Hiring as Exploration*

Danielle Li
MIT & NBER

Lindsey Raymond
MIT

Peter Bergman
Columbia & NBER

July 15, 2020

Abstract

In looking for the best workers over time, firms must balance "exploitation" (selecting from groups with proven track records) with "exploration" (selecting from under-represented groups to learn about quality). Yet modern hiring algorithms, based on "supervised learning" approaches, are designed solely for exploitation. In this paper, we view hiring as a contextual bandit problem and build a resume screening algorithm that values exploration by evaluating candidates according to their statistical upside potential. Using data from professional services recruiting within a Fortune 500 firm, we show that this approach improves the quality (as measured by eventual hiring rates) of candidates selected for an interview, while also increasing demographic diversity, relative to the firm's existing practices. The same is not true for traditional supervised learning based algorithms, which improve hiring rates but select far fewer Black and Hispanic applicants. In an extension, we show that exploration-based algorithms are also able to learn more effectively about simulated changes in applicant quality over time. Together, our results highlight the importance of incorporating exploration in developing decision-making algorithms that are potentially both more efficient and equitable.

JEL Classifications: D80, J20, M15, M51

Keywords: Machine Learning, Hiring, Supervised Learning, Bandit Problems, Algorithmic Fairness, Diversity.

---

*Correspondence to d_li@mit.edu, lraymond@mit.edu, and bergman@tc.columbia.edu. We are grateful to David Autor, Pierre Azoulay, Dan Bjorkegren, Emma Brunskill, Eleanor Dillon, Alex Frankel, Bob Gibbons, Nathan Hendren, Max Kasy, Fiona Murray, Anja Sautmann, Scott Stern, and various seminar participants, for helpful comments and suggestions. The content is solely the responsibility of the authors and does not necessarily represent the official views of Columbia University, MIT, or the NBER.

Algorithms have been shown to outperform human decision-makers across an expanding range of settings, from medical diagnosis to image recognition to game play.[1] However, the rise of algorithms is not without its critics, who caution that automated approaches may codify existing human biases and allocate fewer resources to those from under-represented groups.[2]

A key emerging application of machine learning (ML) tools is hiring, where decisions matter for both firm productivity and individual access to opportunity, and where algorithms are increasingly used to screen job applicants.[3] Modern hiring ML typically relies on "supervised learning," meaning that it models the relationship between applicant covariates and outcomes in a given training dataset, and then applies its model to predict outcomes for subsequent applicants.[4] By systematically analyzing historical examples, these tools can unearth predictive relationships that may be overlooked by human recruiters; indeed, a growing literature has shown that supervised learning algorithms can more effectively identify high quality job candidates than human recruiters.[5] Yet because this approach implicitly assumes that past examples extend to future applicants, firms that rely on supervised learning will tend to select from groups with proven track records rather than taking risks on non-traditional applicants, raising concerns about access to opportunity.

This paper is the first to develop and evaluate a new class of hiring algorithms, one that explicitly values exploration. Our approach begins with the idea that the hiring process can be thought of as a contextual bandit problem: in looking for the best applicants over time, a firm must balance exploitation with exploration as it seeks to learn the predictive relationship between applicant covariates (the "context") and applicant quality (the "reward"). Whereas the optimal solution to bandit problems is widely known to incorporate some exploration, supervised learning based algorithms engage in only in exploitation because they are designed to solve static prediction problems. By contrast, ML tools that incorporate exploration are designed to solve dynamic prediction problems that involve learning from sequential actions: in the case of hiring, these algorithms value exploration because this allows for more optimal choices in the future.

Incorporating exploration into hiring ML may also shift the demographic composition of selected applicants. While "exploration" in the bandit sense—that is, selecting candidates with whatever covariates there is more uncertainty over—need not be the same as favoring demographic diversity,

[1] For example, see McKinney et al. (n.d.); Yala et al. (2019); Mullainathan and Obermeyer (2019); Schrittwieser et al. (2019); Russakovsky et al. (2015)

[2] See Obermeyer et al. (2019); Datta et al. (2015); Lambrecht and Tucker (2019). For additional surveys of algorithmic fairness, see Barocas and Selbst (2016); Corbett-Davies and Goel (2018); Cowgill and Tucker (2019).

[3] A recent survey of technology companies indicated that 60% plan on investing in AI-powered recruiting software in 2018, and over 75% of recruiters believe that artificial intelligence will transform hiring practices (Bogen and Rieke, 2018).

[4] For a survey of commercially available hiring ML tools, see Raghavan et al. (2019).

[5] See, for instance, Hoffman et al. (2018); Cowgill (2018).

it is also the case that Black, Hispanic, and female applicants are less likely to be employed in high-income jobs, meaning that they will also appear less often in the historical datasets used to train hiring algorithms. Because data under-representation tends to increase uncertainty, adopting bandit algorithms that value exploration (for the sake of learning) may expand representation even when demographic diversity is not part of their mandate.

We focus on the decision to grant first-round interviews for high-skill positions in consulting, financial analysis, and data science—sectors which offer well-paid jobs with opportunities for career mobility and which have also been criticized for their lack of diversity. Our data come from administrative records on job applications to these types of professional services positions within a Fortune 500 firm. Like many other firms in its sector, this firm is overwhelmed with applications and rejects the vast majority of candidates on the basis of an initial resume screen. Yet, among those who pass this screen and go on to be interviewed, hiring rates are still relatively low: in our case, only 10% receive and accept an offer. Because recruiting is costly and diverts employees from other productive work (Kuhn and Yu, 2019), the firm would like to adopt screening tools that improve its ability to identify applicants it may actually hire. We therefore define an applicant's quality as her "hiring potential," or likelihood of being hired conditional on being interviewed.

We build three resume screening algorithms—two based on supervised learning, and one based on a contextual bandit approach—and evaluate the candidates that each algorithm selects relative to each other and relative to the actual interview decisions made by human recruiters (resume screeners) in the firm. We observe data on an applicant's demographics (race, gender, and ethnicity), education (institution and degree), and work history (prior firms). Each algorithm is trained to predict an applicant's likelihood of being hired if interviewed, given the covariates we observe. Although we will evaluate the diversity of applicants selected by these algorithms, we do not incorporate any explicit diversity preferences into their design.

Our first algorithm uses a static supervised learning approach (hereafter, "static SL") based on an ensemble LASSO and random forest model.[6] Our second algorithm (hereafter, "updating SL") uses the same model as the static SL model, but updates the training data it uses throughout the test period with the hiring outcomes of the applicants it chooses to interview.[7] While this updating

[6]Training only on data from interviewed applicants may lead to biased predictions because of selection on unobservables. While we believe that there is relatively little scope for selection on unobservables in our setting (because we observe essentially the same information as recruiters, who conduct resume reviews without interacting with candidates), we acknowledge the potential for such bias. That said, we are not aware of any commercially-available hiring AI that does attempt to correct for sample selection. Raghavan et al. (2019), for example, surveys the methods of commercially available hiring tools and finds that the vast majority of products marketed as "artificial intelligence" do not use any ML tools at all, and that the few that do simply predict performance using a static training dataset.

[7]In practice, we can only update the model with data from selected applicants who are actually interviewed (otherwise we would not observe their hiring outcome). See Section 3.2 for a more detailed discussion of how this algorithm is updated.

process allows the updating SL model to learn about the quality of the applicants it selects, it is myopic in the sense that it does not incorporate the value of this learning into its selection decisions.

Our third approach implements an Upper Confidence Bound (hereafter, "UCB") contextual bandit algorithm: in contrast to the static and updating SL algorithms, which evaluates candidates based on their *point estimates* of hiring potential, a UCB contextual bandit selects applicants based on the upper bound of the *confidence interval* associated with those point estimates. That is, there is implicitly an "exploration bonus" that is increasing in the algorithm's degree of uncertainty about quality. Exploration bonuses will tend to be higher for groups of candidates who are under-represented in the algorithm's training data because the model will have less precise estimates for these rarer groups. In our implementation, we allow the algorithm to define "rarity" based on a wide set of applicant covariates: the algorithm can choose to assign higher exploration bonuses on the basis of race or gender, but it is not required to and could choose, instead, to focus on other variables such as education or work history. Once candidates are selected, we incorporate their realized hiring outcomes into the training data and update the algorithm for the next period.[8] Contextual bandit UCB algorithms have been shown to be optimal in the sense that they asymptotically minimize expected regret (Lai and Robbins, 1985; Abbasi-Yadkori et al., 2019; Li et al., 2017). We follow the approach in Li et al. (2017) that extends the contextual bandit UCB for binary outcomes.

We have two main sets of results. First, our SL and UCB models differ markedly in the demographic composition of the applicants they select. Implementing a UCB model would more than double the share of selected applicants who are Black or Hispanic, from 10% to 23%. The static and updating SL models, however, would both dramatically decrease Black and Hispanic representation, to approximately 2% and 5%, respectively. In the case of gender, all algorithms would increase the share of selected applicants who are women, from 35% under human recruiting, to 41%, 50%, and 39%, under static SL, updating SL, and UCB, respectively. Although there are fewer women in our data, increases in female representation under UCB are blunted because men tend to be more heterogeneous on other dimensions—geography, education, and race, for instance—leading them to receive higher exploration bonuses on average. We also show that this increase in diversity is persistent during our test sample; if the additional minority applicants selected by the UCB algorithm were truly weaker, the model would update and learn to select fewer such applicants over time. Instead, we show that the UCB model continues to select more minority applicants relative to both the human and SL models, even as exploration bonuses fall.

[8] Similar to the updating SL approach, we only observe hiring outcomes for applicants who are actually interviewed in practice, we are only able to update the UCB model's training data with outcomes for the applicants it selects who are also interviewed in practice. See Section 3.2 for more discussion.

Our second set of results shows that, despite the differences in the demographics of the candidates that they select, most of our ML models generate substantial and comparable increases in the quality of selected applicants, as measured by their hiring potential. Assessing quality differences between human and ML models is more difficult than assessing diversity because we face a "selective labels" problem (Lakkaraju et al., 2017; Kleinberg et al., 2018a): we do not observe hiring outcomes for applicants who are not interviewed.[9] To address this, we take three complementary approaches, all of which consistently show that ML models select higher quality candidates than human recruiters.

First, we focus on the sample of interviewed candidates for whom we directly observe hiring outcomes. Within this sample, we ask whether applicants preferred by our ML models have a higher likelihood of being hired than applicants preferred by a human recruiter. In order to differentiate recruiter preferences among applicants who are all interviewed, we train a fourth algorithm (a supervised learning model similar to our static SL) to predict human interview decisions rather than hiring likelihood (hereafter, "human SL"). We then correlate algorithm scores with actual hiring outcomes within this set. While scores and hiring outcomes are positively correlated for all ML models, human scores and hiring outcomes are weakly if not negatively related.

One key concern with this approach is that human recruiters may be good at making sure that particularly low quality applicants are never interviewed to begin with. Restricting our analysis to the set of actually interviewed applicants may therefore overstate the relative accuracy of our ML models. Additionally, our human SL model may not perfectly predict actual human interview decisions and may, in fact, be worse. To address both of these concerns, our next approach estimates hiring quality for the full sample of applicants and compares it to actually observed hiring outcomes from human interview decisions. Specifically, we follow DiNardo et al. (1996)'s decomposition approach to recover the mean hiring likelihood among all applicants selected by our ML models. That is, suppose an applicant is selected by an ML model but is not selected by the human and therefore never interviewed. We assign this applicant the average observed hiring outcome among actually interviewed candidates in the same race-gender-and education cell. We then aggregate these estimates across all candidates selected by the ML model, and compare it to actual hiring outcomes among those selected by the human. When we do this, we find that ML approaches select applicants with substantially higher predicted quality: average hiring rates among those selected by the UCB and updating SL models are 25% and 30%, respectively, compared with the observed 10% among observed recruiter decisions. Our static SL model also outperforms human decision-making, with a 15% predicted hiring yield. These results suggest that algorithms are better at selecting

[9] Our diversity results are not subject to these concerns because we observe demographics regardless of whether or not an applicant is interviewed.

candidates who are more likely to receive and accept an offer; using these algorithms, the firm could hire the same number of people while conducting fewer interviews.

This approach assumes that there is no selection on unobservables. In our setting, we believe this is a largely reasonable assumption because interview decisions are made on the basis of resume review only: recruiters never meet or otherwise interact with applicants prior to making a decision, nor does the firm use cover letters for these positions. However, one may be concerned that our covariate cells are too coarse or that there are other variables that recruiters observe (an applicant's programming skills for instance) that we have not coded into our covariate set. Both of these issues can potentially generate biases arising from selection on unobservables.

Our final approach performs an alternative analysis that allows for selection on unobservables. Rather than comparing pure ML and human based interview policies, the key to this approach is to ask whether firms can improve on their current interview choices by following ML recommendations for applicants recruiters are indifferent between interviewing or not. Following Benson et al. (2019); Abadie (2003a); Angrist et al. (1996), we use the random assignment of job applicants to recruiters to identify a group of marginally interviewed applicants (those who are instrument compliers in the sense that they are only interviewed because they were assigned to a lax screener). We then compare the quality and demographics of marginal applicants with high and low ML scores to assess what would happen if the firm were to follow ML recommendations for this set of applicants. We find that following UCB recommendations on the margin would increase both applicant quality and the share of Black, Hispanic, and female interviewees. In contrast, following SL recommendations would increase quality but decrease minority representation. These results are consistent with our earlier results on the interviewed-only subsample.

A key concern with all of these approaches is that hiring likelihood may not be the most appropriate measure of quality. Firms may care about on the job performance and recruiters might sacrifice hiring likelihood and instead choose to interview candidates who would perform better in their roles, if hired. Our ability to address this concern is unfortunately limited by data availability: we observe job performance ratings for very few employees in our training period, making it impossible to train a model to predict on the job performance. We show, however, that all of our ML models are more positively correlated with hiring performance than our human SL model, suggesting that it is unlikely that our results can be explained by human recruiters successfully trading off hiring likelihood to maximize other dimensions of quality.

Together, our main findings show that, given firms' current hiring practices, there need not be an equity-efficiency tradeoff when it comes to expanding diversity in the workplace. Firms' recruiting practices appear to be far from the Pareto frontier, leaving substantial scope for new technologies

to improve both the quality and diversity of selected candidates. Even though our UCB algorithm places no value on diversity in and of itself, incorporating exploration in our setting would lead our firm to interview twice as many under-represented minorities while more than doubling its predicted hiring yield.

Our results, however, caution, against concluding that algorithms are generically equity and efficiency enhancing. In our setting, a supervised learning approach—which is commonly used by commercial vendors of ML-based HR tools—would improve hiring rates, but at the cost of virtually eliminating Black and Hispanic representation. This substantial difference in outcomes underscores the importance of algorithmic design.

In addition, we explore two extensions. First, we examine algorithmic learning over time. Our test data cover a relatively short time period, 2018-2019Q1, so that there is relatively limited scope for applicant quality to evolve. In practice, however, applicant quality can change substantially over time, both at the aggregate level—the increasing share of women with STEM degrees, say—or at the organizational level—as in when firms adopt programs aimed at better retaining and promoting minority talent. To examine how different types of hiring ML adapt to changes in quality, we conduct simulations in which the quality of one group of candidates substantially changes during our test period. By construction, our static SL model does not respond to changes in quality because its training data are fixed. Whereas our updating SL model slightly outperforms UCB in our actual test sample, this changes when we consider an environment in which the quality of minority applicants is changing. For example, when the quality of Black or Hispanic candidates increases, the updating SL is slow to discover this change because it selects relatively few minorities and therefore does not have a chance to see their improved quality. UCB, however, learns more quickly because it actively seeks out under-represented candidates. The pattern is reversed when we simulate data in which hiring rates for Black and Hispanic candidates decreases: UCB continues to select under-presented minorities for their exploration value and does not stop selecting them until its beliefs have sufficiently revised downward.

In a second extension, we explore the impact of blinding the models to demographic variables. Our baseline ML models all use demographic variables—race and gender—as inputs, meaning that they engage in "disparate treatment," a legal gray area (Kleinberg et al., 2018b). To examine the extent to which our results rely on these variables, we estimate a new model in which we remove demographic variables as explicit inputs. We show that this model can achieve similar improvements in quality, but with more modest increases in share of under-represented minorities who are selected; we see a much greater increase in Asian representation because these candidates

are more heterogeneous on other dimensions (such as education and geography) and therefore receive larger "exploration bonuses" in the absence of information about race.

The remainder of the paper is organized as follows. Section 1 discusses our firm's hiring practices and its data. Section 2 presents the firm's interview decision as a contextual bandit problem and outlines how algorithmic interview rules would operate in our setting. Section 3 discuss how we explicitly construct and validate our algorithms. We present our main results on diversity and quality in Section 4, while Sections 5 and 6 discuss our learning and demographics-blinding extensions, respectively.

1 Setting

We focus on recruiting for high-skilled, professional services positions, a sector that has seen substantial wage and employment growth in the past two decades (BLS, 2019). At the same time, this sector has attracted criticism for its perceived lack of diversity: female, Black, and Hispanic applicants are substantially under-represented relative to their overall shares of the workforce (Pew, 2018). This concern is acute enough that companies such as Microsoft, Oracle, Allstate, Dell, JP Morgan Chase, and Citigroup offer scholarships and internship opportunities targeted toward increasing recruiting, retention, and promotion of those from low-income and historically under-represented groups.[10] However, despite these efforts, organizations routinely struggle to expand the demographic diversity of their workforce—and to retain and promote those workers—particularly in technical positions (Jackson, 2020; Castilla, 2008).

