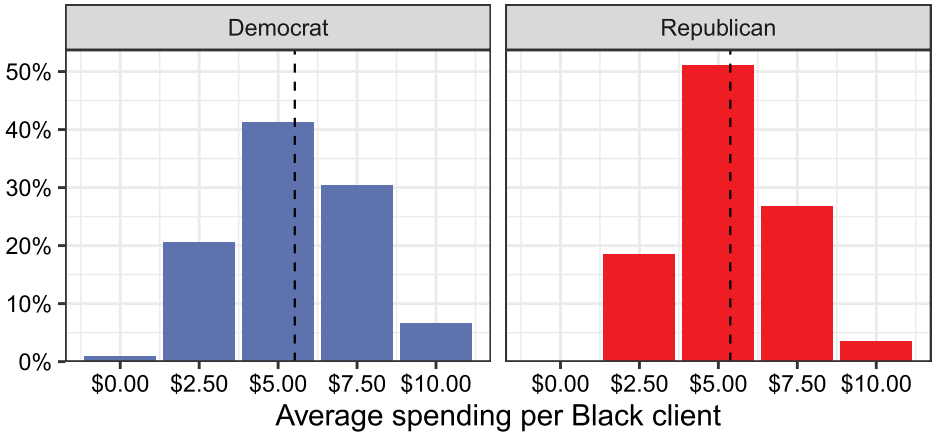
Our survey results are presented in Figure 3(b) and illustrate two key points. First, the vast majority of respondents prefer trading at least some "efficiency" (i.e., as measured by the total number of people who avoid jail because of receiving transportation assistance) in order to spend more money on Black clients. This broad preference for incorporating equity considerations into algorithmic decision making mirrors past results (Koenecke et al. 2023). Second, there is substantial heterogeneity in preferences that elides traditional group boundaries. For example, there is considerable variation in preferences within self-identified Democrats and Republicans; at the same time, the average preference is comparable across these two groups. We observe similar patterns across a number of other demographic characteristics of the respondents, including gender and race/ethnicity, as we show in Online Appendix EC.4. These results suggest that traditional axiomatic approaches to algorithmic fairness—which do not consider the specific context of decisions—risk yielding policies that do not reflect the preferences of stakeholders. In contrast, a more consequentialist perspective allows us to develop algorithms that better

Figure 3. (Color online) Ride Allocation Preferences Among Survey Respondents

(a) Graphic shown to survey participants.



(b) Survey results from 297 respondents, split by self-identified U.S. political party affiliation. Mean preferred allocations are represented by the vertical dashed line.



Notes. The graphic in (a) was shown to survey participants to help them select their preferred ride allocation policy. In this hypothetical scenario, option B maximizes appearances, whereas option C corresponds to spending parity. The survey results in (b) show that both Democrats and Republicans prefer policies that spend roughly equal amounts on Black and White clients, but there is a wide range of preferences among members of both groups.

balance the difficult trade-offs inherent to many policy problems.

4. Computing Equitable Policies

For the rideshare example in the previous section, it is computationally straightforward to trace out the Pareto frontier: for any fixed budget allocated to each group, one can maximize appearances by offering rides to those clients with the largest (estimated) gain in appearance rate per dollar while constraining spending to the allotted per-group budget. (We formally show the optimality of this strategy in Online Appendix EC.2). As a result, given preferences over various outcomes (e.g., trading off appearances with spending parity), one can efficiently determine the utility-maximizing allocation strategy. However, in more complicated scenarios—with more complex preferences and potential actions—it is not immediately clear how to find optimal allocation strategies, even when preferences and treatment effects are fully known. Fortunately, for a large class of preferences, it is indeed feasible to efficiently compute utility-maximizing policies, as we now describe. In Section 6, we consider the problem of learning optimal polices when preferences are known but treatment effects are not.

To generalize from our running example, consider a sequential decision-making setting where, at each time step, one first observes a vector of covariates X_i drawn from a distribution $\mathcal{D}_X$ supported on a finite state space $\mathcal{X}$, and then must select one of K actions from the set $\mathcal{A} = \{a_1, \ldots, a_K\}$. For example, in our motivating application, X_i might encode an individual's demographics, history of appearance, alleged charges, and distance from court, and the set of actions might specify whether 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., $c(x, a_0) = 0$), whereas all other available actions are costly (i.e., $c(x, a_k) > 0$ for $k > 0$).

To arbitrate between feasible policies (i.e., those that adhere to the budget constraint), policymakers might consider both the direct outcomes of a policy (e.g., on appearance rates) and the relative allocation of benefits across demographic groups. To formalize this idea, we suppose 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 one's court date if offered rideshare assistance, and $Y_i(0)$ may indicate the outcome if assistance was not provided.

Now, to facilitate computation, we assume a policymaker's utility $U(\pi)$ of any decision policy π can be approximated by a flexible function of the following form:

$$U(\pi) = \mathbb{E}_{X,Y}[r(X, \pi(X), Y(\pi(X)))] - \sum_{\ell=1}^{L} \sum_{g \in \mathcal{G}} \lambda_{g,\ell} \left| \mathbb{E}_{X,Y}[f_\ell(X, \pi(X), Y(\pi(X))) \mid g \in s(X)] - \mathbb{E}_{X,Y}[f_\ell(X, \pi(X), Y(\pi(X)))] \right|, \quad (1)$$

where r and f_ℓ are fixed functions that parameterize this class of utilities, $|\cdot|$ is an absolute value, $\lambda_{g,\ell}$ are nonnegative constant parameters, 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, and the second term penalizes differences in allocations and outcomes across groups. For example, in our motivating application, we might set

$$r(x, a, y) = (a + c_1 y) \cdot (1 + c_2 \cdot \mathbb{I}_{\text{frequent}}(x)), \quad (2)$$

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. (In Equation (2), we do not multiply a by a constant because the overall scale of r is arbitrary.) 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).

In addition to preferring transportation assistance policies that boost appearance rates, a policymaker might also prefer those for which we spend similar amounts per person across demographic groups to ensure such

investments are broadly applied across an agency's jurisdiction. The second term of $U(\pi)$ can be used to encode these parity preferences. For example, setting $f(x,a,y) = c(x,a)$ would encode a preference for spending parity. Depending on the application, one could imagine similarly penalizing a given policy if the distribution of *actions* or *successes* were unequal across groups.

In practice, to encode preferences in this way, one might first show stakeholders anticipated outcomes of various hypothetical policies, akin to our survey above. We could then sweep over parameters to produce a utility function of the appropriate form that accurately captures the elicited preferences. Importantly, and in contrast with an axiomatic approach, our consequentialist paradigm is predicated on the belief that there are not universal, context-independent constraints on policies. Rather, the utility of a policy depends critically on how much one objective must be sacrificed to achieve another.

Given this setup, our goal is to find a policy π^* that maximizes utility while staying within budget. Formally, we seek to solve the following optimization problem:

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

We next discuss how to efficiently solve this optimization problem.

