Hiring as Exploration

Peter Bergman* Danielle Li † Lindsey Raymond ‡

May 31, 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 upside potential. Using data from professional services recruiting within a Fortune 500 firm, we show that this approach improves both the quality (as measured by eventual offer and acceptance rates) and the diversity of candidates selected for an interview relative to the firm's existing practices. The same is not true for traditional supervised learning based algorithms, which improve quality but select far fewer minority 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 hiring algorithms that are potentially both more efficient and equitable.

*Columbia University and NBER
†MIT and NBER
‡MIT

Algorithms have been shown to outperform human decision-makers across an expanding range of settings, from medical diagnosis to image recognition and game playing.[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, supervised learning tools can unearth predictive relationships that may be overlooked by human recruiters; 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 algorithmic fairness.

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 based on "reinforcement learning" are designed to solve dynamic prediction problems that involve learning from sequential actions: in the case of hiring, such algorithms value exploration because they can learn more efficiently by selecting candidates with covariates there is more uncertainty about. As such, a reinforcement learning-based hiring algorithm can potentially improve the quality of HR decisions over time precisely because it seeks out applicants that are under-represented in historical training datasets. To the extent that training data include

[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).

fewer applicants from traditionally disadvantaged groups, this may expand representation, even though diversity is not part of the algorithm's 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—100 for each opening it fills—and rejects 90% 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: 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 their "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 reinforcement learning—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 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.

[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.

Our third approach implements a classic 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 "rareness" 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 in 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 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 heterogenous on other dimensions—geography, education, and race, for instance—leading them to receive higher exploration bonuses on average.

Our second set of results show 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

[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.

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. While firms may in practice not have the space to hire, these results suggest that algorithms are better at selecting candidates who are more likely to receive and accept an offer. The firm could therefore at least 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

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

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.

Before continuing, we note that 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 in their roles, if hired. Our ability to address this concern is unfortunately limited by data availability: we observe job performance ratings for fewer than 200 employees, 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 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. Adopting ML tools that incorporate exploration 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 expense of virtually eliminating demographic diversity. This substantial difference in outcomes underscores the importance of building algorithms that value exploration.

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 black and hispanic quality 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 heterogenous 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: women, blacks, and hispanics 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]

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 (64%), asian (52%), or white (27%). Black and hispanic candidates comprise 12% 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 an crucial part of the hiring process. Openings for professional services roles are often inundated with applications: our firm receives approximately 100 applications for each worker it hires. Interview slots, by contrast, are scarce: because they are conducted by current employees who are diverted from other types of productive work, firms are extremely 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.

[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.

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. We observe on the job performance measures (ratings and promotions) for a small subset of applicants (180 hired applicants); this number is too small to train a model using, but 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. For now, 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

Firm's interview decision

We begin by introducing a simple model to guide our analysis. Each period t, the firm sees job applicants with covariates X_{it} and makes interview and hiring decisions over time. The firm would only like to interview candidates it would hire, and so an applicant's quality can therefore be thought of in terms of 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.

We assume that the firm receives a payoff of $H - c$ for each candidate it chooses to interview and zero if it passes on the candidate. The firm's objective function is therefore given by:

$$\sum_{t=0}^{\infty} \beta^t \max_{I_t \in \mathbf{I}_t} \sum_{i \in I_t} (E[H_{it}|X_{it}] - c_t) \tag{1}$$

In Equation (1), $\mathbf{I}_t$ denotes the set of all possible subsets of applicants in period t, with generic element I_t. Given a specific interview set I_t, we define, $I_{it} \in \{0, 1\}$ as the indicator function for whether applicant i is interviewed. After each period t, the firm learns H_{it} for each candidate it has chosen to interview.

Contextual bandit solutions

The problem described above is an example of a contextual bandit problem. The firm faces a sequence of applicants and, for each, must choose between one of two actions or "arms": interview or not. If the firm chooses to interview an applicant it receives a reward that is equal to the applicant's

quality, H_{it}, less the cost of interviewing that candidate, c_t. If the firm chooses not to interview that applicant, it receives zero. Unlike multi-arm bandits in which the relationship between action and reward is invariant, the firm here has information about the applicants' covariates X_{it}. These variables provide "context" that can inform the expected returns to an action. For instance, an applicant's educational background can serve as a signal of the applicant's hiring potential. As the firm makes interview decisions over time, it accumulates information about the returns to interviewing candidates with different covariates.

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).

The firm's generic solution to the problem presented in Equation (1) is given by:

$$I_{it} = \mathbb{I}(s_t(X_{it}) > c_t).$$

Here, $s_t(X_{it})$ can be thought of as a score measuring the expected benefit of interviewing a candidate with covariates X_{it} at time t. This score can be a function of the candidate's characteristics, but also the history of prior interview decisions, hiring outcomes, and candidate characteristics.