Our data come from a Fortune 500 company in the United States that hires workers in several job families spanning business and data analytics. All of these positions require a bachelor's degree, with a preference for candidates graduating with a STEM major, a master's degree, and, often, experience with programming in Python, R or SQL. Like other firms in its sector, our data provider faces challenges in identifying and hiring applicants from under-represented groups. As described in Table 1, most applicants in our data are male (68%), Asian (58%), or White (29%). Black and Hispanic candidates comprise 13% of all applications, but under 5% of hires. Women, meanwhile, make up 32% of applicants and 34% of hires.

In our setting, initial interview decisions are a crucial part of the hiring process. Openings for professional services roles are often inundated with applications: our firm receives approximately 200 applications for each worker it hires. Interview slot are scarce: because they are conducted by current employees who are diverted from other types of productive work, firms are extremely

[10]For instance, see here for a list of internship opportunities focused on minority applicants. JP Morgan Chase created Launching Leaders and Citigroup offers the HSF/Citigroup Fellows Award.

selective when deciding which of these applicants to interview: our firm rejects 95% of applicants prior to interviewing them. These initial interview decisions, moreover, are made on the basis of relatively little information: our firm makes interview decisions on the basis of resume review only.

Given the volume of candidates who are rejected at this stage, recruiters may easily make mistakes by interviewing candidates who turn out to be weak, while passing over candidates who would have been strong. Indeed, even after being interviewed, only 10% of applicants are hired (20% receive an offer and 50% of those who receive an offer reject it), suggesting that there may be scope for firms to extend interview opportunities to a stronger set of candidates, or reduce the number of interviews it needs to conduct to achieve its current hiring outcomes.

In this paper, we therefore focus on how firms can improve the quality of its initial interview decisions, as measured by the eventual hiring rates of interviewed workers. We focus on this margin because it is both empirically important given the scarcity of interview slots, and because we have enough data on interview outcomes (hiring or not) to train ML models to predict this outcome. Firms may, of course, care about other dimensions of quality such as on the job performance once hired. In Section 4.2 we provide noisy but suggestive evidence that ML models trained to maximize hiring rates are also positively related to on the job performance measures (ratings and promotions), which we observe for a small subset of test-period hires. We further note that both hiring rates and on the job performance measures are based on the discretion of managers (to hire, promote, or give high performance ratings to) and may exhibit various forms of evaluation bias. With these caveats in mind, we focus on maximizing quality as defined by a worker's likelihood of being hired, if interviewed. We formalize this notion in the following section.

2 Conceptual Framework

2.1 Resume Screening: Contextual Bandit Approach

Model Setup

We model the firm's interview decision as a contextual bandit problem. Decision rules for standard and contextual bandits have been well studied in the computer science and statistics literatures (cf. Bubeck and Cesa-Bianchi, 2012). In economics, bandit models have been applied to study doctor decision-making, ad placement, recommendation systems, and adaptive experimental design (Thompson, 1933; Berry, 2006; Currie and MacLeod, 2020; Kasy and Sautmann, 2019; Dimakopoulou et al., 2018). Our set up follows Li et al. (2017).

Each period t, the firm sees a job applicant and must choose between one of two actions or "arms": interview or not, $I \in \{0, 1\}$. The firm would only like to interview candidates it would hire, so a measure of an applicant's quality is her "hiring potential": $H_{it} \in \{0, 1\}$ where $H_{it} = 1$ if an applicant would be hired if she were interviewed. Regardless, the firm pays a cost, c_t, per interview, which can vary exogenously with time to reflect the number of interview slots or other constraints in a given period. The firm's "reward" each period is therefore given by:

$$Y = \begin{cases} H_{it} - c_t \text{ if } I = 1 \\ 0 \text{ if } I = 0 \end{cases}$$

After each period t, the firm observes the reward associated with its chosen action.

So far, this set up follows a standard multi-armed bandit (MAB) approach, in which the relationship between action and reward is invariant. The optimal solution to MAB problems is characterized by Gittins (1979) and Lai and Robbins (1985). Our application departs from this set up because firms also observe additional information about the applicants' demographics, education, and work history, X. These variables provide "context" that can inform the expected returns to interviewing a candidate. In general, the solutions to *contextual* MABs are complicated by the dimension of the potential context space. To make our model tractable, we follow Li (2010, 2017) and assume that relationship between context X and rewards (e..g hiring potential) follows a generalized linear form. In particular, we write $E[H|X] = \mu(X'\theta^*)$, where $\mu : \mathbb{R} \to \mathbb{R}$ is link function and θ^* is an unobserved vector describing the true predictive relationship between covariates X and hiring potential.

We express a firm's interview policy I as:

$$I_{it} = \mathbb{I}(s_t(X_{it}) > 0) \tag{1}$$

where $s_t(X_{it})$ can be thought of as a score measuring the value the firm places on a candidate with covariates X_{it} at time t, defined so that the interview cutoff can be normalized to zero. This score can reflect factors such as the firm's beliefs about a candidate's hiring potential and can be a function of both the candidate's covariates X_{it}, as well as the data available to the firm at time t.

As is standard in the literature on bandit problems, we express the firm's objective function in terms of choosing an interview policy I to minimize expected cumulative "regret," the difference in

rewards between the best choice at a given time and the firm's actual choice:

$$R(I) = \sum_{t=1}^{\infty} [I_t^*(\mu(X_{it}\theta^*) - c_t) - I_t(\mu(X_{it}\theta^*) - c_t)] \tag{2}$$

In Equation (2), I^* refers to the optimal interview policy that interviews a candidate if and only if $H_{it}-] > c_t$. The firm's goal is to identify a scoring function $s_t(X_{it})$ that leads it to correctly identify and interview applicants with $H = 1$.

"Greedy" solutions

Before turning toward more advanced algorithms, we first note that one class of potential solutions to bandit problems are given by so-called "greedy" or "exploitation only" algorithms. These types of algorithms ignore the dynamic learning problem at hand and simply choose the arm with the highest expected reward in the present. In our case, a firm following a greedy solution would form its best guess of θ^* given the training data it has available, and then score candidates on their basis of their expected hiring likelihood: $s_t(X) = \mu(X'\hat{\theta}_t)$.

Supervised learning algorithms are designed to implement precisely this type of greedy solution. That is, a standard supervised learning model forms $\hat{E}_t[H|X]$ using the data it has available at time t. If the firm uses this model to make interview decisions, it would select applicants based on their expected quality alone. If $\hat{E}_t[H|X]$ is estimated once at $t = 0$ and invariant thereafter, this decision rule would correspond to our "static SL" model; if it is re-estimated in each period to incorporate new data, then this is equivalent to our "updating SL" model.

Exploration-based solutions

It is widely known, however, that greedy algorithms are inefficient solutions to contextual bandit problems because they do not factor the ex post value of learning into their ex ante selection decisions (Dimakopoulou et al., 2018).[11]

While there is in general no generic optimal strategy for contextual bandits, an emerging literature in computer science focuses on developing a range of computationally tractable algorithms that work well in practice.[12] For example, recently proposed contextual bandit algorithms include UCB (Auer, 2003), Thompson Sampling (Agrawal and Goyal (2013), and LinUCB (Li, Chi, Langford

[11] Bastani et al. (2019) show that exploration-free greedy algorithms (such as supervised learning) are generally sub-optimal.

[12] In particular, the best choice of algorithm for a given situation will depend on the number of possible actions and contexts, as well as on assumptions regarding the parametric form relating context to reward.

and Schapire, 2010).[13] All of these algorithms share the feature that they will sometimes select candidates who do not have the highest expected quality, but whose interview outcomes could improve its estimates of hiring potential in the future.

We follow Li et al. (2017) and implement a generalized linear model version of the UCB algorithm, which assumes that $E[H|X]$ follows the functional form given by $\mu(X'\theta^*)$, as discussed above.[14] Given this assumption, Li et al. (2017) shows that the optimal solution assigns a candidate to the arm (interview or not) with the highest combined expected reward and "exploration bonus."[15]

Exploration bonuses are assigned based on the principle of "optimism in the face of uncertainty": the more uncertain the algorithm is about the quality of a candidate based on her covariates, the higher the bonus she receives. This approach encourages the algorithm to focus on reducing uncertainty, and algorithms based on this UCB approach have been shown to be asymptotically efficient in terms of reducing expected regret (Lai and Robbins, 1985; Li et al., 2017; Abbasi-Yadkori et al., 2019). We discuss the specifics of our implementation in the next section.

2.2 SL vs. UCB models

The use of static SL, updating SL, and UCB models can potentially lead to a variety of differences in the composition and quality of selected applicants in both the short and long term. In this section, we focus on the theoretical differences between these models. The empirical specifics of these outcomes—e.g. in the share of minority applicants who are selected for their average hiring rates—will vary depending on the actual data that the models encounter. We discuss model construction and results in Sections 3 and 4.

Hiring Potential of Selected Applicants

Theory predicts that while models that focus on exploration may end up selecting applicants with lower expected hiring potential in the short run (relative to SL models), they should, in the long run, select weakly higher quality candidates via better learning. In some cases, UCB and SL algorithms may differ in the speed at which they learn, but ultimately still arrive at the same estimate of $E[H|X]$. However, it is also possible for the two types of algorithms to hold persistently

[13] In addition, see Agrawal, Hsu, Kale, Langford, Li and Schapire (2014), and Bastani and Bayati (2015). In addition, the existing literature has provided regret bounds—e.g., the general bounds of Russo and Roy (2014), as well as the bounds of Rigollet and Zeevi (2010) and Slivkins (2014) in the case of non-parametric function of arm rewards—and has demonstrated several successful applications areas of application—e.g., news article recommendations (Li, Chi, Langford and Schapire (2010)) or mobile health (Lei, Tewari and Murphy (2017)).

[14] Li, Lu and Zhou (2017) generalizes the classic general LinUCB algorithm for nonlinear relationship between context and reward. Theorem 2 of that paper gives the regret bound and Equation 6 shows the algorithm implementation we follow.

[15] In particular Li et al. (2017) show that the GLM-UCB algorithm has a regret bound of order $\tilde{\mathcal{O}}(d\sqrt{T})$, where d is the number of covariates and T is the number of rounds.

different beliefs about quality, even after many rounds of use. To see this, suppose that applicants from group M are, on average, somewhat higher quality than applicants from a group N. Given information only on applicants' group membership, the firm would prefer to select candidates from group M.

Suppose, however, that the firm's training data includes many N applicants, but only one M applicant, who happens to be of low quality. A static SL model trained on these data would incorrectly assume that group N applicants are superior, and would not change this belief regardless of what data arrive later. An updating SL model would continue selecting from group N as long as their average quality is still greater than the single M applicant in its training data; if this applicant happens to be particularly quality low, an updating SL may never select an M candidate, and therefore never learn that M candidates are in fact higher quality on average. By contrast, a UCB based approach would select more M applicants not because it believes they are higher quality in that moment, but because it also values the opportunity to improve the precision its estimates of their quality. In doing so, it becomes more likely that it would, in this example, positively update its beliefs about the quality of group M applicants.

The rate and extent to which the long term benefits of learning outweigh the short term costs of exploration would depend on the empirical setting. When the θ^*, the parameter that describes the predictive relation between applicant covariates and hiring potential, is fixed and when there is relatively rich training data, theoretically suboptimal SL models may perform as well as if not better than UCB models because there is relatively little scope for additional learning. However, if training data were sparse or if θ^* evolves over time, then the value of exploration is likely to be greater.

Diversity of Selected Applicants

Although none of our models adopt explicit preferences for diversity, SL and UCB models may differ in the demographic composition of the applicants they select. UCB algorithms, by construction, initially select more candidates from groups in which there is greater uncertainty about quality. For example, if there are fewer Hispanic applicants in the model's training data, then—all else equal—the algorithm would have less precise estimates of their quality, meaning that a UCB model should be more likely to select such candidates. This would suggest that UCB models should, at least initially, select more applicants from demographic groups that are under-represented in its training data, relative to SL models. This tendency can be counteracted by several other factors. First, if if there were very little variance in outcomes among the few Hispanics in the training data, then the algorithm may still be more certain about the quality of Hispanics, even if other groups were better represented in the data. Second, Hispanic candidates may have other covariate

traits—such as education or work history—that are common in the data. If this were the case, then average exploration bonuses could be lower for Hispanics even if there are relatively few of them, because the algorithm is also considering rarity as defined by other covariates. Together, this means that while we would theoretically expect UCB models to select more candidates with covariates associated with more variable beliefs about quality, it is not straightforward how this would map into representation by race or gender.

SL and UCB models may also differ in the composition of applicants they select in the long run. The long run behavior of UCB models is governed both by exploration bonuses and by differences in how beliefs evolve. Suppose, for instance, that a UCB model initially selects more Hispanic applicants in order to explore. If this type of exploration leads a UCB model to become more certain that Hispanics have lower hiring rates than Whites, then it will select fewer Hispanic applicants in the long run relative to SL models; if, instead, Hispanics have comparable or better hiring rates, then the initial increase in Hispanic representation would likely persist.

In the next section, we discuss our data and the specifics of our we construct our SL and UCB models.

3 Algorithm Construction

3.1 Data

We have data on 88,666 job applications from January 2016 to April 2019, as described in Table 1. We divide this sample up into a training dataset consisting of 48,719 applicants that arrive before 2018, 2,617 of whom receive an interview, and a test dataset of 43,997 applications that arrive afterward, 2,137 of whom are interviewed. Our models are built on the training data and our analyses of diversity and hiring likelihood are based on out-of-sample model performance in the test data. We split our sample into training and test data by year (rather than taking a random sample) in order to more closely approximate actual applications of hiring ML in which firms would likely provide historical data to train a model that is then applied prospectively.

Input Features

We have information on applicants' educational background, work experience, referral status, basic demographics, as well as information on the type of position to which they applied. Appendix Table A.2 provides a list of these raw variables, as well as some summary statistics. We have self-reported race (White, Asian, Hispanic, Black, not disclosed and other), gender, veteran status,

community college experience, associate, bachelor, PhD, JD or other advanced degree, number of unique degrees, quantitative background (defined having a degree in a science/social science field), business background, internship experience, service sector experience, work history at a Fortune 500 company, and education at elite (Top 50 ranked) US or non-US educational institution. We record the geographic location of education experience at an aggregated level (India, China, Europe). We also track the job family each candidate applied to, the number of applications submitted, and the time between first and most recent application.

To transform this raw information into usable inputs for a machine learning model, we create a series of categorical and numerical variables that serve as "features" for each applicant. We standardize all non-indicator features to bring them into the same value range. Because we are interested in decision-making at the interview stage, we only use information available as of the application date as predictive features. Our final model includes 106 input features.

Interview Outcomes

Each applicant has an indicator for whether they received an interview. Depending on the job family, anywhere from 3-10% of applicants receive an interview. Among candidates chosen to be interviewed, we observe interview ratings, whether the candidate received an offer, and whether the candidate accepted and was ultimately hired. Roughly 20% of candidates who are interviewed receive an offer and, of them, approximately 50% accept and are hired. We will focus on the final hiring outcome as our measure of an applicant's quality, keeping in mind that this is a potential outcome that is only observed for applicants who are indeed interviewed.

Finally, for 180 workers who are hired and have been employed for at least 6 months, we observe a measure of performance ratings on the job. Because this number is too small to train a model on, we will use these data to examine the relationship between maximizing hiring likelihood and on the job performance.

3.2 Models

Here we describe how we construct three distinct interview policies based on static and updating supervised learning, and contextual bandit UCB. For simplicity, we will sometimes write I^{ML} to refer to the interview policy of any of these ML models.

Static Supervised Learning ("S_0" or "static SL")

We first use a standard supervised learning approach to predict an applicant's likelihood of being hired, conditional on being interviewed. At any given time t (which indexes an application round that we observe in the testing period) applicants are selected according to the following interview policy, based on Equation (1) of our conceptual framework:

$$I_t^{S_0} = \mathbb{I}(s^{S_0}(X) > c_t), \text{ where } s^{S_0}(X) = \hat{E}_t[H|X; D_0] \text{ for all } t \tag{3}$$

Here, we emphasize that the firm's estimate of hiring potential at time t depends on the training data that it has available at the time. In the static SL model, we write this data as D_0 to emphasize that it is determined at time $t = 0$ and is not subsequently updated. Using this data, we form an estimate of $s^{S_0}(X)$ using an ensemble model that combines predictions from a L1-regularized logistic regression (LASSO) and a random forest. We first fit the LASSO using three-fold cross validation.[16] We fit the second sub-component, a random forest model, on a second randomly-selected balanced sample from our training set and use three-fold cross validation to choose tree depth, number of trees, and the maximum number of features. We then average the predictions of each model to generate a predicted probability of interview for each applicant.

We evaluate out-of-sample performance on randomly-selected balanced samples from our testing period. Appendix Figure A.1 plots the receiver operating characteristic (ROC) curve and its associated AUC, or area under the curve. These are standard measure of predictive performances that quantify the trade-off between a model's true positive rate and its false positive rate. Formally, the value of the AUC is equal to $\Pr(\mathbb{I}(s(X_i) > c_t) > \mathbb{I}(s(X_j) > c_t) > |I_i = 1, I_j = 0)$.[17] Our model has an AUC of .67, meaning that it will rank an interviewed applicant who is hired higher than an interviewed but not hired applicant 67 percent of the time. We take this not as a measure of optimal ML performance but as an example of what could be feasibly achieved most firms able to organize their administrative records into a modest training dataset with a standard set of CV-level input features.[18]

Updating Supervised Learning ("S_t" or "updating SL")

Our second model presents a variant of the static SL model in which we begin with the same baseline model as the static SL, but update its training data throughout the test period. That is,

[16] Following best practices, as described in Kaebling (2019), we randomly subsample our training data to create a balanced sample, half of whom are interviewed and half of whom are not interviewed.