4.1. Computing Optimal Decision Policies

To compute optimal policies, we assume in this section that one knows the distribution of X and the conditional distribution of the potential outcomes $Y(a_k)$ given X—that is, $\mathcal{D}(X)$ and $\mathcal{D}(Y(a_k)|X)$. (In Section 6, we consider how to learn optimal policies when historical data on treatment effects are not known.) Given this information, we show the optimization problem in Equation (3) can be expressed as a linear program, 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

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

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 $\mathbb{E}_X[c(X,\pi(X))] \le b$ in Equation (3) can be expressed as a linear inequality on the decision variables:

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

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

$$\begin{aligned} U(\pi) = \sum_{x,k} \mathbb{E}_Y[r(x,a_k,Y(a_k))|X=x] \cdot \Pr(X=x) \cdot v_{x,k} \\ - \sum_{\ell} \sum_{g} \lambda_{g,\ell} \left| \sum_{x,k} \left(\frac{\mathbb{I}(g \in s(x)) \Pr(X=x)}{\Pr(g \in s(X))} \right.\right. \\ \cdot \mathbb{E}_Y[f_\ell(x,a_k,Y(a_k))|X=x] \\ - \Pr(X=x) \cdot \mathbb{E}_Y[f_\ell(x,a_k, \\ \left.\left. Y(a_k))|X=x] \right) v_{x,k} \right|. \end{aligned}$$

Because of 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

$$\alpha^T v - \sum_{g,\ell} \lambda_{g,\ell} |\beta_{g,\ell}^T v|, \tag{4}$$

where α and β are constant vectors. We can rewrite this optimization problem as a linear program that includes additional (slack) variables $w_{g,\ell}$:

$$\begin{aligned} \text{Maximize}: \quad & \alpha^T v - \sum_{g,\ell} \lambda_{g,\ell} w_{g,\ell} \\ \text{Subject to}: \quad & 0 \le w_{g,\ell}, \\ & -w_{g,\ell} \le \beta_{g,\ell}^T v \le w_{g,\ell}. \end{aligned} \tag{5}$$

For completeness, we include a proof of this equivalence in Online Appendix EC.1.

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

$$\begin{aligned} &\text{Maximize}: \\ &\sum_{x,k} \mathbb{E}_Y[r(x,a_k,Y(a_k))|X=x] \cdot \Pr(X=x) \cdot v_{x,k} \\ &\qquad - \sum_{g,\ell} \lambda_{g,\ell} w_{g,\ell} \\ &\text{Subject to}: \\ &v_{x,k},\ w_{g,\ell} \ge 0 \qquad \forall x,k,g,\ell \\ &\sum_{k} v_{x,k} = 1 \qquad \forall x, \\ &\sum_{x,k} \Pr(X=x) \cdot v_{x,k} \cdot c(x,a_k) \le b, \text{ and} \\ &-w_{g,\ell} \le \sum_{x,k} \left(\frac{\mathbb{I}(g \in s(x)) \Pr(X=x)}{\Pr(g \in s(X))} \right. \\ &\qquad \cdot \mathbb{E}_Y[f_\ell(x,a_k,Y(a_k))|X=x] \\ &\qquad \left. - \Pr(X=x) \cdot \mathbb{E}_Y[f_\ell(x,a_k,Y(a_k))|X=x] \right) \\ &\qquad \cdot v_{x,k} \le w_{g,\ell} \qquad \forall g,\ell. \end{aligned}$$

Our approach above is a computationally efficient method for finding optimal decision policies. In theory, linear programming is (weakly) polynomial in the size of the input $O(|\mathcal{X}|K + |\mathcal{G}|L)$ variables and constraints in our case. In practice, using open-source software running on conventional hardware, we find it takes approximately one to two seconds to solve random instances of the problem on a state space of size $|\mathcal{X}| = 1{,}000$ with $|\mathcal{G}| = 10$ groups, $K = 5$ treatment arms, and $L = 1$ parity penalties.[3]

5. Sample Complexity Bounds on Learning Optimal Policies

To solve our policy optimization problem, we have thus far assumed perfect knowledge of the distribution of potential outcomes $\mathcal{D}(Y(a_k)|X)$, which allows us to compute the necessary inputs for our linear program. In reality, however, this distribution must typically be learned from observed data. One common approach for estimating the impact of actions is to run an experiment in which actions are randomly allocated, potentially in a way to ensure that all actions are taken an equal number of times or to ensure each group of interest experiences all actions evenly. Note that such data collection strategies do not adapt in response to observed outcomes of actions (such as some actions yielding higher appearance rates partway through data collection). In Section 5.1, we formally analyze these non–outcome-adaptive 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 discuss some considerations relating to experimental cost. We present this initial analysis to highlight how in some cases, the amount of data needed may not differ substantially from simpler objectives that do not involve parity constraints. The other benefit of this first analysis is that it involves creating an experimental design for data collection in advance, which makes the resulting data easily suitable for standard statistical inference. However, in practice, there can be significant benefits to changing the data-gathering strategy over the course of an experiment, as more effective actions can be prioritized faster. In Section 6, we demonstrate this through an alternative, contextual bandit–based data collection strategy that can often learn optimal policies more efficiently by judiciously exploring the effects of actions. We demonstrate the advantages of this alternative strategy in an empirically grounded simulation study.

5.1. Sample Complexity Bounds

A natural concern for practitioners is whether balancing multiple complicated objectives—like the competing outcomes highlighted in our utility function in Equation (1)—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 samples needed to construct near-optimal policies with high probability, focusing on spending parity by setting $f(x, a, y) = c(x, a)$ (following our running example). Our aim in this analysis is not to provide tight sample complexity bounds but, rather, to examine at a high level how additional parity objectives and modeling choices affect the amount of data required. Our results suggest that one may not need much more data to learn a multiobjective policy that incorporates equity preferences compared with a single-objective reward-maximizing policy, and that a known structure on the data-generating process can substantially reduce the amount of data required.

Our work is related to a deep literature in multiarmed bandits and contextual multiarmed bandits (see Lattimore and Szepesvári (2020) for a fairly recent textbook overview). The majority of this research has focused on providing cumulative regret guarantees of online, adaptive algorithms for a wide range of settings, including seminal results for finite armed bandits (Auer et al. 2002) and linear contextual bandits (Abbasi-Yadkori et al. 2011), as well as more recent interest in logistic models (e.g., Li et al. 2017, Jun et al. 2021). Approaches that minimize cumulative regret bounds can be different from algorithms that provide sample complexity bounds that are probably approximately correct (PAC)—that is, methods that, after a sufficient amount of data, output a decision policy that is near optimal with high probability.