A simple class of potential solutions to bandit problems are given by so-called "greedy" algorithms, which always choose to "exploit" the expected reward for pulling a given arm. In our setting, this would be equivalent to scoring candidates based on their conditional likelihood of being hired, $s_t(X_{it}) = P(H_{it} = 1|X_{it})$ and interviewing candidates if $P(H_{it} = 1|X_{it}) > c_t$. This is precisely what a decision-rule based on supervised learning would accomplish: if the scoring function $s_t(X_{it}) = P(H_{it} = 1|X_{it})$ is estimated once and time invariant thereafter, then this decision rule would correspond to our "static SL" model; if $s_t(X_{it}) = P_t(H_{it} = 1|X_{it})$ is re-estimated in each period t to incorporate new data, then this is equivalent to our "updating SL" model.

It is widely known, however, that greedy algorithms are inefficient solutions to contextual bandit problems. Rather, optimal solutions incorporate exploration (Dimakopoulou et al., 2018).[11] In practice, it is often computationally challenging to implement an optimal solution to a given contextual bandit problem (Kendrick et al., 2014). Rather, a growing field of computer science seeks to develop computationally tractable models that achieve solutions which are close to optimal (Bubeck and Cesa-Bianchi, 2012; Sutton and Barto, 2018). Solutions are characterized by a bound

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

on expected regret, i.e. the expected difference between the maximum possible accumulated reward and the actual accumulated reward.

In this paper, we follow Li et al. (2017) and implement a heuristic solution: a generalized linear model version of the Upper Confidence Bound (UCB) algorithm. Under this approach, a candidate is assigned to the arm with the highest expected reward *plus* an exploration bonus. UCB assigns the bonus proportional to the standard error of this estimated reward. In our setting, this is equivalent to assigning a candidate to an interview based on an estimate of $P(H = 1|X)$ and its standard error. This approach implies that if a candidate's characteristics yield a noisy estimate of $P(H = 1|X)$—for instance, because these candidates had rarely been interviewed in the past—the algorithm is more likely to assign them to the interview arm. This "optimism in the face of uncertainty" allows the algorithm to learn and generate more precise estimates of these candidates' hiring potential in the future. 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).

In the next section, we construct three machine learning algorithms aimed at solving the firm's problem, as presented in Equation (1): the first two implement standard supervised learning approaches akin to the greedy solution described above, and the final one implements a UCB approach that incorporates a value of dynamic exploration.

3 Algorithm Construction

We compare outcomes associated with three different ML interview selection policies. Policy $I^{S_0} \in \{0,1\} = \mathbb{I}(s^{S_0}(X) > c_t)$ is based on using supervised learning to predict a candidate's likelihood of being hired conditional on being interviewed given a fixed training dataset. Policy $I^{S_t} \in \{0,1\} = \mathbb{I}(s^{S_t}(X) > c_t)$ takes the same model up continually updates its training data based on the candidates it selects. Policy $I^{UCB}(X) \in \{0,1\} = \mathbb{I}(s^{UCB}(X) > c_t)$ is based on using an exploration-based approach to learn about workers' hiring likelihoods. In our comparisons, we will vary the interview cost c_t to make sure that each algorithm selects the same number of applicants as the human recruiter actually did, reflecting the constrained number of interview slots.

3.1 Training Data

We have data on 100,008 job applications from January 2016 to March 2019, as described in Table 1. We divide this sample up into a training dataset consisting of 54,243 applicants that arrive before 2018, 2,940 of whom receive an interview, and a test dataset of 43,997 applications that

arrive afterward, 2,411 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 60% 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:

$$I_t^{S_0} = \mathbb{I}(s^{S_0}(X) > c_t), \text{ where } s^{S_0}(X) = P(H = 1|X; D_0).$$

$P(H = 1|X; D_0)$ is the algorithm's predicted hiring likelihood for applicants with covariates X, where the model is trained on outcomes for the set of applicants interviewed between 2016-2017, denoted D_0 (we restrict our training data to the subset of interviewed applicants for whom we are able to observe their potential hiring outcome).

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. We estimate $s^{S_0}(X)$ 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. 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 $\Pr(\mathbb{I}(s(X_i) > c_t) > \mathbb{I}(s(X_j) > c_t) > |I_i = 1, I_j = 0)$.[12] Our model has an AUC of .67, meaning that it will rank an interviewed applicant who is hired higher than

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

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.[13]

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, 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

$$I_t^{S_t} = \mathbb{I}(s^{S_t}(X) > c_t), \text{ where } s^{S_t}(X) = P(H = 1|X; D_t^S).$$

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")

Both SL models described above are myopic in the sense that they maximize current period predicted quality, which is based on the current period's training data: the algorithms scoring function $s^{S_t}(x)$ is equal to its beliefs about quality $P(H = 1|X)$. That is, it does not factor the ex post value of learning into its ex ante selection decisions. In contrast, our final model makes selection decisions placing value on an applicant's direct interview performance, as well as the option value to learn more about the interview performances of similar future applicants. To balance exploitation and exploration, we choose the action the maximizes the sum of the estimated quality and its standard error, which can be interpreted as the upper confidence bound of an applicant's

[13] 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.

predicted quality (Li et al., 2017). This criterion for action selection can be interpreted as the additive trade-off between the payoff estimate and model uncertainty reduction (Li et al., 2017).