[17] The AUC is identical to the result of a Wilcoxon or Mann-Whitney U test (Hanley and McNeil, 1982).

[18] We would ideally like to compare our AUC to those of commercial providers, but Raghavan et al. (2019) reports that no firms currently provide information on the validation of their models.

we divide the test data up into "rounds" of 100 applicants. After each round, we take the applicants the model has selected and update its training data with the outcomes of these applicants. Once the training data is updated, we retrain the model and use its updated predictions to make selection decisions in the next round. At any given point t, the updating SL's interview decisions are given by

$$I_t^{S_t} = \mathbb{I}(s^{S_t}(X) > c_t), \text{ where } s^{S_t}(X) = \hat{E}_t[H|X; D_t^S]. \tag{4}$$

Here, D_t^S is the training data available to the algorithm at time t. It is important to emphasize that we can only update the model's training data with *observed* outcomes for the set of applicants selected in the previous period: that is, $D_{t+1}^S = D_t^S \cup (I_t^{S_t} \cap I_t)$. Because we cannot observe hiring outcomes for applicants who are not interviewed in practice, we can only update our data with outcomes for applicants selected by both the model and by actual human recruiters. This will tend to slow down the degree to which the updating SL model can learn about the quality of the applicants it selects, relative to a world in which hiring potential is fully observed for selected applicants.

Upper Confidence Bound ("UCB")

As discussed in Section 2.1, we implement a UCB-GLM algorithm as described in Li et al. (2017). We calculate predicted quality $\hat{E}_t[H|X; D_t^{UCB}]$ using a regularized logistic regression (Cortes, 2019). At time $t = 0$ of the testing sample, our UCB and SL models share the same predicted quality estimate, which is based on the baseline model trained on the 2016-2017 sample. Our UCB model, however, makes interview decisions each period t based on a different scoring function:

$$I_t^{UCB_t} = \mathbb{I}(s_t^{UCB}(X) > c_t), \text{ where } s_t^{UCB}(X) = \hat{E}_t[H|X; D_t^{UCB}] + \alpha B(X; D_t^{UCB}). \tag{5}$$

In Equation (5), the scoring function $s_t^{UCB}(X)$ is a combination of the algorithm's expectations of an applicant's quality based on its training data and an exploration bonus that varies with an applicant's covariates X. Following the model described in Section 2.1, we assume that $E[H|X]$ can be expressed as a generalized linear function $\mu(X'\theta^*)$. In our specific implementation, we assume that μ is a logistic function and, in each round t, estimtate θ^* using a maximum likelihood estimator so that $\hat{E}_t[H|X; D_t^{UCB}] = \mu(X'\hat{\theta}_t)$. Next, we calculate the exploration bonus as $B(X; D_t^{UCB}) = X'V_t^{-1}X$, where $V_t = \sum_{i=1}^{t-1} X_t X_t'$ is the variance covariance matrix.

Intuitively, Equation (5) breaks down the value of an action into an exploitation component and an exploration component. In any given period, the exploitation strategy chooses to interview a candidate if she has an high expected hiring potential $\mu(X'\hat{\theta}_t)$. In contrast, the exploration

interviews a candidate if there is a great deal of variance in her expected outcome. To balance exploitation, Equation (5) indicates that the firm should interview a candidate based on the sum of her estimated mean and variance—equivalently, this simply says that candidates should be evaluated on the basis of the upper confidence bound of their estimated hiring potential, hence the term "UCB." Li et al. (2017) shows that following such a strategy asymptotically minimizes regret in our setting.

As with the updating SL model, we update the UCB model's training data with the outcomes of applicants it has selected—$D_{t+1}^{UCB} = D_t^{UCB} \cup (I_t^{UCB} \cap I_t)$—subject to the limitation that we can only add applicants who are selected by the model and also interviewed in practice. This will tend to slow the rate at which the UCB algorithm learns relative to a live implementation of the algorithm in which all applicants selected by the algorithm are actually interviewed. Based on these new training data, the UCB algorithm updates both its beliefs about hiring potential and the bonuses it assigns.

4 Main Results

4.1 Impacts on Diversity of Interviewed Applicants

Key Findings

We begin by assessing the impact of each policy on the diversity of candidates selected for an interview in our test sample. This is done by comparing $E[X|I = 1]$, $E[X|I^{S_0} = 1]$, $E[X|I^{S_t} = 1]$, and $E[X|I^{UCB} = 1]$, for various demographic measures X, where we choose to interview the same number of people as the actual recruiter. This analysis is straightforward in the sense that we observe demographic covariates such as race and gender for all applicants so that we can easily examine differences in the composition of applicants selected by each of the interview policies described above.

We begin by assessing the racial composition of selected applicants. At baseline, 54% of applicants in our test sample are Asian, 25% are White, 8% are Black, and 4% are Hispanic. Panel A of Figure 1 shows that, from this pool, human recruiters select a similar proportion of Asian and Hispanic applicants (57% and 4%, respectively), but relatively more White and fewer Black applicants (34% and 5%, respectively).

Panels B-D describe our main result with respect to racial diversity: relative to humans, both SL models sharply reduce the share of Black and Hispanics who are selected, while the UCB model sharply increases it. Panel B illustrates this in case of static SL, the approach most commonly used by commercial vendors of hiring ML: the combined share of selected applicants who are Black

or Hispanic falls from 10% to less than 3%. This change is accompanied by an increase in the proportion of interviewed candidates who are White (from 34% to 51%) and a decrease in the share who are Asian (57% to 47%). The updating SL model (Panel C) follows a similar pattern: Black and Hispanic representation falls from 10% to under 5%, White representation increases more modestly from 34% to 40%, and Asian representation stays largely constant. In contrast, Panel D shows that the UCB model increases the Black share of selected applicants from 5% to 14%, and the Hispanic share from 4% to 10%. The White share stays constant, while the Asian share falls from 57% to 41%.

Figure 2 plots the same set of results for gender. Panel A shows that 65% of interviewed applicants are men and 35% are women; this is largely similar to the gender composition of the overall applicant pool. Unlike the case of race, all of our ML models are aligned in selecting more women than human recruiters, increasing their representation to 41% (static SL), 50% (updating SL), or 39% (UCB).

Discussion

The most important question raised by the above analysis is whether these differences in diversity are associated with differences in quality. We will discuss this extensively in the next section and provide evidence that, despite their demographic differences, the quality of applicant pools selected by our SL and UCB models are comparable to each other, and much better than the quality selected by human recruiters.

Another key question relates to how these selection patterns evolve over time. In Figures 1 and 2, we plot demographic characteristics averaged over the entire test period, but this could obscure changes in demographic composition over time. In particular, one may also be concerned that the UCB model engages in exploration by selecting demographically diverse candidates initially, but then "learns" that these candidates are lower quality and selects fewer of them going forward; in this case, the gains we document would erode over time. Appendix Figure A.2 shows that this does not appear to be the case: the proportion of Black and Hispanic candidates selected stays roughly constant over time. This suggests that, in our sample, the quality of minority applicants is high enough that our models do not update downward upon selecting them.

Next, one may also wonder whether the use of exploration bonuses means that demographic minorities are mechanically more likely to be selected by the UCB model. First, we reiterate that all of the ML models we use are designed solely to maximize applicant hiring potential—they do not have a preference for demographic diversity built into their design. That said, a UCB model may favor demographic diversity (above and beyond what is expected by from predicted mean hiring

potential) to the extent that demographic minorities have greater variance in their estimated hiring potential. As discussed in Section 2.2, this may happen—especially, as in the case of Black and Hispanic applicants, when there are few in the training data—but need not mechanically be the case. If minorities do receive higher bonuses, this can occur directly through their demographics, or indirectly insofar as demographics are correlated with work history and other variables that are predictive of higher variance in hiring potential. To demonstrate this point, Appendix Figure A.3 plots the proportion of the total variation in exploration bonuses that can be attributed to different categories of applicant covariates. We find that the greatest driver of variation in exploration bonuses is an applicant's work history variables.

Finally, it is important to note that the results in Figures 1 and 2 are based on the pattern of applicants that the algorithm happens to see in our data. If a different set of applicants had applied to our sample firm—or if a different set had been interviewed—then it is possible that our results would change. In Section 5, we will explore how the SL and UCB algorithms behave under different conditions in which the quality of applicants of different groups is changing over time.

4.2 Impacts on Quality of Interviewed Applicants

Overview

Next, we ask if and to what extent the gains in diversity made by the UCB model come at the cost of quality, as measured by an applicant's likelihood of actually being hired. To assess this, we would ideally like to compare the average hiring likelihoods of applicants selected by each of the ML models to the actual hiring likelihoods of those selected by human recruiters: $E[X|I = 1]$, $E[X|I^{S_0} = 1]$, $E[X|I^{S_t} = 1]$, and $E[X|I^{UCB} = 1]$.

Unlike demographics, however, an applicant's hiring potential H is an outcome that is only observed when applicants are actually interviewed. We therefore cannot directly observe hiring potential for applicants selected by either algorithm, but not by the human reviewer. To address this, we take three complementary approaches, described in turn below. Across all three approaches, we find evidence that both SL and UCB models—despite their differing demographics—would select applicants with greater hiring potential than those select using current human recruiting practices.

Interviewed sample

Our first approach compares the quality of applicants selected by our algorithms among the sample of applicants who are interviewed. However, because all applicants in this sample are—by

definition—selected by human recruiters, we cannot directly compare the accuracy of algorithmic to human choices within this sample, because there is no variation in the latter.

To get around this, we train an additional model to predict an applicant's likelihood of being selected by for an interview, by a human recruiter. That is, we generate a model of $E[I|X]$ where $I \in \{0, 1\}$ are realized human interview outcomes, using same ensemble approach described in Section 3.2.[19]. This model allows us to order interviewed applicants in terms of their human score s^H in addition to their algorithmic scores, s^{S_0}, s^{S_t}, and s^{UCB}.[20] Appendix Figure A.1 plots the ROC associated with this model. Our model ranks a randomly chosen interviewed applicant ahead of a randomly chosen applicant who is not interviewed 76% of the time.[21]

Figure 3 plots a binned scatterplot depicting the relationship between algorithm scores and hiring outcomes among the set of interviewed applicants; each dot represents the average hiring outcome for applicants in a given scoring ventile. Table 2 shows these results as regressions to test whether the relationships are statistically significant. We find that, among those who are interviewed, applicants' human scores are uninformative about their hiring likelihood; if anything this relationship is slightly negative. In contrast, all ML scores have a statistically significant, positive relation between algorithmic priority selection scores and an applicant's (out of sample) likelihood of being hired.

Table 3 examines how these differences in scores translate into differences in interview policies. To do so, we consider "interview" strategies that select the top 25, 50, or 75% of applicants as ranked by each model; we then examine how often these policies agree on whom to select, and which policy performs better when they disagree. Panel A compares the updating SL model to the human interview model and shows that the human model performs substantially worse in terms of predicting hiring likelihood when the models disagree: only 5-8% of candidates favored by the human model are eventually hired, compared with 17-20% of candidates favored by the updating SL model. Panel B finds similar results when comparing the human model to the UCB model. Finally, Panel C shows that, despite their demographic differences, the updating SL and UCB models agree on a greater share of candidates relative to the human model, and there do not appear to be significant differences in overall hiring likelihoods when they disagree: if anything, the UCB model performs slightly better.

[19]The only methodological difference between this model and our baseline static SL model is that, because we are trying to predict interview outcomes as opposed to hiring outcomes conditional on interview, our training sample consists of all applicants in the training period, rather than only those who are interviewed.

[20]Later in this section, we will discuss results that do not require us to model human interview practices.

[21]Although a "good" AUC number is heavily context specific, a general rule of thumb is that tests in the AUC range of $0.75 - 0.85$ have intermediate to good discriminative properties depending on the specific context and shape of the curve (Fischer et al., 2013).

Full sample

Our analysis on the $I = 1$ sample is subject to two important caveats. First, it assesses differences in quality among those who were good enough to be interviewed; if the value of human recruiters is to screen out particularly poor candidates, this value would not be reflected in this analysis. Second, it also requires us to proxy for unobserved human preferences within the set of interviewed candidates by building an ML model to predict interview status. If this model differs prefers worse candidates than the human recruiter actually does, then this would lead us to understate the performance of human recruiters.

In this section, we take a different approach and attempt to estimate the average quality of *all* ML-selected applicants, $E[H|I^{ML} = 1]$. Doing so requires us to assume that there is no selection on unobservables in our sample so that we can infer hiring likelihoods for ML-selected applicants who were not actually interviewed in practice using observed hiring outcomes from applicants with similar covariates who were interviewed: $E[H|I^{ML} = 1, X] = E[H|I^{ML} = 1, I = 1, X]$. Although this is a strong assumption, we believe it is plausible in our setting because recruiters make decisions on the basis of CV variables that we, for the most part, also observe. Importantly, they do not meet, speak with, or otherwise interact with the candidates.[22]

Following DiNardo et al. (1996), we write:

$$\begin{aligned} E[H|I^{ML} = 1] &= \sum_X p(X|I^{ML} = 1)E[H|I^{ML} = 1, X] \\ &= \sum_X \frac{p(I^{ML} = 1|X)p(X)}{p(I^{ML} = 1)} E[H|I^{ML} = 1, X]. \end{aligned}$$

The above expression consists of the following easily observed components: the unconditional distribution of covariates, $p(X)$, the conditional probability of being selected by an ML model for a given set of covariates X, $p(I^{ML} = 1|X)$, and the unconditional probability of selection, $p(I^{ML} = 1)$. Assuming no selection on unobservables, we can proxy for the term $E[H|I^{ML} = 1, X]$ using $E[H|I^{ML} = 1, I = 1, X]$.

Practically, this approach requires common support: for every ML-selected applicant with covariates X, we must be able to find an applicant with the same covariates who is interviewed.

[22] Although recruiters may not observe additional information about the candidate that we cannot, it is possible that the recruiter observes information about the nature of the search that we cannot, such as whether there is a rush to hire. In this case, hiring outcomes for candidates that were not selected by the human may be worse than outcomes for candidates with similar covariates who were selected—not because they candidates were weaker per se, but because they were interviewed when hiring rates were lower. We address this concern by including characteristics of the job search itself our models: e.g. we use information on job family and month of application to predict hiring rates as well. Because there is such strong seasonality in our firm's hiring processes, adding controls for time of year accounts for variation in hiring demand.

In our application we define covariate cells based on race (Black, White, Hispanic, Asian), gender (male, female), and education (bachelors degree or below, masters degree or above).

Using this approach, Figure 4 again shows that ML models outperform human recruiting practices. Among those actually selected (by human recruiters), the average observed hiring likelihood is 10%. Our calculations indicate that, in all cases, ML models select applicants with higher average predicted hiring rates: 15% for static SL, 25% for UCB, and 30% for updating SL. This result is consistent with our findings from the interviewed-only subsample.

Our results also suggest that there can be substantial returns to increasing the size of the training data we use: both dynamic models (updating SL and UCB) do better than the static SL model. Importantly, updating SL and UCB models perform similarly in terms of maximizing hiring likelihoods. The slightly weaker performance of the UCB model may be explained by the fact that an emphasis on exploration means that the algorithm trades off higher performance in earlier periods for increased learning in later periods, which may be relatively more valuable in settings where the relationship between context and rewards (for example, the quality of minority candidates) is changing, a possibility we explore with simulated data in Section 5. Given our actual data, we see no evidence that the gains in diversity that we document in Section 4.1 come at the cost of substantially reducing the quality of selected applicants.

Marginally interviewed sample

The possibility of selection on unobservables can lead to biases in both of our previous sets of analyses. For example, the human SL model we use in Section 4.2 is trained only on features we observe and may therefore miss unobservables that humans may use to correctly predict hiring likelihood, biasing our results on the interviewed-only sample. Similarly, in our decomposition, we use relatively coarse covariate cells in order to assure that there is common support between interviewed and non-interviewed samples, meaning that our full sample analysis may not control for enough covariates to address concerns about selection on unobservables.

In response to these concerns, we consider an alternative approach for valuing the performance of ML models relative to human decisions: instead of asking whether full algorithmic hiring would lead to better outcomes, we ask whether firms can improve quality by adopting a more modest policy of relying on algorithmic recommendations for candidates who are on the margin of being interviewed. We view this approach as complementary in that it allows for selection on unobservables and does not rely on modeling the human decision to grant an interview.

The intuition is as follows: consider a group of candidates who are just at the margin of receiving an interview and, among them, consider those with low ML scores who are just interviewed and

those with high ML scores who are just not interviewed. If the latter group is higher quality, then the firm can increase hiring rates by following algorithmic recommendations more closely and swapping the interview status of these two groups.