Interestingly, prior work (e.g., Jin et al. 2018, section 3.1) has provided an online-to-batch reduction that can be used to convert a contextual multiarmed bandit algorithm with a cumulative regret result to a sample complexity bound on the number of samples needed to extract a near-optimal policy with high probability. However, most contextual MAB algorithms with cumulative regret guarantees rely on selecting actions under the principle of optimism under uncertainty with respect to the immediate estimated reward for the current context. The resulting regret bound is defined with respect to the best action that could have been selected. In contrast, in our setting, the objective is to compute a policy π that maximizes the utility function $U(\pi)$, which includes both a reward maximization term and a spending parity term. In general, the optimal policy in our setting will not match an optimal policy that maximizes only the reward. This implies we cannot directly leverage an online-to-batch reduction from existing algorithms with cumulative regret bounds because the regret bounds provided by those algorithms will not provide regret bounds for our setting. To our

knowledge, none of the existing online contextual bandit algorithms consider additional parity objectives, or a joint policy across contexts, as in our work.

There is fairly limited work on MAB and contextual MAB algorithms that directly provide PAC guarantees. Mannor and Tsitsiklis' (2004) foundational work provided sample complexity bounds for multiarmed bandits with a finite set of arms, and we will build on their work for providing sample complexity bounds for our setting given a finite set of contexts and arms/actions, also known as the tabular setting. Concurrent with the development of this work, there has been some recent interest in sample complexity bounds for contextual bandits (e.g., Zanette et al. 2021, Li et al. 2022, Pacchiano et al. 2023), which we will discuss further under different assumptions of the underlying data-generating process.

We now introduce some additional assumptions. As in Sections 4 and 4.1, we further assume throughout this section that the state space $\mathcal{X}$ is finite and that the costs and distribution of X are known. In practice, information on the distribution of X can often be estimated from historical data before any interventions are attempted. Let π^* be an optimal policy solution, as defined in Equation (3), with corresponding utility $U(\pi^*)$. We define the estimated utility function $\hat{U}(\pi)$ for a particular decision policy as

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

where $\hat{r}$ is the estimated reward function learned from data. Let $\hat{\pi}$ be a solution to the optimization problem in Equation (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 $Y_{X=x}(a_k)$.

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

Lemma 1. *The loss of utility because of using* $\hat{\pi} = \arg\max_\pi \hat{U}(\pi)$ *is bounded by*

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

Proof. Both $\hat{\pi}$ and π^* by definition satisfy any provided constraints. Then,

$$\begin{aligned}U(\pi^*) - U(\hat{\pi}) &= U(\pi^*) - \hat{U}(\hat{\pi}) + \hat{U}(\hat{\pi}) - U(\hat{\pi}) \\ &\le U(\pi^*) - \hat{U}(\pi^*) + \hat{U}(\hat{\pi}) - U(\hat{\pi}),\end{aligned} \tag{7}$$

where the second equation follows because $\hat{\pi} = \arg\max_{pi} \hat{U}(\pi)$, and so, $\hat{U}\pi^* \le \hat{U}\hat{\pi}$.

Because the parity part of the utility function depends only on the policy, and not the rewards, it cancels out in Equation (7), leaving

$$\begin{aligned}&U(\pi^*) - \hat{U}(\pi^*) + \hat{U}(\hat{\pi}) - U(\hat{\pi}) \\ &= \sum_x p_x \sum_k \pi^*_{xk}(r_{xk} - \hat{r}_{xk}) + \sum_x p_x \sum_k \hat{\pi}_{xk}(\hat{r}_{xk} - r_{xk}) \\ &\le 2\sum_x p_x \max_k |r_{xk} - \hat{r}_{xk}|. \qquad \text{Q.E.D.}\end{aligned}$$

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

1. (Tabular rewards) We assume $r(x,k) \overset{d}{=} f(x,k) + \eta$, where $\eta \sim \sigma^2$ is sub-Gaussian, 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 $r(x,k) \overset{d}{=} \phi(x,a_k)^T\theta^* + \eta$, where $\eta \sim \sigma^2$ is sub-Gaussian, 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.

Full proofs for this section are in Online Appendix EC.5.

Theorem 1 (Tabular Rewards). *Assume the reward is tabular. Assume n samples are collected in a round-robin fashion (i.e., for each context x, select the least-sampled action* a_k *in that context, breaking ties arbitrarily). Further assume that the data are used per* (x, a) *pair to estimate a maximum-likelihood reward model* $\hat{r}(x,a)$ *that is used to define* $\hat{U}$ *(see Equation (6)) and* $\hat{\pi} = \arg\max \hat{U}$. *Then, for* $\epsilon > 0$, $\delta > 0$, $\lambda_g \ge 0$, *if*

$$n \ge 16\sigma^2 \frac{|X||A|}{\epsilon^2} \ln\frac{4|X||A|}{\delta} \ln\frac{2|X|}{\delta},$$

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

Standard proofs for tabular multiarmed bandits rely on concentration inequalities on the estimated reward functions (Mannor and Tsitsiklis 2004). Unlike this work, we additionally need to estimate the reward per context to ensure the final estimated utility, which is a weighted sum over contexts, is nearly accurate. We show it suffices to estimate the reward outcome for a particular (x,a) pair to differing levels of accuracy, based on the probability of the context x, which allows our final bounds to be independent of the minimum context probability. This result is identical to finding a policy

such that $\sum_x p(x)r(x,\pi(x))$ is ϵ-close to optimal. Note that this sample bound is identical whether we consider spending parity (i.e., regardless of whether $\lambda_g > 0$ for some g or $\lambda_g = 0$ for all g in Equation (1)). Intuitively, this is the case because the sample complexity is driven by uncertainty in estimating the rewards. The parity component itself depends only on the allocation across subgroups, which can be computed exactly given any policy, independent of the estimated rewards.

Our sample bound in the tabular setting scales linearly with the product of the size of the context space and the action space, which suggests that prohibitively large sample sizes may be needed in practice. When the contexts are independent, this dependence is unavoidable, as a sample complexity lower bound shows at least $|A|/\epsilon^2$ samples are required for a single context (Mannor and Tsitsiklis 2004). Our next theorem proves significantly fewer samples are sufficient if the reward function is a linear model.

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

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

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

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

There exists a static (it does not update as data are gathered) data collection strategy $\tilde{\pi}$ such that, for any $\epsilon > 0, \delta > 0$, $\lambda_g \geq 0$ and

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

with cost incurred $c \leq n\max_{xk} c(x,a_k)$, 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 $\rho_0 \leq \|\phi\|_2/\sqrt{\lambda_{\min}(\Sigma)}$. 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 with $1/\epsilon^2$, and so, the upper bound scales like $\sigma^2 d^2/\epsilon^2$. $\tilde{\pi}$ refers to the data collection strategy in Theorem 2, and $\hat{\pi}$ refers to the performance of a learned decision policy that maximizes the utility given the gathered data.