Specifically, we implement the UCB-GLM algorithm as described in Li et al. (2017). This version of a contextual bandit algorithm is an example of the widely used upper confidence bound (UCB) approach to efficient exploration first introduced in Lai and Robbins (1985). UCB algorithms are efficient and provably optimal in classic multi arm bandits and for generalized linear models (GLM) in the contextual bandit setting (Auer, 2002; Li et al., 2017; Abbasi-Yadkori et al., 2019). In the UCB-GLM case, regret over T rounds is of order $\tilde{\mathcal{O}}(d\sqrt{T})$ where d is the number of covariates and T is the number of rounds, which approaches the $\sqrt{dT}$ lower bound of expected regret for the K-armed linear bandit problem (Li et al., 2017).

We calculate predicted quality $P(H = 1|X; D_t^R)$ 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. There, by construction, $s^{S_t}(X; D_0) = s^{UCB}(X; D_0)$. Over time, however, the UCB learning model is updated based on new data from the applicants it selects, who may be different than those selected by the supervised-learning model.

At the same time, we also calculate exploration bonuses based on the 95th percent confidence interval for an applicant with covariates X.

$$B^R(X; D_t^R) = \alpha X' V_t^{-1} X, \text{ where } V_t = \sum_{i=1}^{t-1} X_t X_t'.$$

Importantly, exploration bonuses are chosen by the algorithm on the basis of any of the model's 106 feature inputs. The point estimate of the expected reward, estimated using the logistic regression, is associated with a range of values for the parameter and all information about the applicant contributes to this calculation. This means that there is no ex-ante sense in which the algorithm is required to grant higher bonuses on the basis of race or gender; the algorithm could instead provide bonuses on the basis of education, geography, or work history—if it does choose to favor applicants of a particular race or gender, this choice is made in a data-dependent fashion; it is not ex ante forced to do so.

As before, we take all applicants the algorithm has chosen and add their outcomes to the data used to train the model in the next period, so that $D_{t+1}^{UCB} = D_t^{UCB} \cup (I_t^{UCB} \cap I_t)$, subject to the limitation that we can only add applicants who are selected by the model and 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 beliefs and bonuses: beliefs update based on outcomes for newly selected applicants and bonuses update based on the precision of estimates within the group.

Human Decision-making

Finally, in addition to the models described above, we also train a model that predicts an applicant's actual interview outcome as chosen by a human recruiter. We will use this model in Section 4.2 to benchmark the predictions of our main algorithms against that of the human recruiter. Creating a model of human decision-making—as opposed to simply using observed interview outcomes I—is useful in some parts of our analysis because it allows us to proxy for variation in recruiter's preferences among applicants who are all interviewed.

To build this model, we use the same functional form specification (an ensemble model with a lasso and random forest model) as we do when constructing our main supervised learning model.[14] Because we are predicting whether or not an applicant is interviewed, we expand our sample to all applicants, not just those who are interviewed.[15]

4 Main Results

4.1 Impacts on Diversity of Interviewed Applicants

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.

In our test data, 54% of applicants are asian, 25% are white, 8% are black, and 4% are hispanic. Figure 1 shows the composition of selected applicants averaged over our test period sample. Panel A considers the composition of those chosen by human recruiters: the proportion of asians and

[14] By choosing the same model structure for both interview and hire, we avoid differences in predictive accuracy driven by choice of machine learning algorithm.

[15] 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. In the appendix, we report the AUC separately for each test sample year, as well as a standard confusion matrix which shows the percentage of correctly classified candidates and the number of type I and II errors (assuming that our model "selects" the top X candidates to be interviewed, where X is the same as the actual number of interviewees). We correctly classify candidates 70% of the time.

hispanics stays the same at 57% and 4%, respectively, and the proportion who are white rises to 34%, largely at the expense of black applicants, whose representation falls in half to just under 5%. Panels B and C display the counterfactual interview choices of our SL models. Panel B shows that a static SL algorithm would increase the proportion of white applicants who are interviewed from 34% to 51%; the majority of this increase comes at the expense of asian applicants whose representation falls from 57% to 47%. By proportion, however, we see the greatest declines among black and hispanic applicants, whose combined representation falls from 10% to less than 3%. The updating SL model (Panel C) follows a similar pattern; white representation increases more modestly from 34% to 40%; asian representation stays largely the same, but black and hispanic representation still falls dramatically from 10% to under 5%.