To show that this alternative policy would improve outcomes, we need to compare the quality of marginally interviewed candidates with high and low ML scores. Following Benson et al. (2019); Arnold et al. (2018); Abadie (2003b), we identify marginal candidates using an instrument, Z, for being interviewed. Instrument compliers can be thought of as marginal: they are only interviewed because they received a lucky draw of the instrument. Just as a standard LATE identifies treatment effects for compliers, we use a similar approach to identify average hiring potential for compliers: $E[H|I^{Z=1} > I^{Z=0}]$.[23]

Our instrument is assignment to initial resume screeners, following the methodology pioneered by Kling (2006). Applicants in our data are randomly assigned to screeners who review their resumes and make initial interview decisions. These screeners vary greatly in their propensity to pass applicants to the interview round: an applicant may receive an interview if she is assigned to a generous screener and that same applicant may not if she is assigned to a stringent one. For each applicant, we form the jackknife mean pass rate of their assigned screener and use this as an instrument, Z, for whether the applicant is interviewed. Marginal applicants are those who only interviewed if they are lucky enough to draw a generous screener. With this in mind, we propose the following counterfactual interview policy:

$$\tilde{I} = \begin{cases} I^{Z=1} \text{ if } s^{ML} > \overline{\tau}, \\ I \text{ if } \underline{\tau} \leq s^{ML} \leq \overline{\tau}, \\ I^{Z=0} \text{ if } s^{ML} < \underline{\tau}. \end{cases}$$

The policy $\tilde{I}$ takes the firm's existing interview policy, I, and modifies it at the margin: $\tilde{I}$ favors applicants with ML scores by asking the firm to evaluate these applicants as if they were assigned to a generous screener.[24] Similarly, $\tilde{I}$ penalizes applicants with ML scores by treating them as if they face a stringent screener. $\tilde{I}$ differs from the status quo I only in its treatment of instrument compliers: in this case, $\tilde{I}$ chooses to interview compliers with high ML scores and chooses not to interview compliers with low ML scores. The performance of $\tilde{I}$ (in selecting applicants with

[23] In standard potential outcomes notation, the LATE effect is $E[Y^1 - Y^0|I^{Z=1} > I^{Z=0}]$. In our case, we are only interested in the average potential outcome of compliers: $E[Y^1|I^{Z=1} > I^{Z=0}]$. Here, Y^1 is equivalent to a worker's hiring outcome if she is interviewed—this is what we have been calling quality, H. Further, we note that, in practice, our instrument will be continuous; we use binary notation for expositional clarity. In the continuous instruments case, the average quality of instrument compliers is written as $E[H|\lim_{z' \downarrow z} I^{z'} = 1, \lim_{z' \uparrow z} I^{z'} = 0]$.

[24] Again, for simplicity in exposition, we let Z be a binary instrument in this example (whether an applicant is assigned to an above or below median stringency screener) though in practice we will use a continuous variable.

greater hiring potential) relative to I therefore depends entirely on whether marginally interviewed candidates with high scores turn out to have greater hiring potential than those with low scores.

Figure A.5 plots the distribution of jackknife interview pass rates in our data, restricting to the 54 recruiters (two thirds of the sample) who evaluate more than 50 applications (the mean in the sample overall is 156). After controlling for job family, job level, and work location fixed effects, the 75th percentile screener has a 50% higher pass rate than the 25th percentile screener. Table 5 shows that this variation is predictive of whether a given applicant is interviewed, but is not related to any of the applicant's covariates.

Given this, Figure 5 plots characteristics of marginally interviewed applicants with high and low scores, in the case when we favor applicants with high UCB scores, s^{UCB}. In Panel A, we see that marginal applicants with high scores are more likely to be hired than marginal applicants with low scores. In addition to examining the quality of marginal candidates, we can also consider their demographics. In Panels B through D, we show that marginal high score applicants are more likely to be Black, Hispanic, and female. As such, the interview policy defined by $\tilde{I}$ would increase quality and diversity on the margin, relative to the firm's current practices. Figure A.6 repeats this exercise using supervised learning scores. Again, we see that marginally interviewed candidates with high scores were more likely to be hired than those with low scores. However, in contrast to the UCB scores, we see that marginal applicants with high supervised learning scores are less diverse: they are less likely to be Black or Hispanic. These results focusing on marginally interviewed applicants are consistent with our earlier results, which examined average interviewed candidates. Again, these results suggest that following UCB recommendations can increase both quality and diversity relative to the firm's present policies.

Other measures of quality

One concern with our analysis so far is that our measure of quality—likelihood of receiving and accepting an offer—may not be the ultimate measure of quality that firms are seeking to maximize. If firms ultimately care about on the job performance metrics, then they may prefer that their recruiters pass up candidates who are likely to be hired in order to look for candidates that have a better chance of performing well, if hired.

Our ability to assess this possibility is limited by a lack of data on tracking on the job performance. Ideally, we would like to train a model to predict on the job performance (instead of or in addition to hiring likelihood) and then compare the performance of that model to human decision-making. However, of the nearly 49,000 applicants in our training data, only 296 are hired and have data on job performance ratings, making it difficult to accurately build such a model.

We take an alternative approach and correlate measures of on the job performance with our ML scores and human SL score, using data from our training period. If it were the case that humans were trading off hiring likelihood with on the job performance, then our human SL model (e.g. predicting an applicant's likelihood of being interviewed) should be positively predictive of on the job performance, relative to our ML models.

Table 4 presents these results using two measures of performance: on the job performance ratings from an applicant's first mid-year review, and an indicator for whether an applicant has been promoted. On the job performance ratings are given on a scale of 1 to 3, referring to below, at, or above average performance; 13% receive an above average rating. We also examine whether a worker is promoted within the time seen in our sample; this occurs for 8% of hires in the test period.

Panel A examines the correlation between our model of human interview behavior, our "human SL" model, and performance rating and promotion outcomes. Columns 1 and 3 present raw correlations and Columns 2 and 4 control for our static SL, updating SL, and UCB scores so that we are examining the relative correlation between the human model and performance outcomes. In all cases, we observe a negatively signed and sometimes statistically significant relationship: if anything, human recruiters are less likely to interview candidates who turn out to do well on the job. By contrast, Panels B through D conduct the same exercise for each of our ML models; Columns 1 and 3 present raw correlations and Columns 2 and 4 control for the human score. In all cases, these correlations are positively signed and occasionally statistically significant. In particular, the UCB score appears to be most positively correlated with on the job performance outcomes.

We caution that these results are potentially subject to strong sample selection—they examine the correlation between applicant scores among the 233 hires in our test sample, only 180 of whom have mid-year evaluation data. That said, our results provide no evidence to support the hypothesis that human recruiters are successfully trading off hiring likelihood in order to improve expected on the job performance among the set of applicants they choose to interview.

Discussion

Our results show that ML tools can be used to increase the hiring yield of applicants, but may have very different implications for demographic representation. In contrast to our SL models—which substantially reduce the share of Black and Hispanic applicants who are interviewed— our UCB model represents a Pareto improvement in quality and diversity to the firm's recruiting practices. Further, the UCB model also represents a "close" to Pareto improvement relative to the updating SL model, dramatically increasing diversity at a relatively small cost to quality. This result raises several additional questions about the viability of our UCB approach, which we explore in extensions.

First, Figure 4 shows that the updating SL model performs better than UCB in terms of improving the hiring likelihoods of selected applicants, in contrast to the theoretical prediction that models which incorporate exploration should outperform greedy models in the long run (Dimakopoulou et al., 2018). Our analysis however, is limited by the short time span of our test period and the fact that we are not running an experiment in which we can actually interview the candidates that an ML selects (who are not otherwise interviewed). Both of these factors can limit the scope for learning in our setting in a way that is not representative of real-life applications. In Section 5, we conduct simulations to see if exploration is more valuable in settings where these constraints on learning are not in place.

Second, our ML algorithms all make explicit use of race, ethnicity, and gender as model inputs, raising questions about their legality under current employment law. One reason our UCB algorithm is able to expand diversity may be that it is allowed to provide exploration bonuses based on race and gender; taking those away may restrict its ability to explore along demographic dimensions. A growing literature (c.f. Rambachan et al. (2020), Corbett-Davies and Goel (2018), and Kleinberg et al. (2016)) considers how information on protected should be used in algorithmic design. In Section 6, we show how our UCB algorithm is impacted when we restrict its access to demographic information.

5 Learning over time

5.1 Changes in applicant quality

In this section, we examine the value of exploration and learning in greater depth. In particular, our main analysis—based on the actual applicants that the firm receives from January 2018 to April 2019—is limited in two ways. First, we are only able to observe how our algorithms would behave on this relatively short span of data, making it difficult to understand whether our results would hold in other instances, particularly those where the quality of applicants changes substantially over time. Second, the degree of learning in our models is limited by our updating procedure, which only allows us to add in hiring outcomes for ML-selected candidates who are actually interviewed. If ML models choose candidates that are very different from those selected by human recruiters, they will not be able to observe hiring outcomes for these candidates and this will slow down their learning.

To address these issues, we conduct simulations in which the hiring potential of one group of applicants (by race) changes during the test sample and observe how quickly our ML models are able to learn about this change. For example, in one case, we assume that all Black applicants in the test period would be hired if interviewed, while the hiring potential of all other groups remains

the same. In addition to creating a change that increases the value of learning, this simulation comes closer to a live-implementation in which we would be able to observe hiring outcomes for all applicants the ML selects—and update the model's training data accordingly—instead of only for those who are interviewed in reality.

To evaluate how the models learn over time, we examine how they would assess the *same* candidates at different points throughout the simulation test period. Specifically, we take the actual set candidates who applied between January 2019 and April 2019 (hereafter, the "evaluation cohort"), and estimate their model scores $s_t^{ML}(X)$ for each of the three ML models (static SL, updating SL, and UCB), at different points t throughout 2018. By keeping the evaluation cohort the same, we are able to isolate changes in the algorithm's scores that arise from differences in learning and exploration over time.

For intuition, consider the scores of candidates on January 1, 2018, the first day of the test period. In this case, all three ML algorithms would have the same beliefs about the hiring potential of candidates in the evaluation cohort, because they share the same estimate of $E[H|X; D_0)$ trained on the initial data D_0. The static SL and updating SL models would therefore have the same scores; the UCB would have the same "beliefs" but a different score, because it also factors in its exploration bonus. On December 31, 2018, however, the models may have different scores for the same set of evaluation cohort candidates. Because its training data is never updated, the static SL model would have the same scores as it did on January 1. The updating SL and UCB algorithms would have both different beliefs—updated based on hiring outcomes for the potentially different sets of applicants that they selected throughout 2018—and have different scores, because the UCB model in addition continues to factor in exploration bonuses.

We next report how the demographics and quality of selected applicants evolves under our three original models—the static SL, updating SL, and UCB. To better understand how the UCB model differs from the the updating SL, we also consider a forth variant, which tracks who the UCB model would have selected based on its estimates of $E[H|X; D_t^{UCB}]$ alone; this model allows us to track the evolution of the UCB model's beliefs separately from its exploration behavior.

5.2 Results

We first report results from simulations in which we assign a hiring potential of $H = 1$ to all 2018 test period applicants in one racial group, leaving all else equal. To measure learning, we track the share of applicants selected by each algorithm from the group whose quality we increase. Because no other covariates so perfectly predict hiring likelihood in our data, the quality maximizing algorithm would simply select only applicants from this group.

Panel A of Figure 6, plots the share of Black applicants who are selected in the case when all Black applicants are assigned $H = 1$. The flat line, which hovers at just over 1%, represents the proportion of evaluation cohort applicants who would be selected by the static SL algorithm if they arrived at time t between Jan 1, 2018 and December 31, 2018. This line is flat by construction because the static supervised algorithm's beliefs do not change, so it maintains the same rankings of applicants in this cohort regardless of the date on which they are evaluated. By contrast, the UCB model rapidly increases the share of Black candidates it selects: initially, it seeks out Black candidates because they are relatively rare and, upon selecting them, learns that they are all high quality. Over time, as the number of high quality Black examples accumulates, the algorithm updates its beliefs and selects an increasing proportion of Black candidates. In this simulation, the UCB model learns enough about the improved quality of Black applicants to always select them after only a couple months (e.g. after seeing about 800 candidates). In this case, the percentage of Black candidates selected using the UCB model is similar to the share who would have been selected had we focused on beliefs only: the share of Black applicants increases because the algorithm's beliefs about their quality increases, not because Black applicants receiver higher exploration bonuses.

The orange line shows this same process using the updating SL model. While it is able to do learn about the improved quality of Black applicants, it does so at a slower rate. Because supervised learning algorithms apply an exploitation only approach to selection, the updating SL model does not go out of its way to select Black candidates. As such, it has a harder time learning that these candidates are of high quality. This difference in speed is meaningful. To see this, consider a Black applicant from the evaluation cohort who applies in April of 2018. If this person is evaluated by the UCB algorithm at this point, she would have an 19% chance of receiving an interview. If, instead, she were evaluated by the updating SL model, she would only have an 3% chance.

This same pattern can also be seen in Panel B of Figure 6, which plots the percentage of Hispanic applicants who are selected in the case when all Hispanic applicants in 2018 are all assigned $H = 1$. The UCB model quickly picks up on this change while the updating SL model is much slower. This is unsurprising considering Panel C of Figure 1, which shows that only 1.5% of candidates selected by the updating SL model are Hispanic. When this model is allowed to learn on the simulated data, it selects very few Hispanic applicants in the earlier months of 2018, making it difficult to learn that quality has changed for this group.

Panels C and D of Figure 6 plot outcomes for Asian and White applicants under the counterfactual that these groups are assigned quality $H = 1$ in 2018. Here, we see the opposite pattern: the updating SL model quickly learns to select only White or only Asian applicants. The UCB model, by contrast, initially selects fewer White and Asian applicants. The reasons for this behavior can be

seen by comparing the difference between the set of applicants selected by the UCB model and the set that would be selected based on UCB beliefs only: the UCB algorithm learns that Asians and Whites are higher quality almost as quickly as the updating SL, but it selects fewer of them because it is still allocating higher exploration bonuses to other groups. In this way, exploration bonuses initially come at the expense of selecting the highest quality candidates; however, as UCB beliefs about the quality of White and Asian applicants become sufficiently high, the exploration bonuses it grants to other groups are no longer large enough to matter.

In Appendix Figure A.7, we repeat this exercise for cases in which we assume that applicant quality declines, so that all candidates of a specific group are assigned $H = 0$. When Black and Hispanic quality falls, the UCB model is a bit slower to respond because it continues providing higher exploration bonuses to these groups. However, the proportion of selected Black or Hispanic applicants falls to zero within a couple months, as it learns about their poor quality. When White or Asian candidates become low quality, both the updating SL and UCB models reduce the share of such applicants that they select at approximately the same rate, reaching zero within 6 months. We note that this behavior differs from a quota-based system; just because under-represented (in the training data) candidates tend to receive higher exploration bonuses, this does not mean that the algorithm will necessarily select them at any specific rate, especially if its beliefs about the quality of these candidates decreases.

6 Blinding the Model to Applicant Demographic Characteristics

So far, our algorithms have used race, ethnicity, and gender as explicit model inputs. This means that our algorithms engage in "disparate treatment" on the basis of protected categories, in possible violation of employment and civil rights law (Kleinberg et al., 2018b).[25] A natural question, then, is how much of our results would hold if we eliminated the use of race and gender as model inputs (as a practical matter, we continue to allow the inclusion of other variables, such as geography, which may be correlated). In particular, our UCB model is able to increase diversity and quality, relative to human selection practices: would this still hold if we restricted the use of demographic inputs?

In our UCB model, race and gender enter in two ways: first, as predictive features of the model that are used to predict an applicant's chances of being hired if interviewed; and second, as inputs into how exploration bonuses are assigned. The model, may, for instance, be able to select more Black applicants by recognizing race as a dimension on which these applicants are rare, relative to

[25] A number of recent papers have considered the impacts of anonymizing applicant information on employment outcomes (Goldin and Rouse, 2000; Åslund and Skans, 2012; Behaghel et al., 2015; Agan and Starr, 2018; Alston, 2019; Doleac and Hansen, 2020; Craigie, 2020).

those that are Asian or White. If this were the case, then restricting the use of race as a model input could hinder the algorithm's ability to assign higher bonuses to minorities on average; whether this is the case or not depends on whether Black and Hispanic applicants are under-represented on other dimensions that the model can still use.

In this section, we re-estimate the UCB model without the use of applicants' race, gender, and ethnicity in either prediction or bonus provision. Figure 7 shows how blinding affects diversity. Panels A and C reproduce the race and gender composition of applicants selected by the unblinded UCB model and Panels B and D track the blinded results. Blinding reduces the share of selected applicants who are Black or Hispanic, from 23% to 14%, although there is still greater representation relative to human hiring (10%). The most stark differences, however, come in the treatment of White and Asian applicants. In the non-blinded model, White and Asian applicants make up approximately the same share of of interviewed applicants (35% and 41%, respectively), even though there are substantially more Asian applicants. When the algorithm is blinded, however, many more Asian applicants are selected relative to Whites (61% vs. 26%). In our data, this likely arises for two reasons. First, Asian applicants are more likely to have a master's degree or above, a trait that is more strongly rewarded for White applicants; blinding the algorithm to race therefore increases the returns to education among Asian applicants. Second, in the race-aware model, Asian applicants received smaller exploration bonuses because they comprised a majority of the applicant pool; when bonus provision is blinded, exploration bonuses for Asian applicants increase because they are more heterogenous on other dimensions (such as having niche majors) that lead to higher bonuses. In Panels C and D, we find little impact of blinding on gender composition.