The above result was motivated in part by, as we noted earlier, that, in general, we cannot directly leverage cumulative regret bounds for contextual bandits because the bounds relate empirical decisions to the optimal decision for the current context, with no further constraints or objectives. However, concurrent research by Zanette et al. (2021) on contextual linear bandits provides a sample complexity result sufficient to upper bound the expected reward error,

$$\int_x p_x \max_k |\phi(x,a_k)^T(\theta^* - \hat{\theta})|. \tag{8}$$

From our Lemma 1, we can use this to directly bound our expected utility. Therefore, we could also use their data collection strategy and bound and obtain a $O(d^2/\epsilon^2)$ sample complexity result, which does not depend on $\rho_0(\pi)$. Their sample complexity results is minimax (in the dominant term, up to constants and log terms) optimal for linear contextual bandits (for both static data collection strategies that do not update based on the observed rewards and for adaptive ones that do update as rewards are observed). This implies that using their algorithm also yields a minimax (in the dominant term, up to constants and log terms) optimal sample complexity result for our setting because crucially, the parity objective depends only on the policy.

Given the practical importance of binary rewards, bounds for this setting would also be beneficial. However, whereas there has been some recent attention to logistic bandits (Li et al. 2017, Dong et al. 2019, Jun et al. 2021), these papers have focused on cumulative regret guarantees. Jun et al. (2021) provide some PAC bounds on returning the optimal arm for logistic bandits. We are not aware of sample complexity results for contextual logistic bandits. In Online Appendix EC.5, we provide some preliminary bounds on the suboptimality of the performance of the resource allocation strategy derived from using estimated plug-in parameters for the logistic reward model (Theorem EC.4). Our results require strong assumptions and depend on problem-specific properties and the data collection strategy, suggesting there is significant room for similar results under more relaxed, and constructive, conditions. Contextual multiarmed bandits are an active research area in the machine learning community, and it is likely our results can benefit from future results on sample complexity algorithms and bounds for contextual bandits.

5.2. Cost-Aware Sample Complexity

Whereas we have provided sufficient sample bounds, it is also useful to consider bounds on the experimental cost that are sufficient to learn a near-optimal policy. Note that by this, we mean a bound on the cost required to learn a near-optimal policy, not the budget constraint on the learned decision policy. In general, the amount of resources available during the experimental period may be different than the resources available during sustained deployment.

In the tabular case, we can prove that an experimental budget of $O(\sum_{x,k} c(x,a_k)\log(1/\delta)/\epsilon^2)$ is sufficient (see Corollary EC.1 in the e-companion). When the domain

can be modeled with a linear or logistic reward model, we can simply multiply our sample bounds by the maximum cost $\max_{xk} c(x, a_k)$ to get sufficient upper bounds on the experimental cost. In some settings, these bounds are likely order optimal in the dominant terms. For example, in the tabular case without parity preferences, when costs are homogeneous across contexts and actions and the context distribution is uniform, the expected experimental cost must be at least $c|X||A|\log(1/\delta)/\epsilon^2$ in the worst case (Theorem EC.5 in the e-companion).

However, in general, we expect that there are alternate strategies with tighter bounds that are cost aware. As an illustration, consider a setting with two contexts, two actions, bounded rewards, no parity preferences ($\lambda_g = 0$), and a budget b (for the final learned decision policy) that is very large. As shown in Table 1, let costs be \$1 for both actions in context 1, \$0 for action 1 in context 2, and \$500 for action 2 in context 2. Using a round-robin data collection strategy to obtain an ϵ-optimal policy, which we analyzed previously, will take both actions in both contexts an equal number of times. However, if the probability of context 2 is very small compared with context 1, depending on the reward structure, it may be possible to learn a policy that yields a utility that is ϵ-optimal by only learning the optimal action in context 1 and always taking action 1 in context 2 (the zero cost action). Given the high cost of sampling action 2 in context 2, such an alternate data-gathering strategy might be preferable if it is important to optimize the cost incurred when learning the decision policy.

Generally, we expect that a cost-aware data-gathering strategy would depend on the interaction between context probabilities, cost functions, and bounds on the potential outcome (reward) ranges. This is an interesting direction for future work, and the technical innovations required will likely further increase when we use parametric assumptions on the reward models and when budget or parity preferences ($\lambda_g > 0$) are in place.

6. Adaptively Learning Optimal Policies

The results from Section 5 suggest the feasibility of solving our desired optimization problem even when the distribution of potential outcomes must be estimated from data. However, learning from the type of nonadaptive strategies considered above is typically not the most efficient approach to learning from data. For instance, 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, a nonadaptive learning strategy will still allocate a proportional amount of limited resources to such individuals. In contrast, 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.[4]

Table 1. Setup for Hypothetical Cost-Aware Example

		Costs	
	Probability	Action 1	Action 2
Context 1	0.98	\$1	\$1
Context 2	0.02	\$0	\$500

Algorithm 1 (Policy Learning Procedure)

input: Actions a_k, budget b, parity preferences $\lambda_{g,\ell}$ and f_ℓ, reward function r, covariate distribution $\mathbb{P}(X = x)$, group membership function s, bandit algorithm, n_{init}
initialize: Randomly treat first n_{init} 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 $\mathcal{D}_{k,x}(Y(a_k)|X = x)$ with a parametric family $g(x, a; \theta)$ fit on $\mathcal{D}_i$
  if ε-greedy **then**
    Estimate $\mathbb{E}_Y[r(x, a, Y(a_k))|X = x]$ and $\mathbb{E}_Y[f_\ell(x, a, Y(a_k))|X = x]$ using $g(x, a; \hat{\theta}_i)$, where $\hat{\theta}_i$ is the MLE
  else if Thompson sampling **then**
    Estimate $\mathbb{E}_Y[r(x, a, Y(a_k))|X = x]$ and $\mathbb{E}_Y[f_\ell(x, a, Y(a_k))|X = x]$ using $g(x, a; \hat{\theta}_i^*)$, where $\hat{\theta}_i^*$ is drawn from the posterior of $\hat{\theta}_i$
  else if UCB **then**
    Estimate $\mathbb{E}_Y[r(x, a, Y(a_k))|X = x]$ using the α-percentile of the posterior of $g(x, a; \hat{\theta}_i)$ and estimate $\mathbb{E}_Y[f_\ell(x, a, Y(a_k))|X = x]$ using the $(1 - \alpha)$-percentile of the posterior of $g(x, a; \hat{\theta}_i)$
  end if
  Compute nominal budgets b_i^* according to Equation (EC.42)
  Find solution π_i^* of the LP in Section 4.1 with b_i^* and the input values estimated above
  if ε − greedy & BERNOULLI(ε) == 1 **then**
    Take random action A_i according to Equation (EC.43)
  else
    Take action $A_i \sim \pi_i^*(X_i)$
  end if
  Observe outcome Y_i
end for

To more efficiently learn decision policies in the real world, we now outline our procedure to integrate the LP formulation from Section 4 with three common contextual bandit approaches: ε-greedy, Thompson sampling, and upper confidence bound (UCB), as described in Algorithm 1. 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.