Panel D presents our UCB results. Here, we see a substantial increase in the proportion of interviewed candidates who are black or hispanic: the black share rises from 5% to 14% while the hispanic share rises from 4% to 10%. The white share stays constant and these gains largely come at the expense of Asian applicants, whose representation is cut from 57% to 41%.

Figure 2 plots the same set of results, for gender. Among those who report a binary gender (97% of the sample), the majority of applicants in our sample are male (64%); Panel A shows that the interviewed class maintains this same composition with 65% men and 35% women. All of our ML models would interview women at a higher rate than men, increasing their representation among those selected to 41% (static SL), 50% (updating SL), and 39% (UCB). Unlike race and ethnicity, our supervised and reinforcement algorithms are aligned. This suggests that, in the historical training data, women who were interviewed were more likely to be hired than men. Although there are more men than women in our training data, the UCB model appears to assign higher exploration bonuses to men, indicating that men are likely more heterogenous on other dimensions such as education, race, or work history.

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.

First, we simply restricts to the set of applicants who are interviewed for whom we observe hiring outcomes. Among this set, we ask whether applicants ranked highly by one interview policy are more likely to be hired than those ranked highly by another. Across a range of specifications, our results will show that all ML algorithms perform better than human screening practices.

A concern with this type of approach is that the relative performance of ML algorithms among the selected set of interviewed applicants may not be representative of its performance among the applicant pool at large. Our second approach therefore attempts to infer the quality of ML-selected applicants who are not actually interviewed in order to provide an estimate of $E[X|I^{ML} = 1]$ that applies to the full set of applicants. To do this, follow a decomposition-reweighting approach based on DiNardo et al. (1996). Again, we find that all ML models outperform humans.

A concern with our decomposition approach is that it assumes no selection on unobservables. To address this, our final approach uses random variation in receiving an interview (leniency of randomly assigned recruiters) to identify the quality returns to more closely following algorithmic recommendations *on the margin*. That is, rather than asking whether ML models perform better on the full sample of applicants, we ask whether they perform better for applicants on the margin of receiving an interview. If so, then the firm can improve quality outcomes by following algorithmic recommendations for marginal cases. Following the logic outlined in Abadie (2003b) and Kling (2006), we identify the characteristics of marginally interviewed candidates using random assignment to initial screeners as an instrument for receiving an interview. We then explicitly construct an alternative interview policy that improves both the diversity and quality of selected workers by following UCB learning recommendations for marginal applicants. Given a valid instrument, this analysis is robust to the presence of selection on unobservables.

Interviewed sample

In this section, we compare the quality of applicants selected by our algorithms among the sample of applicants who are interviewed. Because all applicants in this sample are, by construction, selected by human recruiters, we cannot directly compare the accuracy of algorithmic to human choices within this sample. To get around this, we construct another algorithm designed to predict human interview decisions (rather than hiring likelihood) using the standard LASSO approach described in Section 3.2. 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}.[16] 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.[17]

Figure 3 plots a binned scatterplot depicting the correlation 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. Panel A focuses on applicants as ordered by their human-predictive model score, s^H. Table 3 shows these results as regressions to test whether the relationships are statistically significant. We see a slightly negative correlation between human model scores and an applicant's actual likelihood of being hired, but this disappears when we include controls in the model (column (2)). In contrast, Panels B and C show large and statistically significant positive relationships between algorithmic priority selection scores and an applicant's (out of sample) likelihood of being hired. Interviewed applicants with algorithmic scores in the bottom half the distribution have hiring rates of less than 10%, while those in the upper half of the distribution have hiring rates of closer to 20%. All of our algorithms are particularly good at identifying interviewed candidates with the greatest hiring likelihoods: over 30% of interviewed applicants with algorithmic scores in the top decile end up being hired.

Figure 5 examines how much these algorithms agree on whom to interview, and which model is correct when they disagree. The top panel looks at applicants selected from the top 25% of the algorithms. All three algorithms agree on only 6% of candidates (left diagram). Most of this discrepancy is from the human-interview model; the UCB and updating SL models agree 31% of the time. Moreover, the right column of the top panel shows that the human model performs substantially worse in terms of predicting hiring likelihood when the models disagree. For the human versus UCB model: only 13% of candidates favored by the human model are eventually hired compared to 40% of candidates favored by the algorithm. For the human versus updating SL model, the same share (13%) of candidates favored by the human model are eventually hired compared to 34% of candidates favored by the algorithm. The remaining lower two panels of the figure show similar results for the top 50% and the top 75% of selected candidates.

Full sample with decomposition-reweighting

Our analysis above 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

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

[17] 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).