Figure 8 examines the predictive accuracy of various blinded algorithms, using the decomposition-reweighting approach described in Section 4.2. Here, blinding—if anything—improves the quality of the UCB algorithms. In our setting, this arises because Asian applicants, who make up the majority of applicants selected by the blinded UCB model, tend to have a higher likelihood of hire in the test period, conditional on being interviewed. As discussed above, the race-aware UCB model implicitly restricted the number of Asian applicants who are selected because it could see that these applicants shared a covariate variable—being Asian—that was very common in the sample. When the algorithm is no longer permitted to do this explicitly, the share of selected applicants who are Asian increases, increasing hiring yield on average. Theoretically, this could reduce the efficiency of future learning because blinding restricts the information the algorithm can use to decide which candidates are rare, but, as discussed in Section 5, the value of active learning in our actual test sample is relatively small, making this a small cost relative to the gain that comes from selecting more applicants from a higher yield group.

7 Conclusion

While a growing body of papers have pointed out potential gains from following algorithmic recommendations, our paper goes further to compare outcomes between standard supervised learning approaches and a contextual bandit based approaches that value exploration. Our results are consistent with a range of recent papers showing that traditional supervised learning approaches may over-weight historical successes, at the expense of learning about present and future relationships between applicant covariates and outcomes. Indeed, the value of simply updating one's training data (even without a preference for exploration), can yield significant gains in accuracy, relative to a static approach.

Our results shed new light on the value of exploration and the the relationship between efficiency and equity in the provision of job opportunities. We show, for instance, that SL models substantially increase quality and decrease diversity relative to the firm's actual practices. One way to interpret this finding is to say that firms may be engaging in some type of de-facto affirmative action that prioritizes equity at the expense of efficiency. Implicitly, this framing implies that firms are making a tradeoff at the Pareto frontier.

Our UCB model results, however, indicate that such an explanation would be misleading. Rather than making a tradeoff, our results suggest that human recruiters may simply be inefficient at valuing equity: in seeking a more diverse set of candidates than a SL approach would, they end up selecting weaker candidates minority candidates over stronger minority candidates.

This paper shows that a data-driven approach to valuing exploration can lead firms to improve their hiring outcomes by identifying stronger candidates from traditionally under-represented groups—even when the algorithm is not explicitly charged with increasing diversity, and even when it is blinded to demographic inputs. These findings go against the idea that gains in equity must always come at the expense of efficiency; rather, firms appear to be operating inside the Pareto frontier, and a contextual bandit approach to designing algorithms can lead to gains along both dimensions.

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 2003, *113* (2), 231–263.

_ , "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 2003, *113* (2), 231–263.

Abbasi-Yadkori, Yasin, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz, "POLITEX: Regret bounds for policy iteration using expert prediction," in "International Conference on Machine Learning" 2019, pp. 3692–3702.

Agan, Amanda and Sonja Starr, "Ban the box, criminal records, and racial discrimination: A field experiment," *The Quarterly Journal of Economics*, 2018, *133* (1), 191–235.

Alston, Mackenzie, "The (Perceived) Cost of Being Female: An Experimental Investigation of Strategic Responses to Discrimination," *Working paper*, 2019.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Arnold, David, Will Dobbie, and Crystal S Yang, "Racial Bias in Bail Decisions*," *The Quarterly Journal of Economics*, 2018, p. qjy012.

Åslund, Olof and Oskar Nordströum Skans, "Do anonymous job application procedures level the playing field?," *ILR Review*, 2012, *65* (1), 82–107.

Barocas, Solon and Andrew D. Selbst, "Big Data's Disparate Impact," *SSRN Electronic Journal*, 2016.

Bastani, Hamsa, Mohsen Bayati, and Khashayar Khosravi, "Mostly Exploration-Free Algorithms for Contextual Bandits," *arXiv:1704.09011 [cs, stat]*, November 2019. arXiv: 1704.09011.

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon, "Unintended effects of anonymous resumes," *American Economic Journal: Applied Economics*, 2015, *7* (3), 1–27.

Benson, Alan, Danielle Li, and Kelly Shue, "Promotions and the peter principle," *The Quarterly Journal of Economics*, 2019, *134* (4), 2085–2134.

Berry, Donald A, "Bayesian clinical trials," *Nature reviews Drug discovery*, 2006, *5* (1), 27–36.

BLS, "Industries with the largest wage and salary employment growth and declines," 2019.

Bogen, Miranda and Aaron Rieke, "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias," 2018.

Bubeck, Sébastien and Nicolo Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

Castilla, Emilio, "Gender, Race, and Meritocracy in Organizational Careers," *American Journal of Sociology*, 2008, *113* (6) 1479–1526.

Corbett-Davies, Sam and Sharad Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *arXiv:1808.00023 [cs]*, August 2018. arXiv: 1808.00023.

Cortes, David, "Adapting multi-armed bandits policies to contextual bandits scenarios," *arXiv:1811.04383 [cs, stat]*, November 2019. arXiv: 1811.04383.

Cowgill, Bo, "Bias and productivity in humans and algorithms: Theory and evidence from resume screening," *Columbia Business School, Columbia University*, 2018, *29*.

_ **and Catherine E Tucker**, "Economics, fairness and algorithmic bias," *preparation for: Journal of Economic Perspectives*, 2019.

Craigie, Terry-Ann, "Ban the Box, Convictions, and Public Employment," *Economic Inquiry*, 2020, *58* (1), 425–445.

Currie, Janet M. and W. Bentley MacLeod, "Understanding Doctor Decision Making: The Case of Depression Treatment," *Econometrica*, 2020, *88* (3), 847–878.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta, "Automated experiments on ad privacy settings," *Proceedings on privacy enhancing technologies*, 2015, *2015* (1), 92–112.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens, "Estimation Considerations in Contextual Bandits," *arXiv:1711.07077 [cs, econ, stat]*, December 2018. arXiv: 1711.07077.

DiNardo, John, Nicole M Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 1996, *64* (5), 1001–1044.

Doleac, Jennifer L and Benjamin Hansen, "The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden," *Journal of Labor Economics*, 2020, *38* (2), 321–374.

Fischer, Christine, Karoline Kuchenbäcker, Christoph Engel, Silke Zachariae, Kerstin Rhiem, Alfons Meindl, Nils Rahner, Nicola Dikow, Hansjörg Plendl, Irmgard Debatin et al., "Evaluating the performance of the breast cancer genetic risk models BOADICEA, IBIS, BRCAPRO and Claus for predicting BRCA1/2 mutation carrier probabilities: a study based on 7352 families from the German Hereditary Breast and Ovarian Cancer Consortium," *Journal of medical genetics*, 2013, *50* (6), 360–367.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of" blind" auditions on female musicians," *American economic review*, 2000, *90* (4), 715–741.

Hanley, James A and Barbara J McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, 1982, *143* (1), 29–36.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in hiring," *The Quarterly Journal of Economics*, 2018, *133* (2), 765–800.

Jackson, Summer, "Not Paying for Diversity: Repugnance and Failure to Choose Labor Market Platforms that Facilitate Hiring Racial Minorities into Technical Positions," 2020.

Kaebling, Leslie P, "Lecture Notes in 6.862 Applied Machine Learning: Feature Representation," February 2019.

Kasy, Maximilian and Anja Sautmann, "Adaptive treatment assignment in experiments for policy choice," 2019.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2018, *133* (1), 237–293.

_ , Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein, "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 2018, *10*.

_ , Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv:1609.05807 [cs, stat]*, November 2016. arXiv: 1609.05807.

Kling, Jeffrey R, "Incarceration length, employment, and earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Kuhn, Peter J and Lizi Yu, "How Costly is Turnover? Evidence from Retail," Technical Report, National Bureau of Economic Research 2019.

Lai, Tze Leung and Herbert Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, 1985, *6* (1), 4–22.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables," in "Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2017, pp. 275–284.

Lambrecht, Anja and Catherine Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Management Science*, 2019, *65* (7), 2966–2981.

Li, Lihong, Yu Lu, and Dengyong Zhou, "Provably Optimal Algorithms for Generalized Linear Contextual Bandits," in "Proceedings of the 34th International Conference on Machine Learning - Volume 70" ICML'17 JMLR.org 2017, p. 2071–2080.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C. Young, Jeffrey De Fauw, and Shravya Shetty, "International evaluation of an AI system for breast cancer screening," *Nature*, *577* (7788), 89–94, year=2020.

Mullainathan, Sendhil and Ziad Obermeyer, "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error," *NBER WP*, 2019.

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

Pew Research Center, *Women and Men in STEM Often at Odds Over Workplace Equity* January 2018.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy, "Mitigating bias in algorithmic employment screening: Evaluating claims and practices," *arXiv preprint arXiv:1906.09208*, 2019.

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig, "An Economic Approach to Regulating Algorithms," Working Paper 27111, National Bureau of Economic Research May 2020.

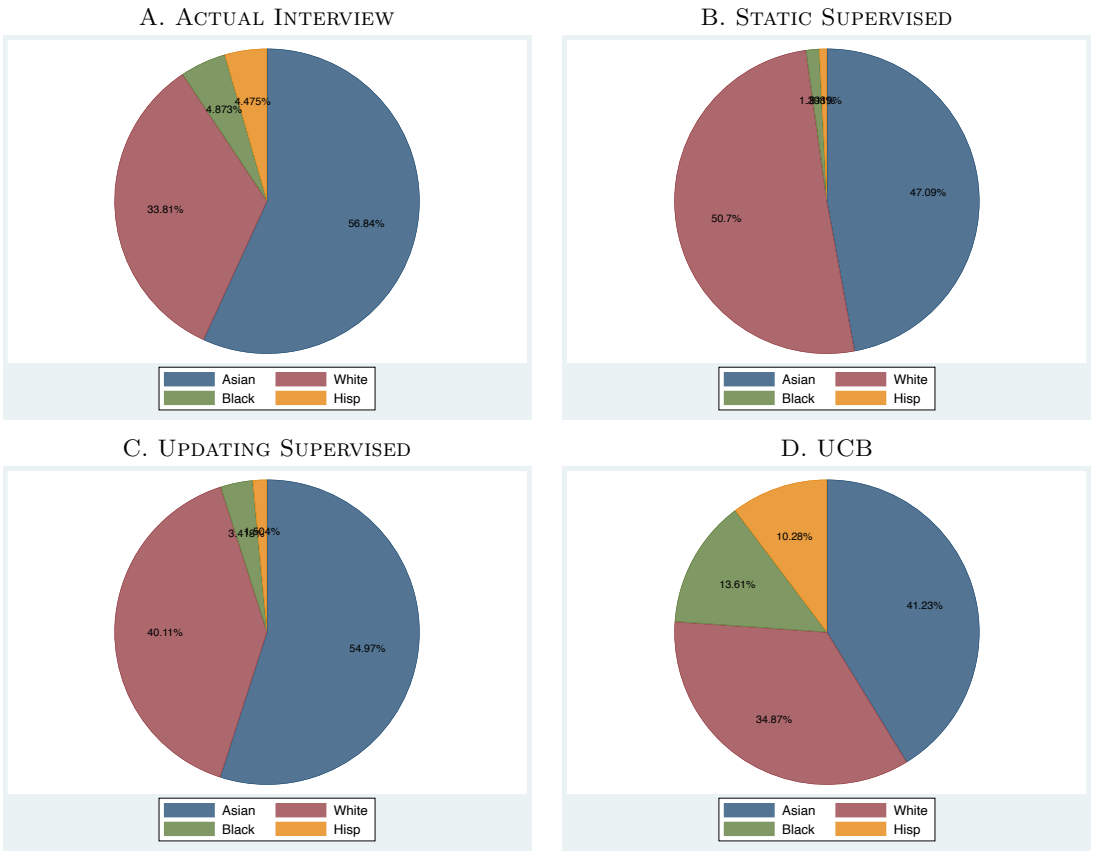
Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Apr 2015, *115* (3), 211–252.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver, "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," 2019.

Thompson, William R, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933, *25* (3/4), 285–294.

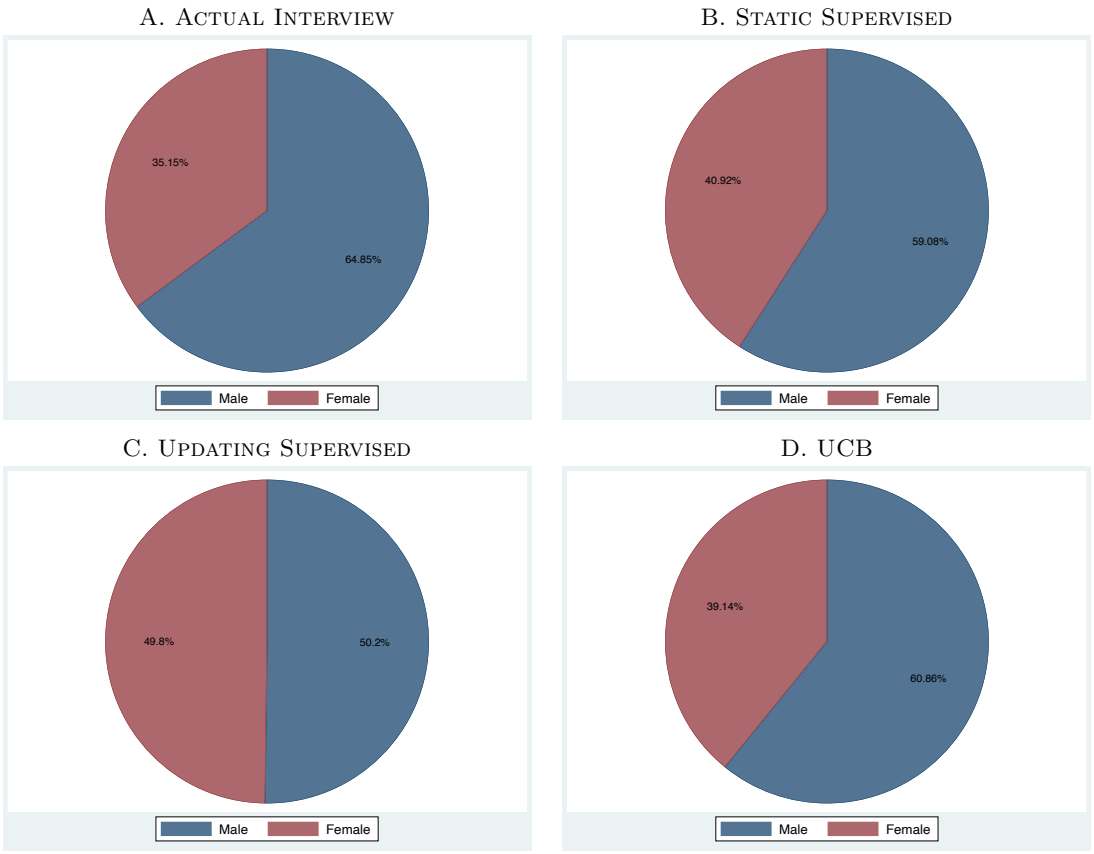
Yala, Adam, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay, "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction," *Radiology*, 2019, *292* (1), 60–66. PMID: 31063083.

FIGURE 1: RACIAL COMPOSITION



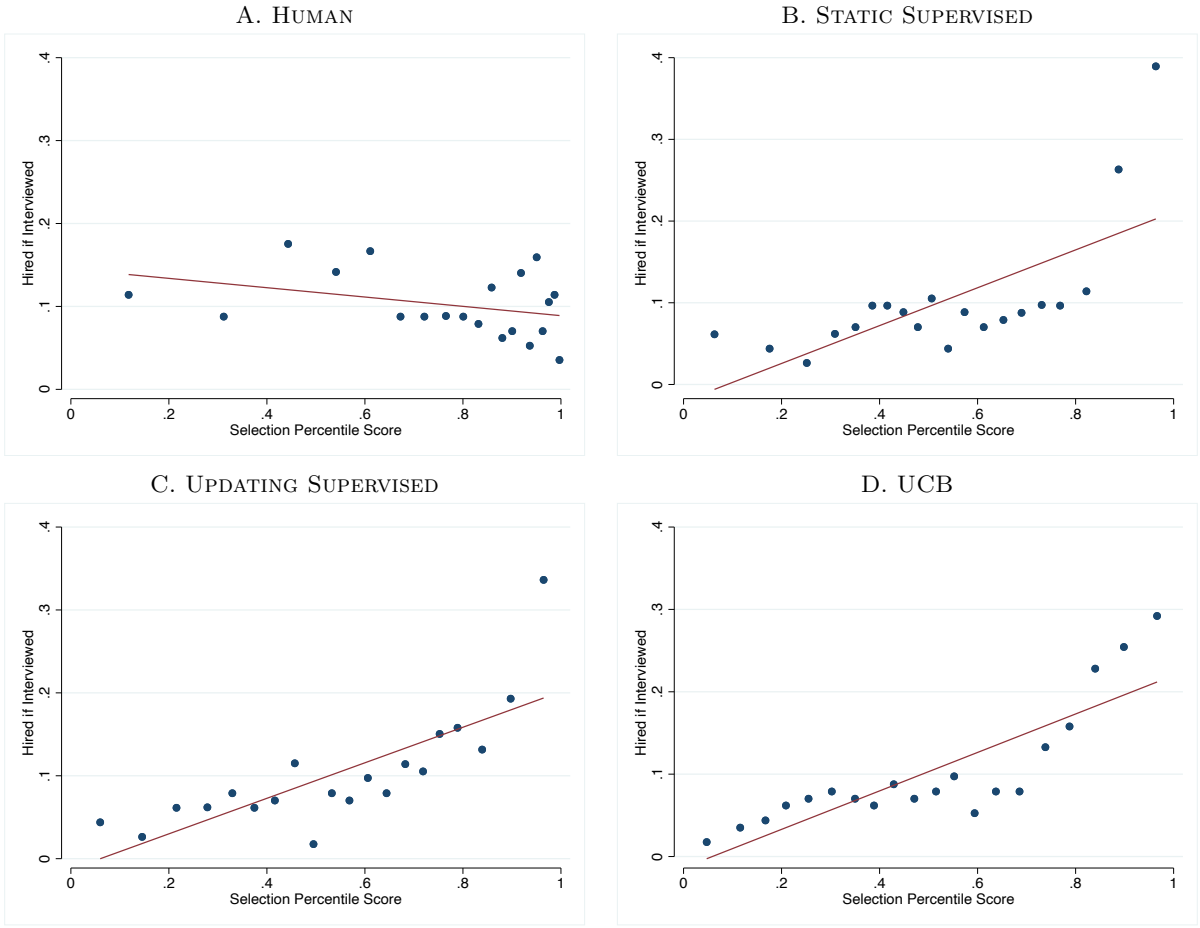
NOTES: This figure shows the racial composition of applicants to professional-services positions at the firm from 2018-2019. Panel A shows the racial composition of applicants selected for an interview by the firm. Panel B shows the racial composition of applicants selected for an interview by the static supervised learning algorithm. Panel C shows the racial composition of applicants selected for an interview by the supervised learning algorithm that updates the training data each. Panel D shows the racial composition of applicants selected for an interview by the UCB algorithm. See the text for details on the construction of each algorithm. All data come from the firm's application and hiring records.