At a high level, at each step i, our ε-greedy approach first estimates $\mathcal{D}(Y(a_k)|X)$ using the maximum-likelihood estimates of a chosen parametric family. We next use these estimates 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 $\mathcal{D}(Y(a_k)|X)$, samples from this posterior; uses the posterior draw to construct the inputs for our LP, yielding 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 optimistic estimates of r and the parity penalties (e.g., using the 97.5th percentile of the posterior of the former and the 2.5th percentile of the latter).

6.1. Simulation Study

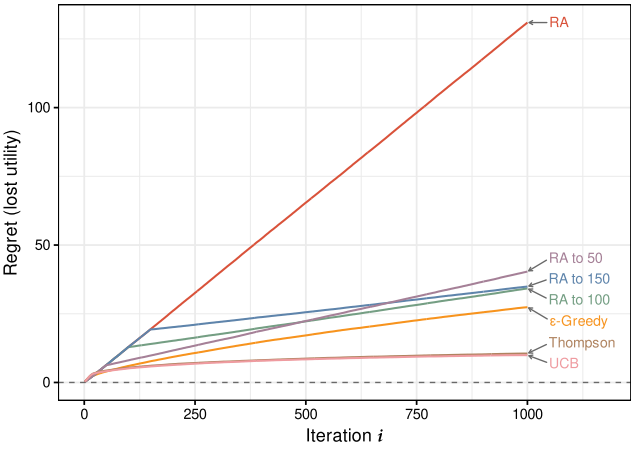
To evaluate our learning approach above, we conducted a simulation study using data from a sample of clients served by the Santa Clara County Public Defender Office. In this example, clients can receive one of three mutually exclusive treatments a_k: round-trip rideshare assistance, a transit voucher, or no transportation assistance. We fix our budget to \$5,000 and limit our population to 1,000 clients, resulting in an average per-person budget of \$5. In line with many government pilot programs, we assume that this funding is dedicated to learning suitable policies and that our hypothetical public defender would be able to provision separate funding later to operate a more permanent program. We set the cost of rides to \$5 per mile. We limit the client population to White and Vietnamese individuals to reflect our running example. The utility of a policy is described by Equation (1), where we set $r(x,a,y)=y, f(x,a,y)=c(x,a)$, and $\lambda=0.0006$. This choice yields an oracle policy that balances between maximizing appearances and achieving parity in per-capita spending across groups. The data-generating process for this population and additional experiment parameters are described in detail in Online Appendix EC.6.

We compare our contextual bandit approaches against several baselines. First, we compare with nonadaptive random assignment (RA) in which treatment is randomly selected (in accordance with the budget) throughout the entire learning phase. The simplicity and versatility of RA make it a common strategy for learning optimal policies. We also include partially adaptive variations on this approach, where we run RA 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 appearance probabilities.

We repeat this evaluation 2,000 times each on 1,000 randomly selected individuals from our data set and compare the performance of all approaches using two different metrics. Our two main bandit approaches—Thompson sampling and UCB—significantly reduce regret when compared with nonadaptive and partially adaptive approaches during the learning phase (Figure 4). Our bandit approaches also learn policies that, if used for future populations, would outperform nonadaptive and partially adaptive approaches (Figure 5). 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. RA and its variations illustrate the limits of the conventional randomized approach. For example, it is possible to learn a near-oracle policy using classic RA, but this incurs substantial regret during the learning phase. Though it is possible to reduce this regret by ending RA early, these alternatives learn poorer-performing policies.

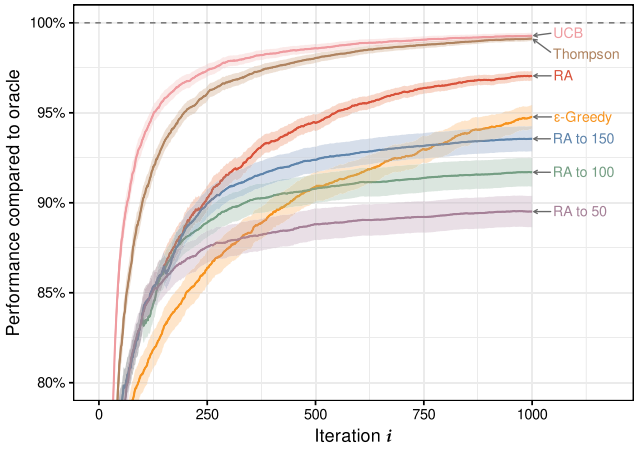
By design, the bandit methods discussed above reduce spending disparities during the course of the simulation. We demonstrate this by comparing our main simulation to an alternate set of simulations where $\lambda_g=0$ (Table 2). For example, with a choice of $\lambda_g=0.004$, reflecting a mild preference for more equal spending, we observe that UCB methods spent \$2.21 less on Vietnamese clients than the \$5 population

Figure 4. (Color online) Mean Regret, Across 2,000 Simulations, Incurred by Different Learning Approaches



Notes. We define regret here as the difference between the observed utility and the utility obtained by an oracle during the same experiment. Values are tightly estimated at each i, with the 95% confidence interval no more than 1.1 units off the estimate, so we omit uncertainty bands for this figure. We note that the three bandit approaches—ε-greedy, Thompson sampling, and UCB—incur substantially less regret than random assignment (RA). It is possible to reduce the regret incurred from RA by stopping randomization early and following the optimal estimated policy from that point forward. However, these stop-early RA approaches produce worse policies than other approaches (Figure 5).

Figure 5. (Color online) Mean Performance, Across 2,000 Simulations, of Optimal Policies Estimated with Data Available at Each Iteration i



Notes. 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 random assignment (RA) at any given iteration i. In contrast, the ε-greedy approach and the stop-early versions of RA generate policies that are slower to (or may never) reach near-oracle performance.

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 with the population average.

The bandit approaches we discuss above aim to maximize utility during the learning phase but do not explicitly try to minimize money spent during learning. As discussed in Section 5.2, it is possible that alternate approaches may spend less while achieving similar outcomes. One could imagine a learning strategy in which nearly all participants were offered the no-cost treatment, with only a small number offered a costly treatment (a ride or transit voucher). With these alternate approaches, we may be able to learn the structure of appearance behavior by using the no-cost treatment for most participants and then learn the impact of costly treatments with a small number of remaining participants. To explore such alternate approaches empirically, we considered a range of policies that assign the first 1,000 clients in each experiment to one of our three treatment arms in different random allocations. For example, one variation randomly allocated rides to only 2% of clients and transit vouchers to only 2% of clients, with the remaining 96% of clients receiving the no-cost control action. Another variation randomly allocated rides for 10% of clients and vouchers for 40% of clients, with the remaining 50% of clients receiving the no-cost control action. Additional details describing this simulation are included at the end of Online Appendix EC.6.