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 estimate the average quality of all ML-selected applicants, including those who are not interviewed. To do this, we infer hiring likelihood using data on hiring outcomes among applicants with similar covariates who did not. Doing so requires us to assume that there is no selection on unobservables in our sample, which implies that $E[H|I^{ML}=1,X] = E[H|I^{ML}=1,I=1,X]$. We believe that this is a plausible assumption 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.

We are interested in recovering the unconditional mean $E[H|I^{ML}=1]$, given observed data on $E[H|I^{ML}=1,I=1,X]$. We can write:

$$E[H|I^{ML}=1] \quad = \quad \sum_X p(X|I^{ML}=1)E[H|I^{ML}=1,X]$$

In general, the term $p(X|I^{ML}=1)$ is difficult to estimate empirically. We follow DiNardo et al. (1996) and transform this using Bayes Rule: $p(X|I^{ML}=1) = \frac{p(I^{ML}=1|X)p(X)}{p(I^{ML}=1)}$. Given this, we can write:

$$E[H|I^{ML}=1] \quad = \quad \sum_X \frac{p(I^{ML}=1|X)p(X)}{p(I^{ML}=1)} E[H|I^{ML}=1,X]. \tag{2}$$

Equation (2) is compromised mostly of easily observed components: the unconditional distribution of covariates, $p(X)$, the conditional probability of being selected by an ML model, $p(I^{ML}=1|X)$, and the unconditional probability of selection, $p(I^{ML}=1)$. The term that is most difficult to observe is $E[H|I^{ML}=1,X]$ because we observe only $E[H|I^{ML}=1,I=1,X]$—outcomes only among the set of applicants who are actually interviewed. However, assuming no selection on unobservables, we note that $E[H|I^{ML}=1,I=0,X] = E[H|I^{ML}=1,I=1,X]$ so that we can infer the hiring likelihood of ML-selected applicants who are not interviewed using observed hiring outcomes for applicants with similar covariates who are interviewed.

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. 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).

Figure 4 shows our results. We consider an interview policy that selects the number of candidates as were actually interviewed by human recruiter. Among those selected, the average observed hiring likelihood, as selected by human recruiters, is 10%. In all cases, our 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.

Comparing within the ML models suggests that there are substantial returns to increasing the size of the training data we use, even when the newly added applicants are not necessarily that under-represented (as in the case of the updating SL model): both dynamic models do better than the static SL model. Importantly, though there do not appear to be substantial differences in overall hiring likelihoods between the updating SL and UCB models, even though, as discussed earlier, these models differ markedly in the diversity of the candidates they select. 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. We will investigate this possibility in Section 5.

Following algorithmic recommendations on the marginal 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. 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. The downside of this, however, is that we are able to control for relatively few of the covariates we actually observe.

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: $E[H|I^{Z=1} > I^{Z=0}]$.[18]

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:

$$\tilde{I} = \begin{cases} I^{Z=1} \text{ if } s^{ML} > \overline{\tau},^{19} \\ I \text{ if } \underline{\tau} \leq s^{ML} \leq \overline{\tau}, \\ I^{Z=0} \text{ if } s^{ML} < \underline{\tau}. \end{cases}$$

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. 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 greater hiring

[18] 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 $E[H|\lim_{z' \downarrow z} I^{z'} = 1, \lim_{z' \uparrow z} I^{z'} = 0]$.

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.3 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 6 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.4 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.

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 43,000 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 for 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 particular, standard supervised learning based approaches reduce the representation of black and hispanic applicants by over 70% relative to baseline human decisions. A UCB-based approach that emphasizes exploration, meanwhile, more than doubles representation. These results raise several additional questions about the viability of our UCB approach, which we explore in extensions.

First, our results so far indicate that our UCB model presents a Pareto improvement in quality and diversity relative to both human learning and the static SL model. Further, it also represents a "close" to Pareto improvement relative to the updating SL model, dramatically increasing diversity

at a relatively small cost to quality. However, the fact that our quality estimates are higher for the updating SL model suggests that the value of active learning is limited in our setting, in contrast with theoretical bandit predictions. 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 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. Yet 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 the use of demographic information.

5 Learning over time

In this section, we examine the value of exploration bonuses in greater depth. Our results so far indicate that our UCB model presents a Pareto improvement in quality and diversity relative to both human learning and the static SL model. Further, it also represents a "close" to Pareto improvement relative to the updating SL model, dramatically increasing diversity at a relatively small cost to quality. However, the fact that our quality estimates are higher for the updating SL model suggests that the value of active learning is limited in our setting.

There are, however, reasons to believe that learning can play a more important role in other settings. First, we note that 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. While this constraint limits updating for both the SL and UCB models, it is more binding in the case of the UCB model. Intuitively, this makes sense because the UCB favors candidates that are under-represented in its training data, and is therefore more likely to select candidates who look different from what human recruiters have chosen in the past an who are precisely the applicants whose hiring outcomes are likely to be unobserved. Second, the degree of learning is limited by changes in the relation between context (covariates) and rewards (hiring likelihood) over time.