FIGURE 2: GENDER COMPOSITION



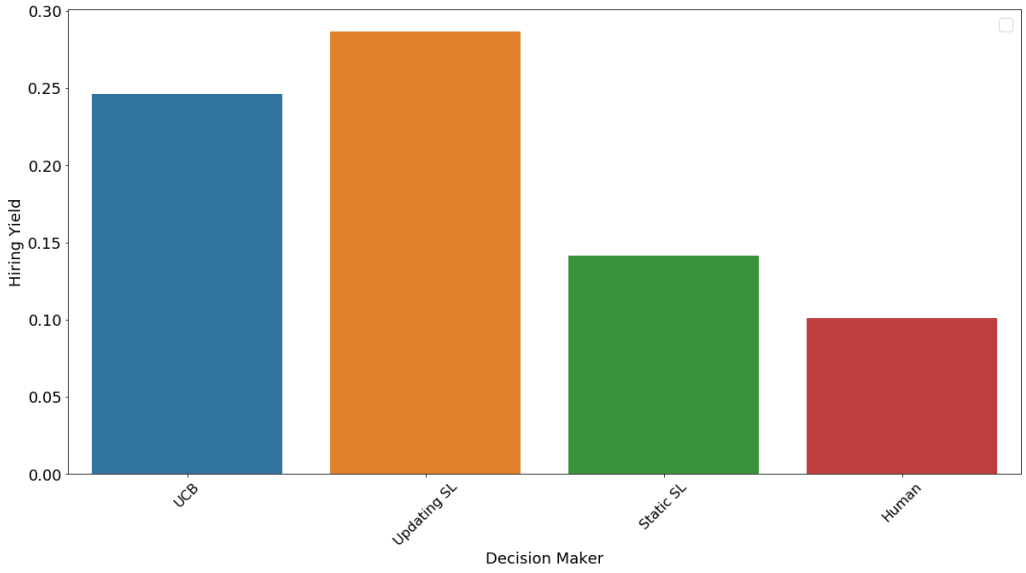
NOTES: This figure shows the gender composition of applicants to professional-services positions at the firm from 2018-2019. Panel A shows the gender composition of applicants selected for an interview by the firm. Panel B shows the gender composition of applicants selected for an interview by the static supervised learning algorithm. Panel C shows the gender composition of applicants selected for an interview by the supervised learning algorithm that updates the training data each. Panel D shows the gender composition of applicants selected for an interview by the UCB algorithm. See the text for details on the construction of each algorithm. All data come from the firm's application and hiring records.

FIGURE 3: CORRELATIONS BETWEEN ALGORITHM SCORES AND HIRING LIKELIHOOD



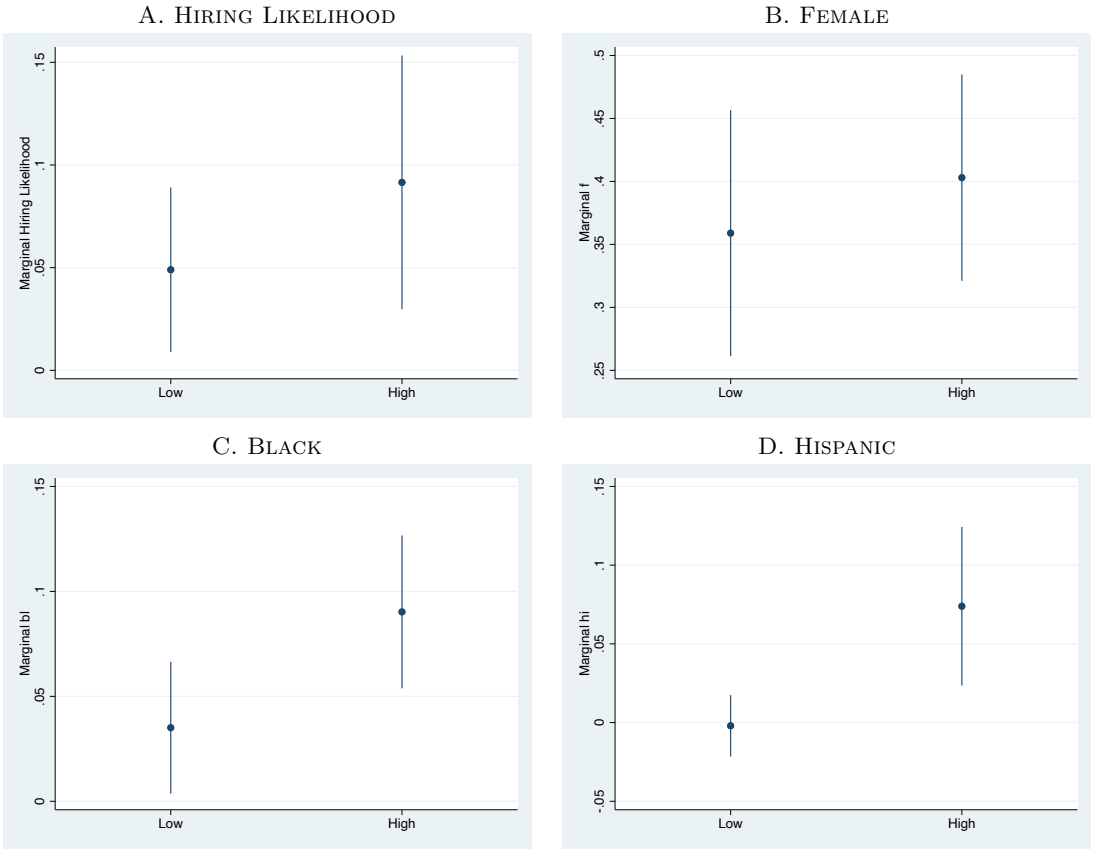
NOTES: Each panel of this figure plots algorithm selection scores on the x-axis and the likelihood of an applicant being hired if interviewed on the y-axis. Panel A shows the selection scores from an algorithm that predicts the firm's actual selection of which applicants to interview. Panel B shows the selection scores from the supervised-learning algorithm that predicts whether an applicant will be hired if interviewed. Panel C shows the selection scores from the UCB algorithm that predicts whether an applicant will be hired if interviewed. See the text for details on the construction of each algorithm. All data come from the firm's application and hiring records.

FIGURE 4: AVERAGE HIRING LIKELIHOOD, FULL SAMPLE



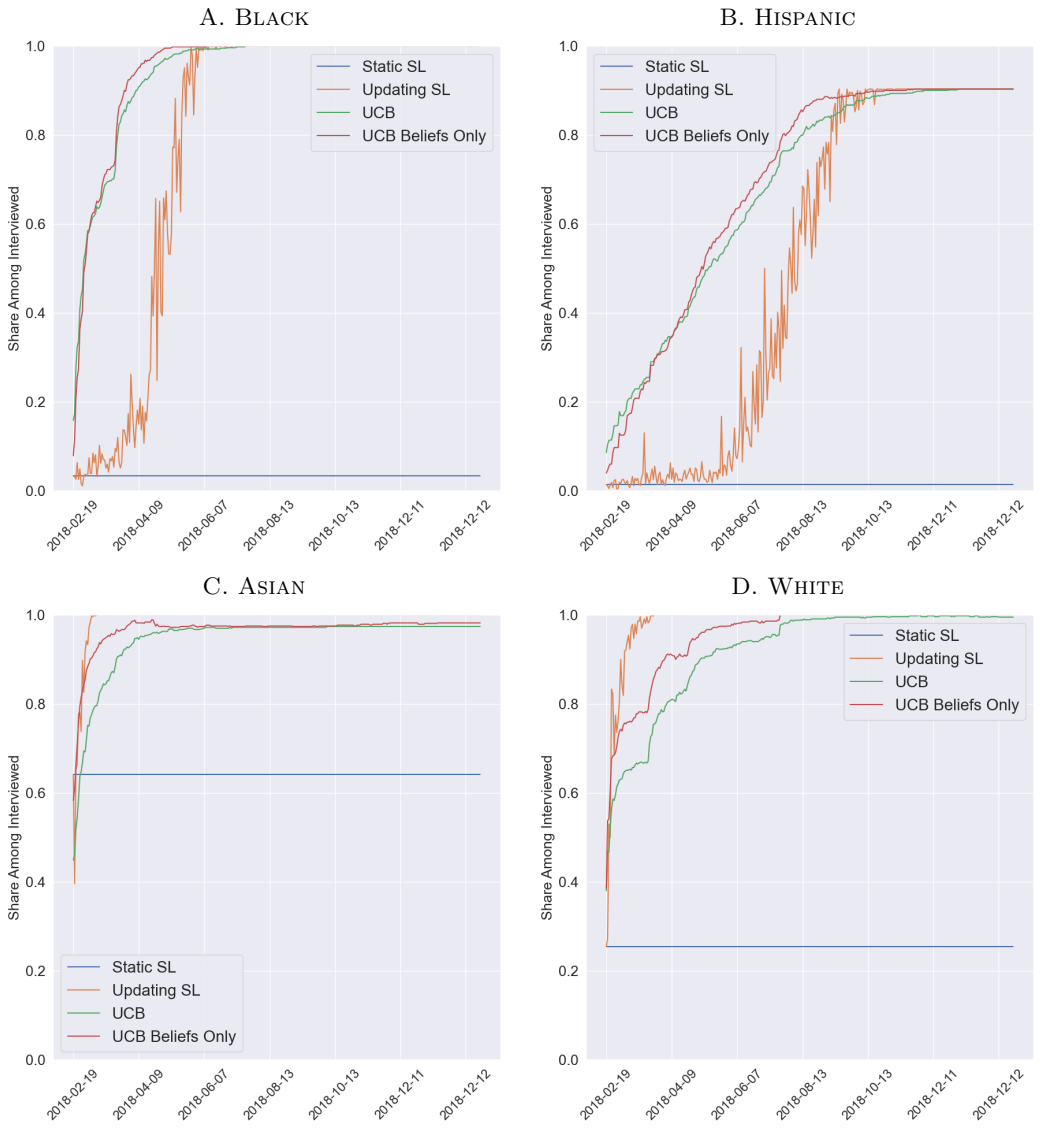
NOTES: This figure shows the unconditional average hiring yield of each unblinded algorithmic selection strategy and the human recruiters. The first bar shows average hiring yield for applicants selected by the contextual bandit UCB algorithm, the second by the updating supervised learning algorithm and the third bar plots the static supervised learning version. The fourth bar shows the unconditional hiring yield for candidates selected by human recruiters. See text for details on the construction of the unweighted average. All data come from the firm's application and hiring records.

FIGURE 5: CHARACTERISTICS OF MARGINAL INTERVIEWEES, BY UCB SCORE



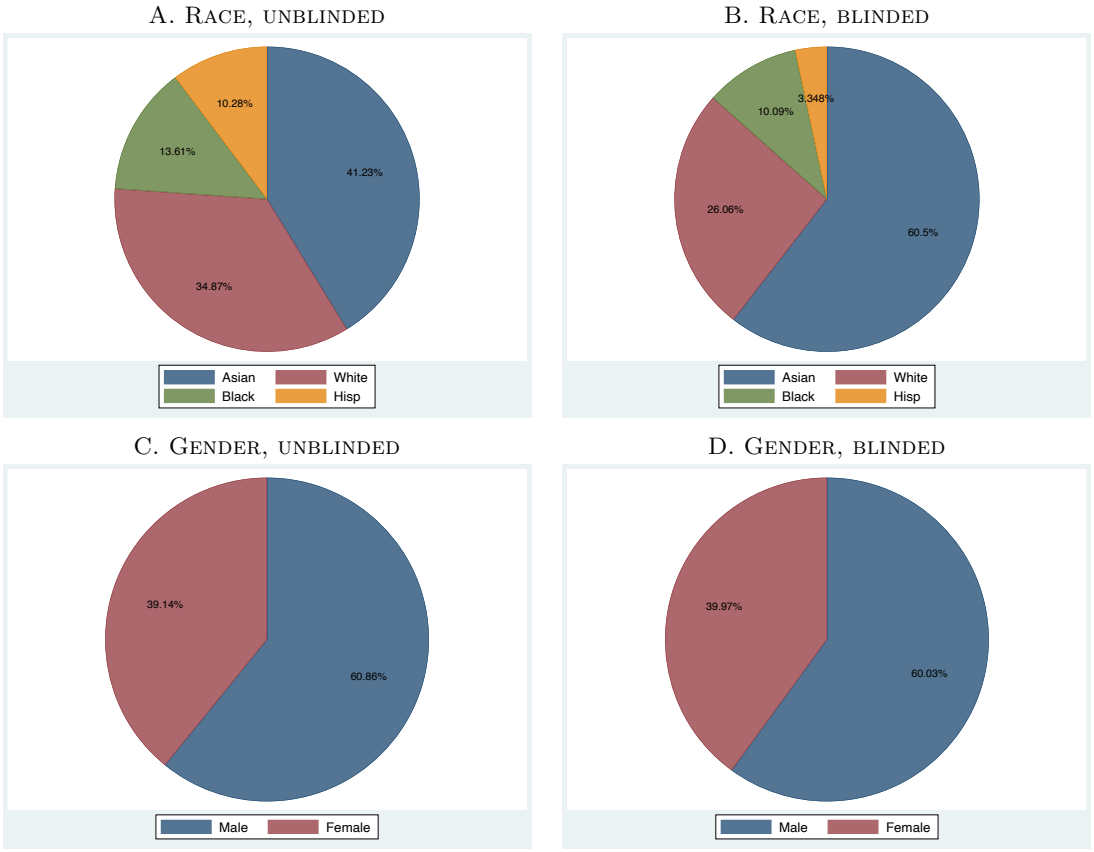
NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics are estimated separately for applicants in the top tercile of the supervised-learning algorithm's score and for applicants in the lowest tercile of the supervised-learning algorithm's score. The y axis in each panel is the magnitude of the effect. The outcome for Panel A is an indicator for being hired. The outcome for Panel B is an indicator for being hired and female. The outcome for Panel C is an indicator for being hired and Black. The outcome for Panel D is an indicator for being hired and Hispanic. The confidence intervals shown in each panel are derived from robust standard errors clustered at the recruiter level.

FIGURE 6: DYNAMIC UPDATING, INCREASED QUALITY



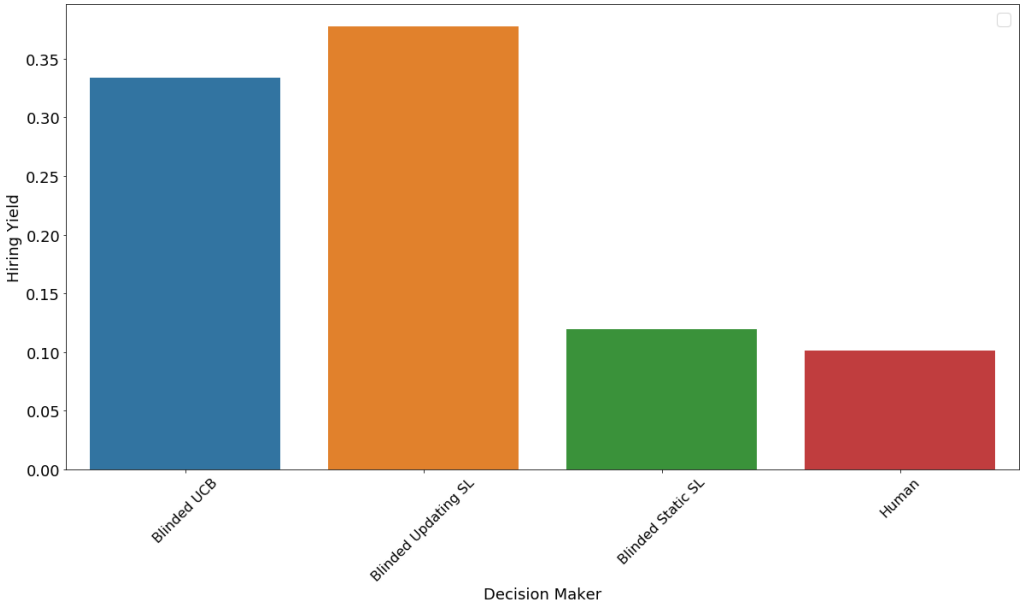
NOTES: This figure shows the racial composition of applicants recommended for interviews under by the UCB algorithm, UCB without the exploration bonus (exploitation component only), updating supervised learning and a static version of supervised learning, when quality of a given race dramatically improves. The graph plots each algorithm's belief about quality for a fixed set of applicants from 2019 over time as each algorithms learns about improved quality of applicants in 2018. Panel A shows the percent of 2019 applicants selected to be interviewed who are Black when all Black applicants in 2018 are hired and interviewed. Panel B shows the same results for 2019 Hispanic applicants when the quality of Hispanic applicants in 2018 dramatically improves. Panels C and D repeat the same procedure for White and Asian applicants. See text for details on construction of algorithm and training process. All data come from the firm's application and hiring records.