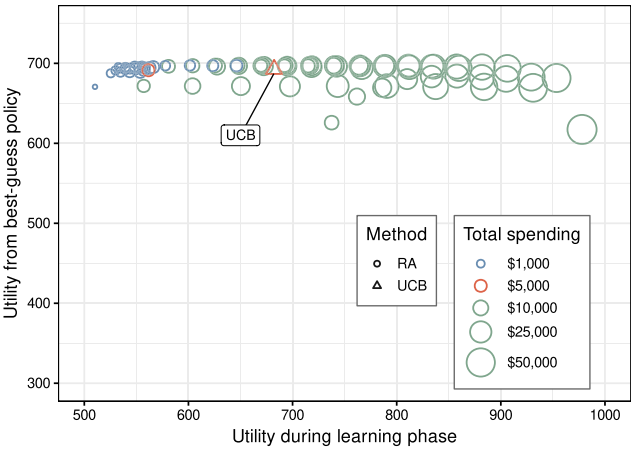
We show the results of this exercise in Figure 6. This plot compares three dimensions on which we evaluate each policy: first, the utility achieved during the learning phase; second, the quality of the policy learned by the end of the phase; and third, the total amount spent during learning. We see that varying spending mostly affects the utility observed during learning, with more expensive allocations resulting in higher utility during learning. Spending appears to have little impact on the quality of the final policy learned. For the sake of

Table 2. 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 $\lambda_g = 0$)

	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

Notes. 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.

Figure 6. (Color online) The Effect of Varying Spending on Outcomes of Interest



Notes. Each circle represents average outcomes across 125 simulations of random assignment (RA) with a given allocation. For the sake of comparison, we also included the average outcome across 2,000 simulations of UCB from earlier in this section, represented here by a single triangle. Each glyph is sized and color coded by total spending, with color indicating if these approaches spent less (blue), equivalent (red), or more (green) money compared with the methods discussed earlier in this section. We find that varying spending mostly affects the utility observed during the learning phase but has little effect on the quality of the final policy learned. Random assignment strategies that save money (when compared with UCB) do not achieve as much utility during the learning phase, though both approaches result in similar-quality policies.

comparison, Figure 6 also includes UCB results from the simulations at the beginning of this section. Among the spending variations we tested, no variation spent less money than UCB while achieving similar utility during the learning phase. This suggests that UCB can be a cost-effective approach to maximizing utility while learning a high-quality policy.

7. 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 equalizing spending across groups. Popular rule-based approaches to algorithmic fairness—such as enforcing spending parity or 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 literature. 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 random assignment or acting greedily based on the available information.

Our learning algorithm requires access to a well-specified utility function that reflects stakeholder preferences. In practice, inferring this utility is a complex task in its own right, but the illustrative survey that we conducted shows how one can begin to operationalize this task. 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 multifaceted preferences less directly. One family of approaches selects pairs of similar realistic scenarios, asks stakeholders to pick their preferred outcome, and infers their preferences from these choices (Chu and Ghahramani 2005, Fürnkranz and Hüllermeier 2010, Jung et al. 2019, Lin et al. 2020, Koenecke et al. 2023).

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. Thompson sampling methods have been observed to be more robust to delayed outcomes than upper confidence bound strategies in contextual bandit scenarios (Chapelle and Li 2011). Another 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. 2016). For example, with rideshare assistance to clients, one might use intermediate responses (like a client's confirmation to attend the appointment) as a proxy for appearance. A third approach might be 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 aid the entire population of interest. If policymakers had enough funding available to assist an entire population, it may not make sense to even consider equalizing 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 (Fishbane et al. 2020, Chohlas-Wood et al. 2023b, Zottola et al. 2023).

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.

Acknowledgments

The authors thank Johann Gaebler, Jonathan Lee, Hamed Nilforoshan, Julian Nyarko, and Ariel Procaccia for helpful comments. The authors 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.

Endnotes

[1] Optimizing for parity across protected demographic groups, including race groups, is legally impermissible in some contexts in the United States, as we discuss more in Section 7.

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

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

[4] Nonadaptive strategies are particularly useful when testing statistical hypotheses post hoc, which is most easily done with data that are independently and identically distributed across treatments. We note that there is considerable interest in developing suitable inference methods for this latter goal using data gathered with adaptive multi-armed bandit strategies (e.g., Hadad et al. 2021, Zhang et al. 2021).

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, eds. *NIPS'11 Proc. 24th Internat. Neural Inform. Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY), 2312–2320.

Agrawal S, Devanur NR, Li L (2016b) An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. Feldman V, Rakhlin A, Shamir O, eds. *29th Annual Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 49 (PMLR, New York), 4–18.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2016a) A near-optimal exploration-exploitation approach for assortment selection. *Proc. 2016 ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 599–600.

Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A, Rieke A (2019) Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proc. ACM Human-Comput. Interaction (CSCW)*, vol. 3 (Association for Computing Machinery, New York), 1–30.

Allen S (2024) Interview research with people in jail: Challenges and possibilities. *Handbook on Prisons and Jails* (Routledge, London), 387–398.

Athey S, Chetty R, Imbens GW, Kang H (2016) Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate Index. Preprint, submitted March 30, https://arxiv.org/abs/1603.09326.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine Learn.* 47(2):235–256.

Badanidiyuru A, Langford J, Slivkins A (2014) Resourceful contextual bandits. Balcan MF, Feldman V, Szepesvári C, eds. *Proc. 27th Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 35 (PMLR, New York), 1109–1134.

Barabas C, Virza M, Dinakar K, Ito J, Zittrain J (2018) Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. Friedler SA, Wilson C, eds. *Proc. 1st Conf. Fairness, Accountability Transparency*, Proceedings of Machine Learning Research, vol. 81 (PMLR, New York), 62–76.

Barocas S, Hardt M, Narayanan A (2023) *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, Cambridge, MA).

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Oper. Res.* 59(1):17–31.

Blodgett SL, O'Connor B (2017) Racial disparity in natural language processing: A case study of social media African-American English. Preprint, submitted June 30, https://arxiv.org/abs/1707.00061.

Brams SJ, Brams SJ, Taylor AD (1996) *Fair Division: From Cake-Cutting to Dispute Resolution* (Cambridge University Press, Cambridge, UK).

Brough R, Freedman M, Ho DE, Phillips DC (2022) Can transportation subsidies reduce failures to appear in criminal court? Evidence from a pilot randomized controlled trial. *Econom. Lett.* 216:110540.

Brown A, Chouldechova A, Putnam-Hornstein E, Tobin A, Vaithianathan R (2019) Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. *Proc. 2019 CHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 1–12.

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. Friedler SA, Wilson C, eds. *Proc. 1st Conf. Fairness, Accountability Transparency*, Proceedings of Machine Learning Research, vol. 81 (PMLR, New York).