Because our test data cover a relatively short period (Jan 2018 to March 2019), there may be limited scope for this relationship to evolve, limiting the value of learning. In a world where the quality of applicants does change, models that encourage exploration may be better equipped to identify these changes.

In this section, we conduct simulations designed to address both of these issues. Specifically, we consider cases in which the quality of one group of applicants (by race) changes during the test sample. For example, we consider a case in which we assume that the quality of other candidates remains fixed but now assign a value of $H = 1$ to all black applicants. This increases the value of learning by changing applicant quality in the test period, relative to what the models were trained on. This approach makes it easier for our algorithms to update because we assume that $H = 1$ for these candidates so we are able to incorporate their outcomes into the model's new training data even if they were never interviewed in practice.

To examine how our models update over time, we consider the actual cohort of candidates who applied in 2019, and ask each model to evaluate this *same* cohort of candidates at different points in time throughout 2018. On Jan 1, 2018, all three ML algorithms would have the same beliefs about quality because they all share the same estimate of $P(H = 1|X; D_0)$ trained on the initial data D_0. If, instead they are evaluated on December 31, 2018, the static SL model would have the same beliefs as it did on January 1 (because its training set never updates), but the updating SL and UCB algorithms would have different beliefs, based on the different subsets of 2018 applicants that they selected, whose outcomes were added to their respective training datasets.

Panel A of Figure 7, reports results from the exercise described above where we allow our algorithms to update based on simulated data that is identical to the set of actual 2018 applicants, except that all black applicants during this period are assigned a quality value of $H = 1$. Then, having allowed this process to happen, we consider how each algorithm would treat the fixed cohort of 2019 applicants. The y-axis represents the proportion of candidates selected from this fixed cohort who are black. The flat line, which hovers at just over 1%, represents the proportion of 2019 cohort applicants who would be selected by the static supervised algorithm if they arrived at time t between Jan 1, 2018 and December 31, 2019. 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).

The green line shows this same process using the updating SL model. It, too, manages to learn about the quality of black applicants, but 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. The same pattern can also be seen in Panel B of Figure 7, which plots the percentage of hispanics who are selected in the case where now all hispanic applicants in 2018 are all assigned high quality. 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 7 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 whites and asians because it is still attempting to explore the quality of other candidates. Eventually, however, its beliefs about the quality of whites and asians is sufficiently high that the exploration bonuses it grants to other groups are no longer large enough to make up for its certainty about the increased quality of asians or whites.

Finally, in Appendix Figure A.5, 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.

Together, these results show that the value of exploration is most apparent when the quality of applicants—particularly less commonly selected applicants—changes. When the quality of candidates from more represented groups changes, an updating SL model learns just as quickly.

More generally, any change in applicant quality illustrates the advantages that dynamic updating has relative to baseline static SL models, which are commonly used in commercial applications.

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). 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 asians or whites. 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 8 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 whites and asians. In the non-blinded model, whites and asians 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 asians are selected relative to whites (61% vs. 26%). In our data, this is likely to arise for two reasons. First, asians tend to be more likely to have a master's degree or above, a trait that is strongly rewarded for whites; blinding the algorithm to race increases the returns to education for asians. Second, in the race-aware model, asians received smaller exploration bonuses because they comprised a majority of the applicant pool; when bonus provision is blinded, exploration bonuses for asians 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 9 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 asians who are selected because it could see that asian applicants shared a covariate—race—that was very common in the sample. When the algorithm is no longer permitted to do this explicitly, the share of selected asians 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 an active learning 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 are, in part, inefficient at valuing equity: just as they are likely to pass up on high quality white male candidates in order to interview

lower quality white male candidates, they are also choosing to select weaker under-represented minorities ahead of stronger under-represented minorities.

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 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.

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.

Auer, Peter, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, 2002, *3* (Nov), 397–422.

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.

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, "ndustries 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.

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.

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.

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.

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.

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.

Kendrick, David A, Hans M Amman, and Marco P Tucci, "Learning about learning in dynamic economic models," in "Handbook of Computational Economics," Vol. 3, Elsevier, 2014, pp. 1–35.

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.