FIGURE 7: DEMOGRAPHICS BLIND VS. AWARE: UCB MODEL



NOTES: This figure shows the race and gender composition of applicants recommended for interviews by the UCB algorithm when this algorithm explicitly incorporates race and gender in estimation (race and gender "unblinded") and when it excludes these characteristics in estimation (race and gender "blinded"). Panel A shows the racial composition of applicants recommended for an interview when the algorithm is unblinded. Panel B shows the racial composition of applicants selected for an interview when the algorithm is blinded. Panel C shows the gender composition of applicants selected for an interview when the algorithm is unblinded. Panel D shows the gender composition of applicants selected for an interview when the algorithm is blinded. See the text for details on the construction of the algorithm. All data come from the firm's application and hiring records.

FIGURE 8: AVERAGE HIRING LIKELIHOOD, FULL SAMPLE, BLINDED ALGORITHMS



NOTES: This figure shows the unconditional average hiring yield of each blinded algorithmic selection strategy and the human recruiters. The first bar shows average hiring yield for applicants selected by the blinded contextual bandit UCB algorithm, the second by the blinded updating supervised learning algorithm and the third bar plots the blinded static supervised learning version. The fourth bar shows the unconditional hiring yield for candidates selected by human recruiters. See text for details on the construction of the unweighted average. All data come from the firm's application and hiring records.

Table 1: Applicant Summary Statistics

Variable	Mean Training	Mean Test	Mean Overall
Black	0.09	0.09	0.09
Hispanic	0.04	0.04	0.04
Asian	0.57	0.59	0.58
White	0.30	0.28	0.29
Male	0.68	0.66	0.67
Female	0.32	0.34	0.33
Referred	0.14	0.11	0.13
B.A. Degree	0.23	0.24	0.24
Associate Degree	0.01	0.01	0.01
Master's Degree	0.61	0.64	0.63
Ph.D.	0.07	0.07	0.07
Attended a U.S. College	0.75	0.80	0.77
Attended Elite U.S. College	0.13	0.14	0.13
Interviewed	0.05	0.05	0.05
Hired	0.01	0.01	0.01
Observations	48,719	39,947	88,666

Notes: This table shows applicants' demographic characteristics, education histories, and work experience. The sample in column (1) is all applicants who applied to a business analyst or data scientist in the training data (2016 and 2017). Column (2) is comprised of all applicants in the test data (2018 to Q1 2019). Column (3) is comprised of all applicants (2016 to Q1 2019). All data come from the firm's application and hiring records.

Table 2: Correlations between algorithm scores and hiring likelihood

	Hired			
	(1)	(2)	(3)	(4)
Human	-0.0562**			
	(0.0277)			
Static SL		0.232***		
		(0.0304)		
Updating SL			0.214***	
			(0.0271)	
UCB				0.233***
				(0.0261)
Observations	2275	2275	2275	2275
Mean of DV: .102				

Notes: This table presents the results of regressing an indicator for being hired on the algorithm scores on the sample of interviewed applicants.. Control variables include are fixed effects for the job family, application month and year and seniority level. All data come from the firm's application and hiring records. Robust standard errors shown in parentheses.

Table 3: Algorithm Agreement

A. Human vs. Updating SL

Selectivity (Top X%)	Overlap %	Both	Human Only	SL Only
	(1)	(2)	(3)	(4)
25	11.89	12.30	7.74	20.13
50	36.30	11.51	6.37	17.07
75	66.29	10.55	4.61	17.29

B. Human vs. UCB

Selectivity (Top X%)	Overlap %	Both	Human Only	UCB Only
	(1)	(2)	(3)	(4)
25	13.33	16.23	6.38	21.41
50	34.22	11.00	7.14	17.89
75	60.83	11.00	4.32	16.31

C. Updating SL vs. UCB

Selectivity (Top X%)	Overlap %	Both	SL Only	UCB Only
	(1)	(2)	(3)	(4)
25	47.12	24.18	8.25	13.11
50	58,25	16.43	7.64	8.64
75	76,49	12.94	5.26	7.89

Notes: This table shows the hiring rates of each algorithm when they make the same recommendation or differing recommendations. The top panel compares the human versus updating SL algorithm, the middle panel compares the human versus the UCB algorithm, and the lower panel compares the updating SL versus the UCB algorithm. Each row of a given panel conditions on selecting either the top 25%, 50%, 75% of applicants according to each of the models. For the two algorithms being compared in a given panel, Column 1 shows the percent of applicants selected by either algorithm that both agree on. Column 2 shows the share of applicants hired when both algorithms recommend an applicant, and Columns 3 and 4 show the share hired when applicants are selected by only one of two algorithms being compared. See text for the exact specification of each algorithm. All data come from the firm's application and hiring records.

Table 4: Correlations between Human Scores and On the Job Performance

A. Human Scores

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
Human SL Score	-0.288**	-0.309**	-0.0707	-0.0934
	(0.121)	(0.124)	(0.0756)	(0.0772)
Observations	180	180	233	233
Controls for ML Scores		X		X

B. Static SL Scores

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
Static SL	0.0144	0.0161	0.0462	0.0494
	(0.113)	(0.108)	(0.0598)	(0.0613)
Observations	180	180	233	233
Controls for Human SL		X		X

C. Updating SL Scores

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
Updating SL	0.0366	0.0780	0.0978	0.115*
	(0.112)	(0.102)	(0.0644)	(0.0693)
Observations	180	180	233	233
Controls for Human SL		X		X

D. UCB Scores

	Top Rating		Promoted	
	(1)	(2)	(3)	(4)
UCB Score	0.0903	0.0802	0.132**	0.132**
	(0.0941)	(0.0903)	(0.0589)	(0.0587)
Observations	180	180	233	233
Controls for Human SL		X		X

Notes: This table presents the results of regressing measures of on-the-job performance on algorithm scores, for the sample of applicants who are hired and for which we have available information on the relevant performance metric. "High performance rating" refers to receiving a 3 on a scale of 1-3 in a mid-year evaluation. Columns marked with an "X" reflect the added controls as indicated in the panel. Robust standard errors shown in parentheses.

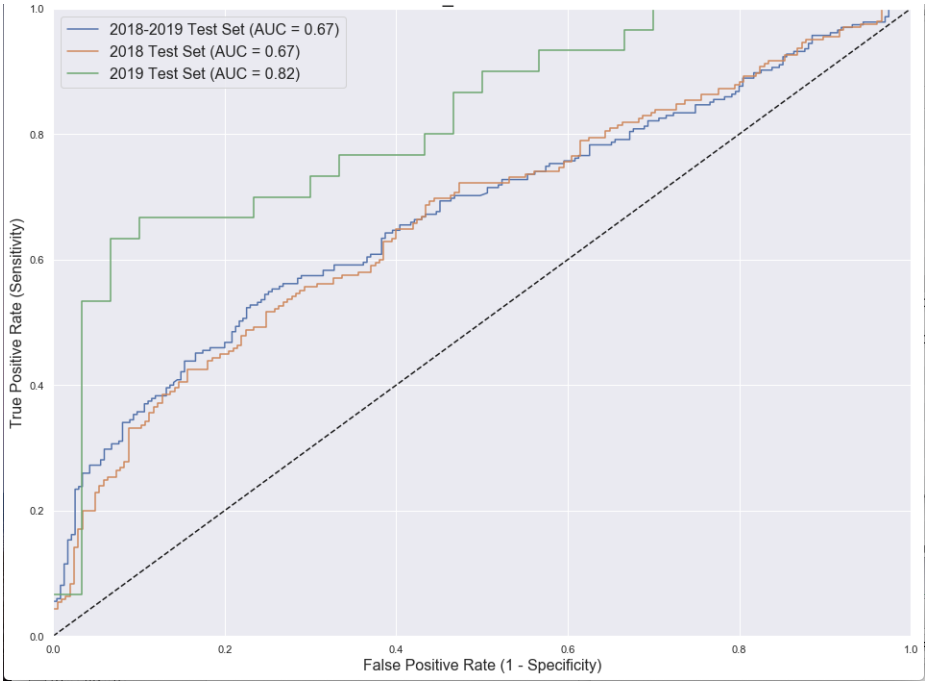
TABLE 5: INSTRUMENT VALIDITY

	Interviewed	Black	Hispanic	Asian	White	Female	Ref.	MA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
JK interview rate	0.0784***	0.000767	-0.000234	0.00812	-0.00939	-0.000348	0.00987	-0.00888
	(0.00881)	(0.00439)	(0.00221)	(0.0108)	(0.00740)	(0.00461)	(0.00814)	(0.0104)
Observations	26281	26281	26281	26281	26281	26281	26281	26281

NOTES: This table shows the results of regressing the baseline applicant characteristics and an indicator for being interviewed on the instrumental variable, which is the jack-knife mean-interview rate for a recruiter assigned to an applicant, while controlling for the job family, management level, application year and location of the job opening. This leave out mean is standardized to be mean zero standard deviation one. The outcome in column (1) is an indicator variable for being interviewed. The outcomes in columns (2)–(7) are indicators for baseline characteristics of the applicant. The sample is restricted to recruiters who screened at least 50 applicants. All data come from the firm's application and hiring records. Standard errors clustered at the recruiter level shown in parentheses.

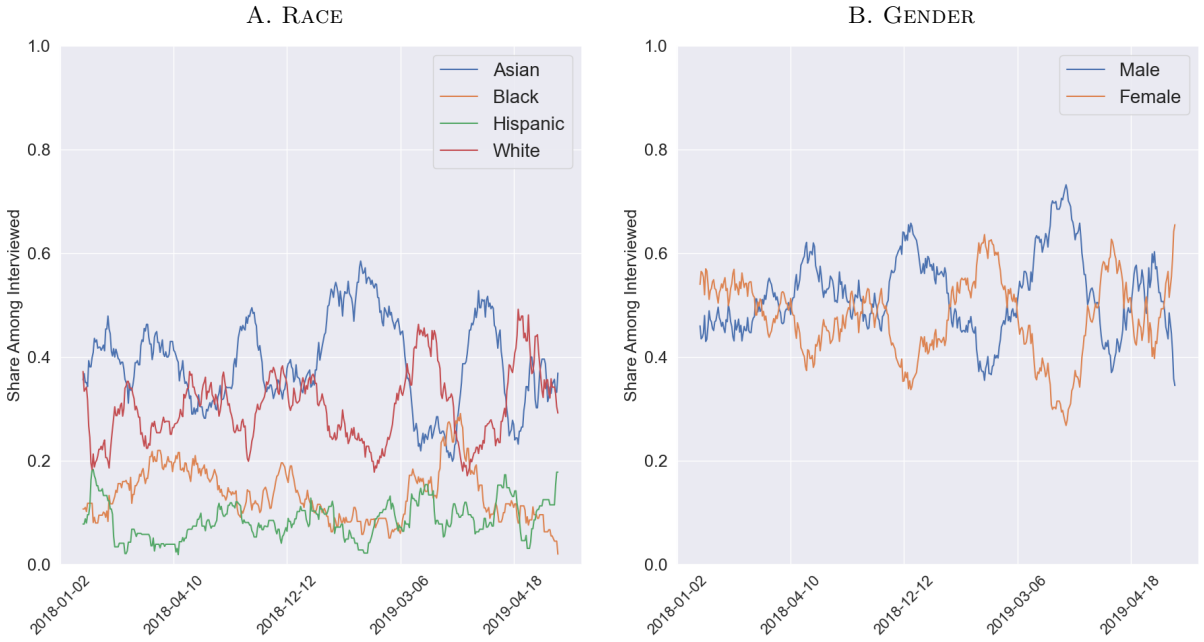
8 Appendix

FIGURE A.1: MODEL PERFORMANCE: PREDICTING HIRING, CONDITIONAL ON RECEIVING AN INTERVIEW



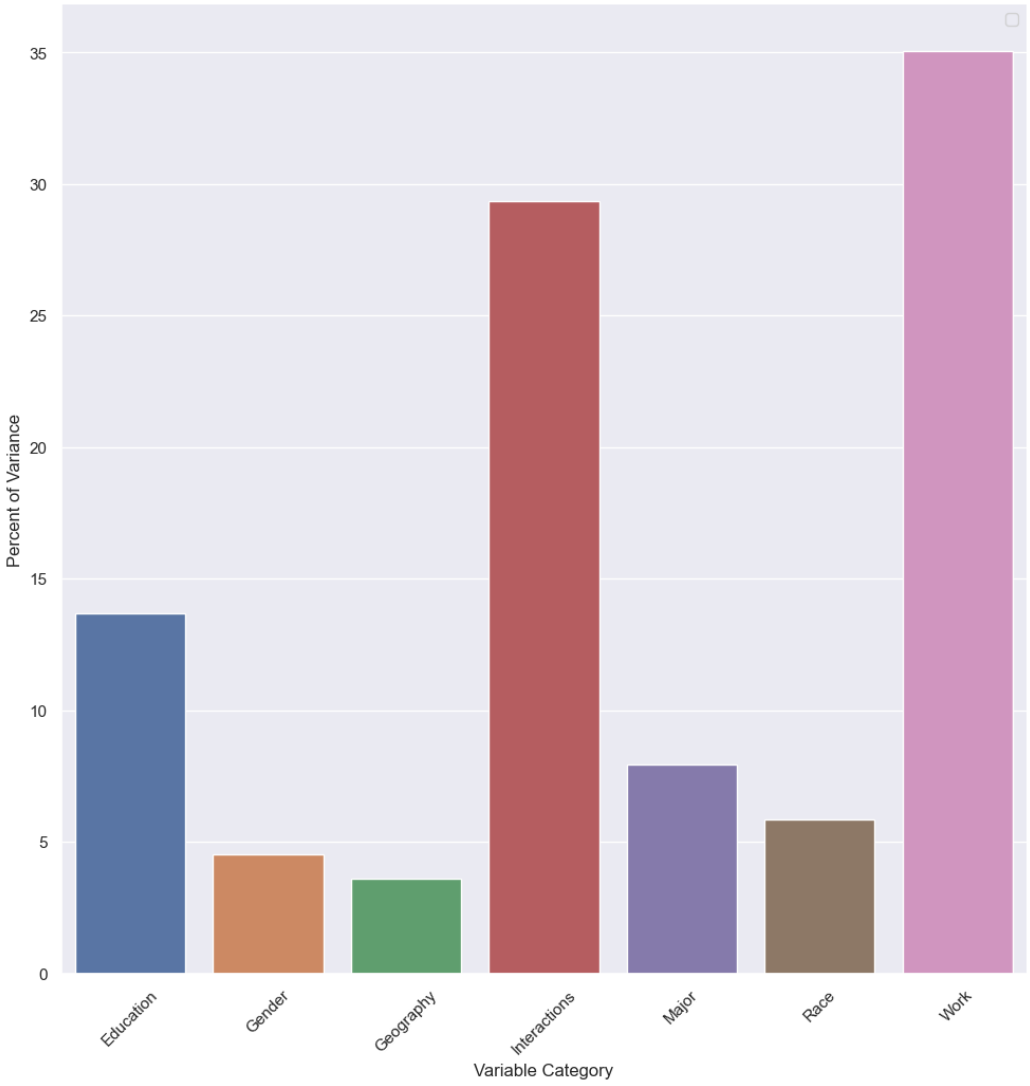
NOTES: This figure shows Receiver-Operating Characteristic (ROC) Curve for the baseline static supervised learning model, which predicts hiring potential. The ROC Curve plots the false positive rate on the x-axis and the true positive rate on the y axis. For each model, we plot this curve for different test data: the green line shows the ROC curve using data from 2019 year, the orange line uses data from the 2018 year, and the blue line uses data from both the 2018 and 2019 years. For reference, the 45 degree line is shown with a Black dash in each plot. All data come from the firm's application and hiring records.