Cadigan TP, Lowenkamp CT (2011) Implementing risk assessment in the federal pretrial services system. *Federal Probation* 75(2): 30–34.

Cai W, Gaebler J, Garg N, Goel S (2020) Fair allocation through selective information acquisition. *Proc. AAAI/ACM Conf. AI Ethics Soc.* (Association for Computing Machinery, New York), 22–28.

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

Caragiannis I, Kaklamanis C, Kanellopoulos P, Kyropoulou M (2012) The efficiency of fair division. *Theory Comput. Systems* 50(4): 589–610.

Card D, Smith NA (2020) On consequentialism and fairness. *Frontiers Artificial Intelligence* 3:34.

Chaiyachati KH, Hubbard RA, Yeager A, Mugo B, Shea JA, Rosin R, Grande D (2018) Rideshare-based medical transportation for Medicaid patients and primary care show rates: A difference-in-difference analysis of a pilot program. *J. General Internal Medicine* 33(6):863–868.

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, eds. *NIPS'11 Proc. 24th Internat. Neural Inform. Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY).

Chiappa S, Isaac WS (2018) A causal Bayesian networks viewpoint on fairness. Kosta E, Pierson J, Slamanig D, Fischer-Hübner S, Krenn S, eds. *Privacy and Identity Management. Fairness Accountability Transparency in the Age of Big Data. Privacy and Identity 2018*, IFIP Advances in Information and Communication Technology, vol. 547 (Springer, Cham, Switzerland), 3–20.

Chohlas-Wood A, Coots M, Goel S, Nyarko J (2023a) Designing equitable algorithms. *Nature Comput. Sci.* 3(7):601–610.

Chohlas-Wood A, Coots M, Nudell J, Nyarko J, Brunskill E, Rogers T, Goel S (2023b) Automated reminders reduce incarceration for missed court dates: Evidence from a text message experiment. Preprint, submitted June 21, https://arxiv.org/abs/2306.12389.

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. *Comm. ACM* 63(5):82–89.

Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. Friedler SA, Wilson C, eds. *Proc. 1st Conf. Fairness, Accountability Transparency*, Proceedings of Machine Learning Research, vol. 81 (PMLR, New York), 134–148.

Chu W, Ghahramani Z (2005) Preference learning with Gaussian processes. *Proc. 22nd Internat. Conf. Machine Learning* (Association for Computing Machinery, New York), 137–144.

Cohler YJ, Lai JK, Parkes DC, Procaccia AD (2011) Optimal envy-free cake cutting. *25th AAAI Conf. Artificial Intelligence* (AAAI Press, Washington, DC), 626–631.

Corbett-Davies S, Gaebler J, Nilforoshan H, Shroff R, Goel S (2023) The measure and mismeasure of fairness. *J. Machine Learning Res.* 24(1):14730–14846.

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 797–806.

Coston A, Mishler A, Kennedy EH, Chouldechova A (2020) Counterfactual risk assessments, evaluation, and fairness. *Proc. 2020 Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 582–593.

Cowgill B, Tucker CE (2020) Algorithmic fairness and economics. Preprint, submitted April 4, 2019, http://dx.doi.org/10.2139/ssrn.3361280.

Datta A, Datta A, Makagon J, Mulligan DK, Tschantz MC (2018) Discrimination in online advertising: A multidisciplinary inquiry. Friedler SA, Wilson C, eds. *Proc. 1st Conf. Fairness, Accountability Transparency*, Proceedings of Machine Learning Research, vol. 81 (PMLR, New York), 20–34.

De-Arteaga M, Fogliato R, Chouldechova A (2020) A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proc. 2020 CHI Conf. Human Factors Comput. Systems* (Association for Computing Machinery, New York), 1–12.

De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai AT (2019) Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proc. Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 120–128.

Donahue K, Kleinberg J (2020) Fairness and utilization in allocating resources with uncertain demand. *Proc. 2020 Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 658–668.

Dong S, Ma T, Van Roy B (2019) On the performance of Thompson sampling on logistic bandits. Beygelzimer A, Hsu D, eds. *Proc. Thirty-Second Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 99 (PMLR, New York), 1158–1160.

Fang EX, Wang Z, Wang L (2022) Fairness-oriented learning for optimal individualized treatment rules. *J. Amer. Statist. Assoc.* 118(543):1733–1746.

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. *Proc. 21th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 259–268.

Fishbane A, Ouss A, Shah AK (2020) Behavioral nudges reduce failure to appear for court. *Science* 370(6517):eabb6591.

Fraade-Blanar L, Koo T, Whaley CM (2021) Going to the doctor: Rideshare as nonemergency medical transportation. Technical report, RAND Corporation, Santa Monica, CA.

Friedewald JJ, Samana CJ, Kasiske BL, Israni AK, Stewart D, Cherikh W, Formica RN (2013) The kidney allocation system. *Surgical Clinics* 93(6):1395–1406.

Fürnkranz J, Hüllermeier E (2010) Preference learning and ranking by pairwise comparison. Fürnkranz J, Hüllermeier E, eds. *Preference Learning* (Springer, Berlin, Heidelberg), 65–82.

Gal Y, Mash M, Procaccia AD, Zick Y (2017) Which is the fairest (rent division) of them all? *J. ACM* 64(6):1–22.

Goel S, Shroff R, Skeem JL, Slobogin C (2018) The accuracy, equity, and jurisprudence of criminal risk assessment. Vogl R, ed. *Research Handbook on Big Data Law* (Edward Elgar Publishing, Cheltenham, UK), 9–28.

Goodman SN, Goel S, Cullen MR (2018) Machine learning, health disparities, and causal reasoning. *Ann. Intern. Med.* 169(12):883–884.

Grgić-Hlača N, Lima G, Weller A, Redmiles EM (2022) Dimensions of diversity in human perceptions of algorithmic fairness. *EAAMO '22 Proc. 2nd ACM Conf. Equity Access Algorithms Mechanisms Optim.* (Association for Computing Machinery, New York), 1–12.

Gupta S, Jalan A, Ranade G, Yang H, Zhuang S (2020) Too many fairness metrics: Is there a solution? Preprint, submitted April 9, http://dx.doi.org/10.2139/ssrn.3554829.

Hadad V, Hirshberg DA, Zhan R, Wager S, Athey S (2021) Confidence intervals for policy evaluation in adaptive experiments. *Proc. Natl. Acad. Sci. USA* 118(15):e2014602118.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *NIPS'16 Proc. 30th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY).

Hsu D, Kakade SM, Zhang T (2014) Random design analysis of ridge regression. *Foundations Comput. Math.* 14:569–600.

Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? *NIPS'18 Proc. 32nd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 3323–3331.

Jun KS, Jain L, Mason B, Nassif H (2021) Improved confidence bounds for the linear logistic model and applications to bandits. Meila M, Zhang T, eds. *Proc. 38th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 139 (PMLR, New York), 5148–5157.