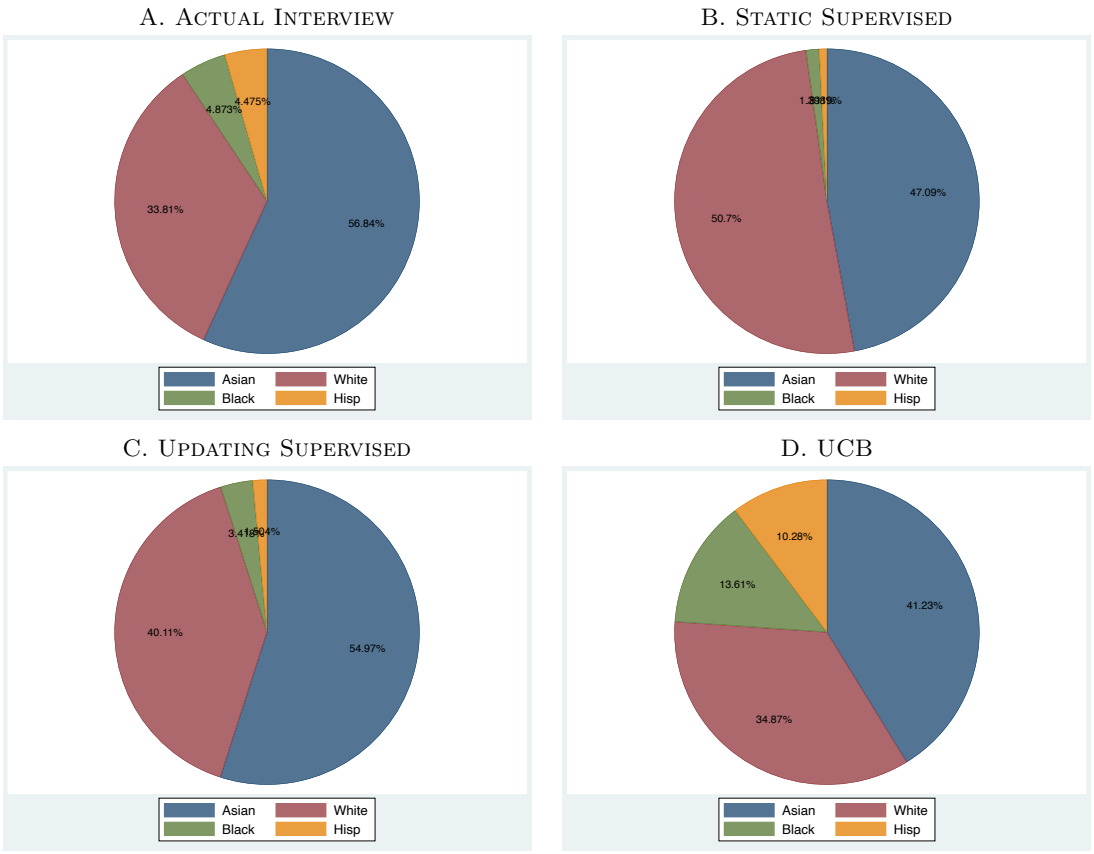
Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver, "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," 2019.

Sutton, Richard S and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

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.

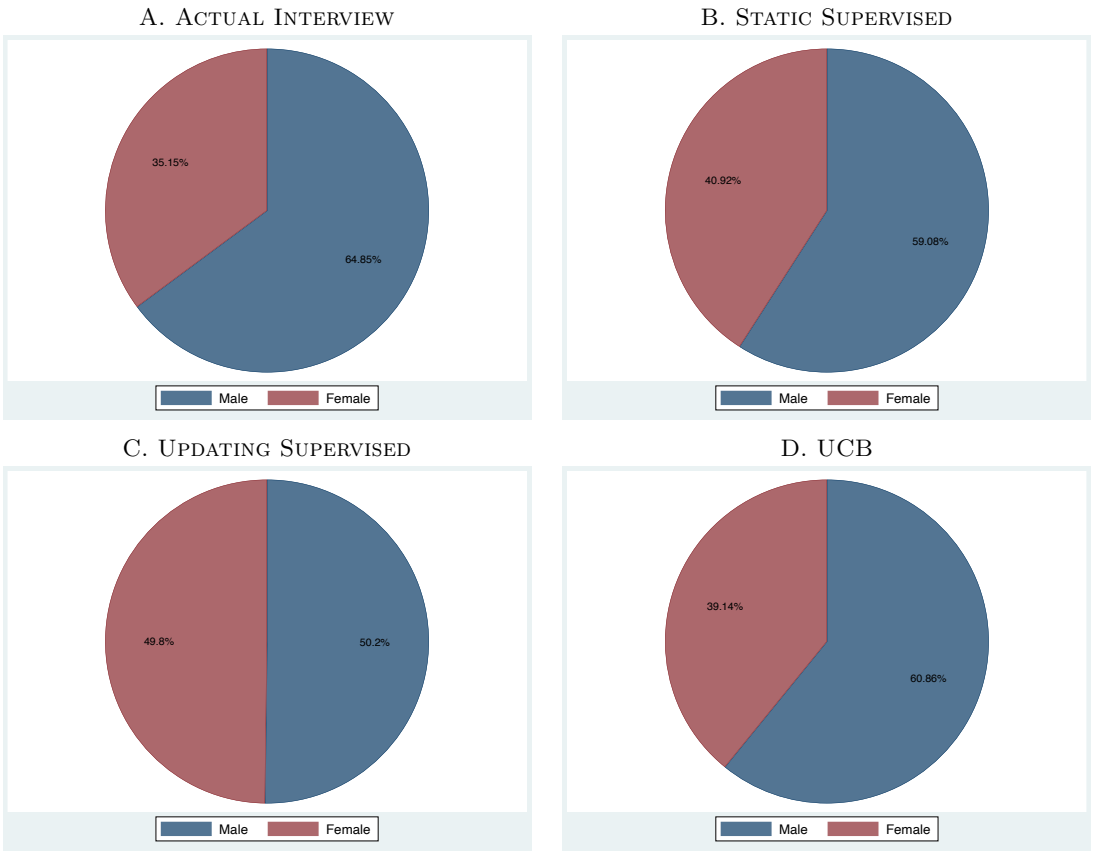
Yala, Adam, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay, "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction," *Radiology*, 2019, *292* (1), 60–66. PMID: 31063083.