FIGURE A.2: UCB COMPOSITION OF SELECTED CANDIDATES, OVER TIME



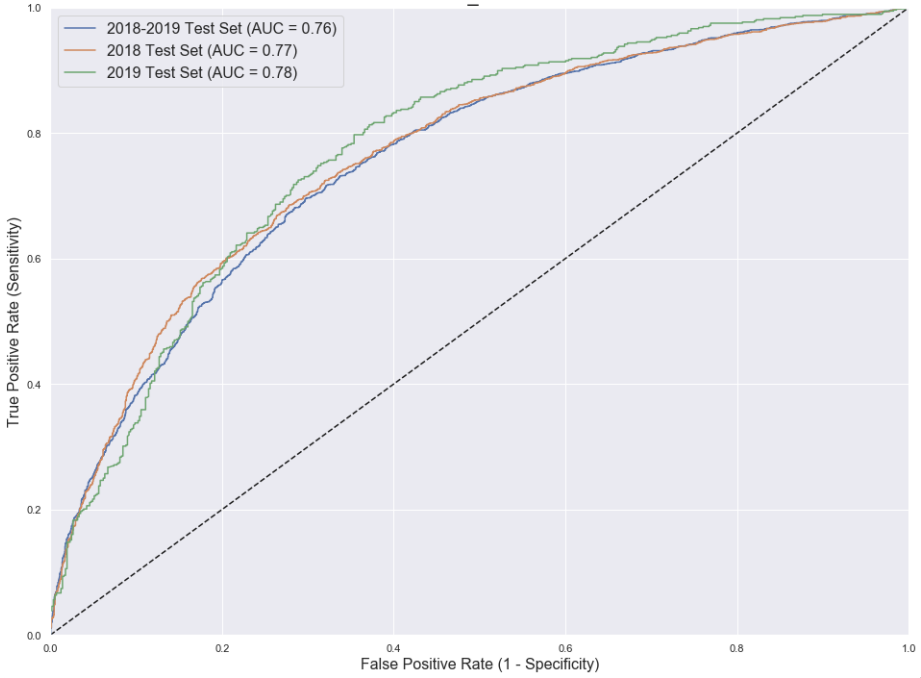
NOTES: This figure shows the composition of applicants selected to be interviewed by the UCB model at each point during the test period. Panel A focuses on race while Panel B focuses on gender.

FIGURE A.3: DRIVERS OF VARIATION IN EXPLORATION BONUSES



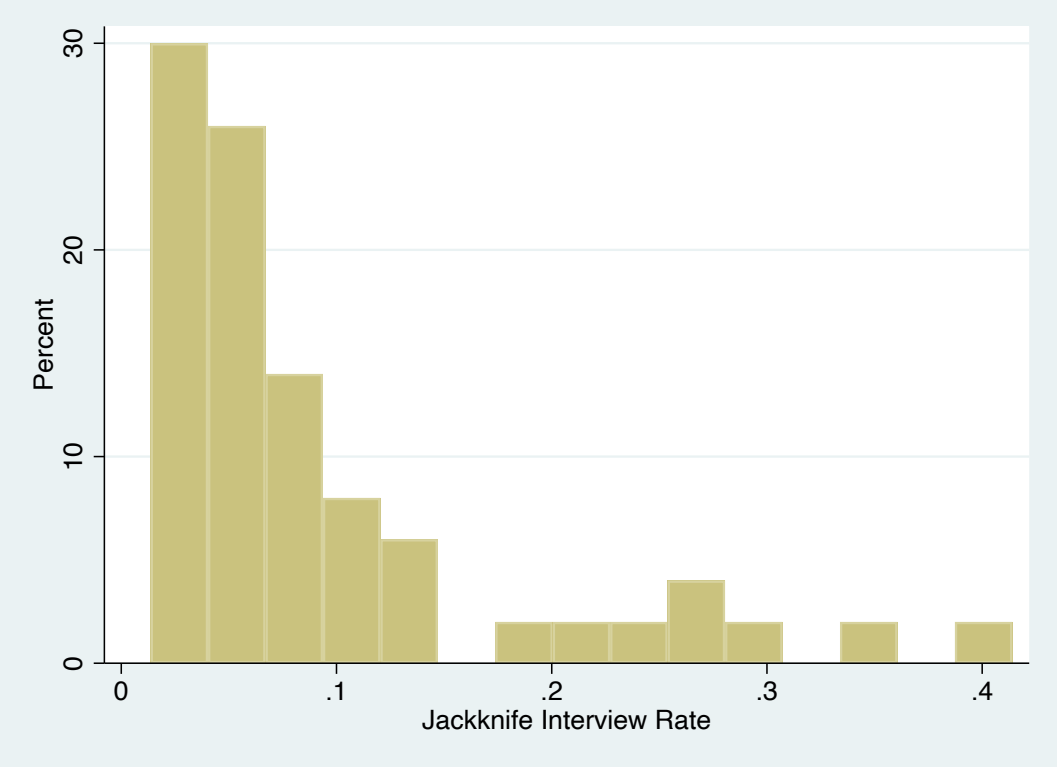
NOTES: This figure shows the percent of variance of applicant information in 2018-2019 attributed to each category of applicant characteristics. The education category refers to information on candidate educational background, such as degree and ranking of college attended. Geography captures the geographic location of educational experience, such as India, China or US, Major includes coding of majors for each educational degree above high school. The interactions category includes race and gender by degree and ranking of college or university interaction variables and the work category includes information on previous work experience.

FIGURE A.4: MODEL PERFORMANCE: PREDICTING INTERVIEW SELECTION



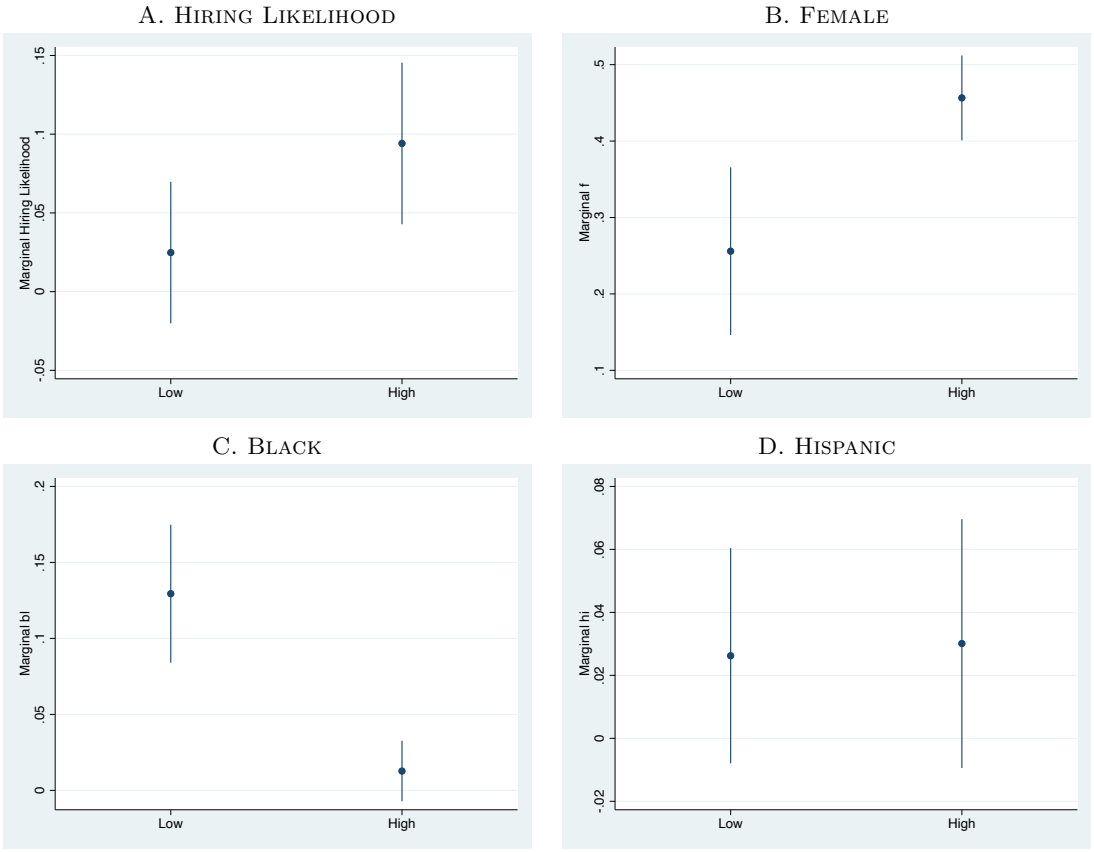
NOTES: This figure shows Receiver-Operating Characteristic (ROC) Curve for the human-decision making model, which predicts who is interviewed. The ROC Curve plots the false positive rate on the x-axis and the true positive rate on the y axis. For each model, we plot this curve for different test data: the green line shows the ROC curve using data from 2019 year, the orange line uses data from the 2018 year, and the blue line uses data from both the 2018 and 2019 years. For reference, the 45 degree line is shown with a Black dash in each plot. All data come from the firm's application and hiring records.

FIGURE A.5: DISTRIBUTION OF INTERVIEW RATES



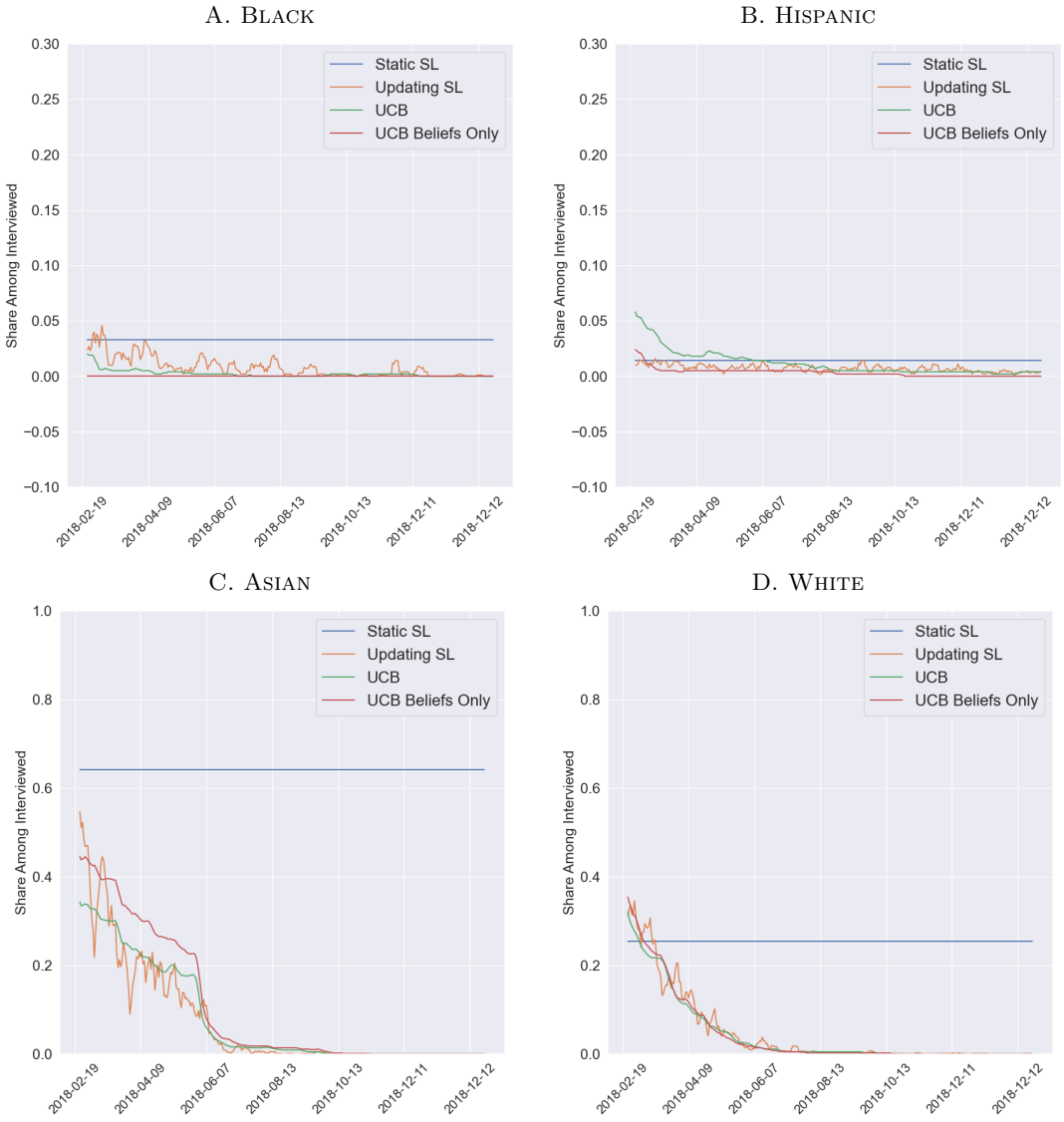
NOTES: This histogram shows the distribution of jack-knife interview rates for the 54 screeners in our data who evaluate more than 50 applicants. All data come from the firm's application and hiring records.

FIGURE A.6: CHARACTERISTICS OF MARGINAL INTERVIEWEES, BY UPDATING SUPERVISED SCORE



NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics is estimated separately for applicants in the top half of the supervised-learning algorithm's score and for applicants in the lower half of the updating supervised-learning algorithm's score. The y axis in each panel is the magnitude of the effect. The outcome for Panel A is an indicator for being hired. The outcome for Panel B is an indicator for being hired and female. The outcome for Panel C is an indicator for being hired and Black. The outcome for Panel D is an indicator for being hired and Hispanic. The confidence intervals shown in each panel are derived from robust standard errors clustered at the recruiter level.

FIGURE A.7: DYNAMIC UPDATING, DECREASED QUALITY



NOTES: This figure shows the racial composition of applicants recommended for interviews under by the UCB algorithm, UCB without the exploration bonus (exploitation component only), constantly updating supervised learning and a static version of supervised learning when quality of a given race decreases. The graph plots each algorithm's belief about quality for a fixed set of applicants from 2019 over time as each algorithms learns about reduced quality of applicants in 2018. Panel A shows the percent of 2019 applicants selected to be interviewed who are Black when all Black applicants in 2018 are interviewed but not hired. Panel B shows the same results for 2019 Hispanic applicants when the quality of Hispanic applicants in 2018 falls. Panels C and D repeat the same procedure for White and Asian applicants. See text for details on construction of algorithm and training process. All data come from the firm's application and hiring records.

TABLE A.1: MODEL PERFORMANCE: CONFUSION MATRICES

		Predicted Interview			Predicted Hire	
		not interviewed	interviewed		not hired	hired
Actual Outcome	not interviewed	70 %	30 %	not hired	70 %	30 %
	interviewed	30 %	70 %	hired	30 %	70 %

NOTES: This table shows the true negatives (top left), true positives (bottom right), false positives (top right) and false negatives (bottom left) for predicted interviews and predicted hires derived from the human-decision making algorithm in the test data. See text for the exact specification of the algorithm. All data come from the firm's application and hiring records.

TABLE A.2: APPLICANT FEATURES AND SUMMARY STATISTICS

Variable	Mean Training	Mean Test	Mean Overall
Black Applicants	0.08	0.08	0.08
Hispanic Applicants	0.04	0.04	0.04
Asian Applicants	0.52	0.54	0.53
White Applicants	0.27	0.25	0.26
Male Applicants	0.65	0.64	0.64
Female Applicants	0.31	0.33	0.32
Referred Applicants	0.14	0.11	0.13
B.A. Degree	0.24	0.25	0.24
Associate Degree	0.01	0.01	0.01
Master's Degree	0.60	0.63	0.62
Ph.D.	0.07	0.08	0.07
Attended a U.S. College	0.74	0.80	0.77
Attended Elite U.S. College	0.13	0.15	0.14
Worked at a Fortune 500 Co.	0.02	0.02	0.02
Has a Quantitative Background	0.23	0.27	0.25
Attended School in China	0.07	0.08	0.08
Attended School in Europe	0.05	0.05	0.05
Attended School in India	0.21	0.24	0.22
Attended School in Latin America	0.01	0.01	0.01
Attended School in Middle East/Africa	0.01	0.02	0.02
Attended School in Other Asian Country	0.02	0.02	0.02
Attended Elite International School	0.09	0.10	0.10
Attended US News Top 25 Ranked College	0.14	0.14	0.14
Attended US News Top 50 Ranked College	0.27	0.28	0.28
Military Experience	0.04	0.04	0.04
Number of Applications	3.5	3.8	3.5
Number of Unique Degrees	1.7	1.75	1.7
Number of Work Histories	3.8	4.0	3.9
Has Service Sector Experience	0.01	0.01	0.01
Major Description Business Management	0.17	0.15	0.17
Major Description Computer Science	0.14	0.13	0.14
Major Description Finance/Economics	0.14	0.13	0.14
Major Description Engineering	0.06	0.06	0.06
Major Description None	0.20	0.25	0.22
Observations	54,243	43,997	98,240

NOTES: This table shows more information on applicants' demographic characteristics, education histories, and work experience. The sample in column (1) is all applicants who applied to a business analyst or data scientist in the training data (2016 and 2017). Column (2) is comprised of all applicants in the test data (2018 to Q1 2019). Column (3) is comprised of all applicants (2016 to Q1 2019). All data come from the firm's application and hiring records.

TABLE A.3: EFFECT OF THE QUALITY OF APPLICANT POOL ON INTERVIEW DECISIONS

	All	Black	Hispanic	Asian	White	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Applicant Pool Quality	0.000249	-0.00304	-0.00630	0.00151	0.000137	0.000849
	(0.00125)	(0.00341)	(0.00557)	(0.00175)	(0.00258)	(0.00198)
Observations	39087	3171	1554	20685	10130	12868

NOTES: This table shows how the quality of the applicant pool affects the likelihood of being interviewed. Applicant Pool Quality is the jack-knife mean quality in each round according to the human interview model, standardized to be mean zero standard deviation one. This variable is regressed on an indicator for being interviewed or not, controlling for application month, job family, and management-level fixed effects. The sample in column (1) is all applicants, column (2) Black applicants, column (3) is Hispanic applicants, column (4) is Asian applicants, column (5) is White applicants, and column (6) is female applicants. All data come from the firm's application and hiring records.