Jung C, Kearns M, Neel S, Roth A, Stapleton L, Wu ZS (2019) An algorithmic framework for fairness elicitation. Preprint, submitted May 25, https://arxiv.org/abs/1905.10660.

Kasy M, Abebe R (2021) Fairness, equality, and power in algorithmic decision-making. *Proc. 2021 ACM Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 576–586.

Kilbertus N, Rojas Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. *NIPS'17 Proc. 31st Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 656–666.

Koenecke A, Giannella E, Willer R, Goel S (2023) Popular support for balancing equity and efficiency in resource allocation: A case study in online advertising to increase welfare program awareness. *Proc. Internat. AAAI Conf. Web Social Media*, vol. 17, (AAAI Press, Washington, DC), 494–506.

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S (2020) Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. USA* 117(14):7684–7689.

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. *NIPS'17 Proc. 31st Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 4069–4079.

Latessa EJ, Lemke R, Makarios M, Smith P (2010) The creation and validation of the Ohio risk assessment system (ORAS) *Federal Probation* 74(1):16–22.

Lattimore T, Szepesvári C (2020) *Bandit Algorithms* (Cambridge University Press, Cambridge, UK).

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

Li L, Lu Y, Zhou D (2017) Provably optimal algorithms for generalized linear contextual bandits. Precup D, Teh YW, eds. *Proc. 34th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 70 (PMLR, New York), 2071–2080.

Li Z, Ratliff L, Nassif H, Jamieson KG, Jain L (2022) Instance-optimal PAC algorithms for contextual bandits. *NIPS'22 Proc. 36th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 37590–37603.

Lin ZJ, Obeng A, Bakshy E (2020) Preference learning for real-world multi-objective decision making. Bogunovic I, Neiswanger W, Yue Y, eds. *Workshop Real World Experiment Design Active Learn.* (ICML).

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. Dy J, Krause A, eds. *Proc. 35th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 80 (PMLR, New York), 3150–3158.

Lyft (2020) Modernizing medical transportation with rideshare. Technical report, FierceHealthcare, Washington, DC.

Mahoney B, Beaudin BD, Carver JA III, Ryan DB, Hoffman RB (2001) *Pretrial Services Programs: Responsibilities and Potential* (National Institute of Justice, Washington, DC).

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. *J. Machine Learn. Res.* 5:623–648.

Metevier B, Giguere S, Brockman S, Kobren A, Brun Y, Brunskill E, Thomas P (2019) Offline contextual bandits with high probability fairness guarantees. *NIPS'19 Proc. 33rd Internat. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 14922–14933.

Milgram A, Holsinger AM, Vannostrand M, Alsdorf MW (2014) Pretrial risk assessment: Improving public safety and fairness in pretrial decision making. *Federal Sentencing Reporter* 27(4): 216–221.

Nabi R, Shpitser I (2018) Fair inference on outcomes. *Proc. AAAI Conf. Artificial Intelligence*, vol. 32 (AAAI Press, Washington, DC), 1931–1940.

Nilforoshan H, Gaebler JD, Shroff R, Goel S (2022) Causal conceptions of fairness and their consequences. Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, eds. *Proc. 39th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 162 (PMLR, New York), 16848–16887.

Nyarko J, Goel S, Sommers R (2021) Breaking taboos in fair machine learning: An experimental study. *EAAMO '21 Proc. 1st ACM Conf. Equity Access Algorithms Mechanisms Optim.* (Association for Computing Machinery, New York), 1–11.

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

Pacchiano A, Lee J, Brunskill E (2023) Experiment planning with function approximation. *NIPS'23 Proc. 37th Internat. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 9409–9421.

Patil V, Ghalme G, Nair V, Narahari Y (2021) Achieving fairness in the stochastic multi-armed bandit problem. *J. Machine Learn. Res.* 22:1–31.

Procaccia AD (2013) Cake cutting: Not just child's play. *Comm. ACM* 56(7):78–87.

Raji ID, Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proc. 2019 AAAI/ACM Conf. AI Ethics Soc.* (Association for Computing Machinery, New York), 429–435.

Rolf E, Simchowitz M, Dean S, Liu LT, Bjorkegren D, Hardt M, Blumenstock J (2020) Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. Daumé H III, Singh A, eds. *Proc. 37th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 119 (PMLR, New York), 8158–8168.

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

Slivkins A (2019) *Introduction to Multi-Armed Bandits*, Foundations and Trends in Machine Learning, vol. 12 (now Publishers, Hanover, MA), 1–286.

Thomas PS, da Silva BC, Barto AG, Giguere S, Brun Y, Brunskill E (2019) Preventing undesirable behavior of intelligent machines. *Science* 366(6468):999–1004.

Vais S, Siu J, Maru S, Abbott J, Hill IS, Achilike C, Wu WJ, Adegoke TM, Steer-Massaro C (2020) Rides for refugees: A transportation assistance pilot for women's health. *J. Immigrant Minority Health* 22(1):74–81.

Viviano D, Bradic J (2024) Fair policy targeting. *J. Amer. Statist. Assoc.* 119(545):730–743.

Wang Y, Sridhar D, Blei DM (2019) Equal opportunity and affirmative action via counterfactual predictions. Preprint, submitted May 26, https://arxiv.org/abs/1905.10870.

Wilder B, Onasch-Vera L, Diguiseppi G, Petering R, Hill C, Yadav A, Rice E, Tambe M (2021) Clinical trial of an AI-augmented intervention for HIV prevention in youth experiencing homelessness. *Proc. AAAI Conf. Artificial Intelligence*, vol. 35 (AAAI Press, Washington, DC), 14948–14956.

Wu H, Srikant R, Liu X, Jiang C (2015) Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *NIPS'15 Proc. 28th Internat. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 433–441.

Wu Y, Zhang L, Wu X, Tong H (2019) PC-fairness: A unified framework for measuring causality-based fairness. *Proc. 33rd Internat. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 3404–3414.

Zanette A, Dong K, Lee JN, Brunskill E (2021) Design of experiments for stochastic contextual linear bandits. *NIPS'21 Proc. 35th Internat. Neural Inform. Processing Systems*, vol. 34 (Curran Associates, Red Hook, NY), 22720–22731.

Zhang J, Bareinboim E (2018) Fairness in decision-making—The causal explanation formula. *Proc. AAAI Conf. Artificial Intelligence*, vol. 32 (AAAI Press, Washington, DC), 2037–2045.

Zhang K, Janson L, Murphy S (2021) Statistical inference with m-estimators on adaptively collected data. *NIPS'21 Proc. 35th Internat. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 7460–7471.

Zottola SA, Crozier WE, Ariturk D, Desmarais SL (2023) Court date reminders reduce court nonappearance: A meta-analysis. *Criminology Public Policy* 22(1):97–123.

Zuluaga M, Sergent G, Krause A, Püschel M (2013) Active learning for multi-objective optimization. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 28 (PMLR, New York).