Figure 1: Racial Composition



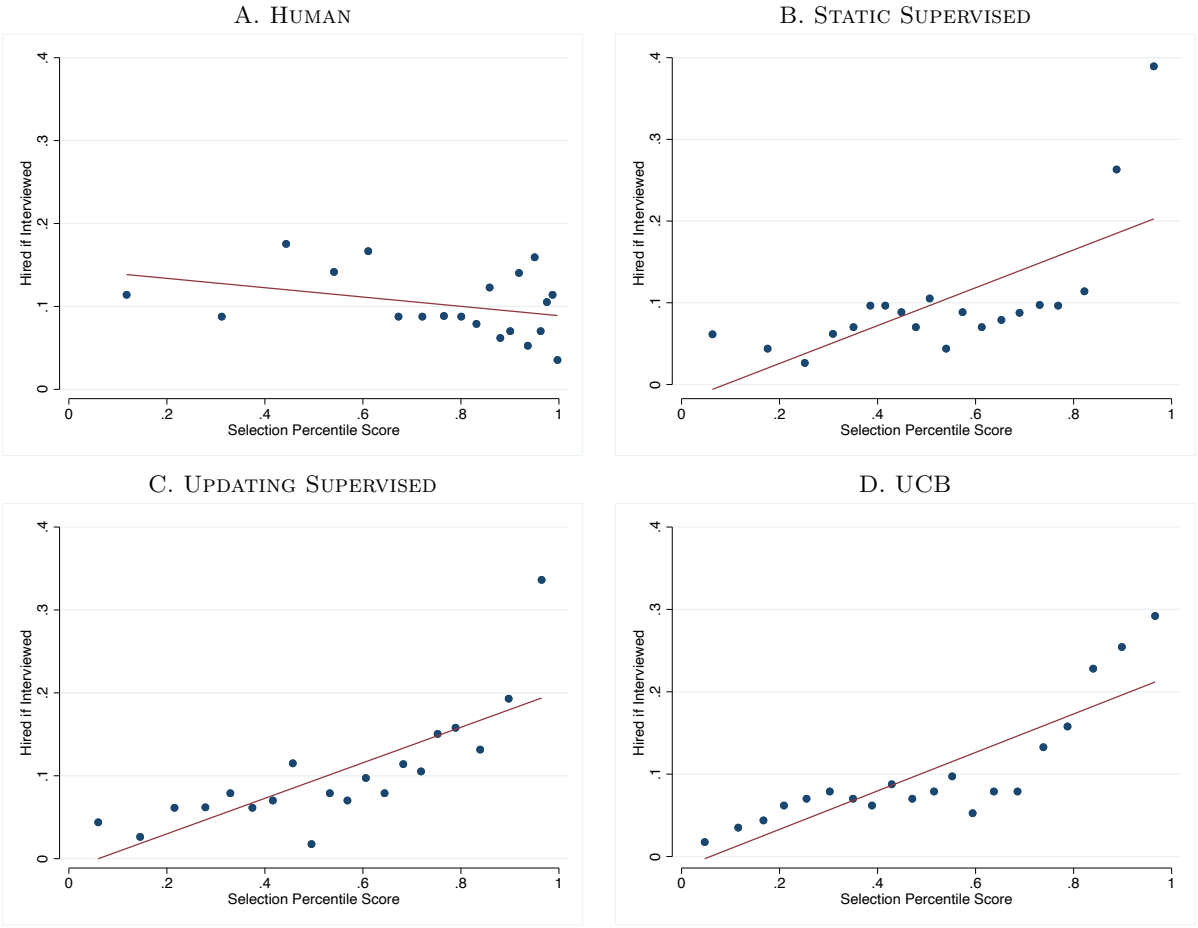
Notes: This figure shows the racial composition of applicants to professional-services positions at the firm from 2018-2019. Panel A shows the racial composition of applicants selected for an interview by the firm. Panel B shows the racial composition of applicants selected for an interview by the static 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



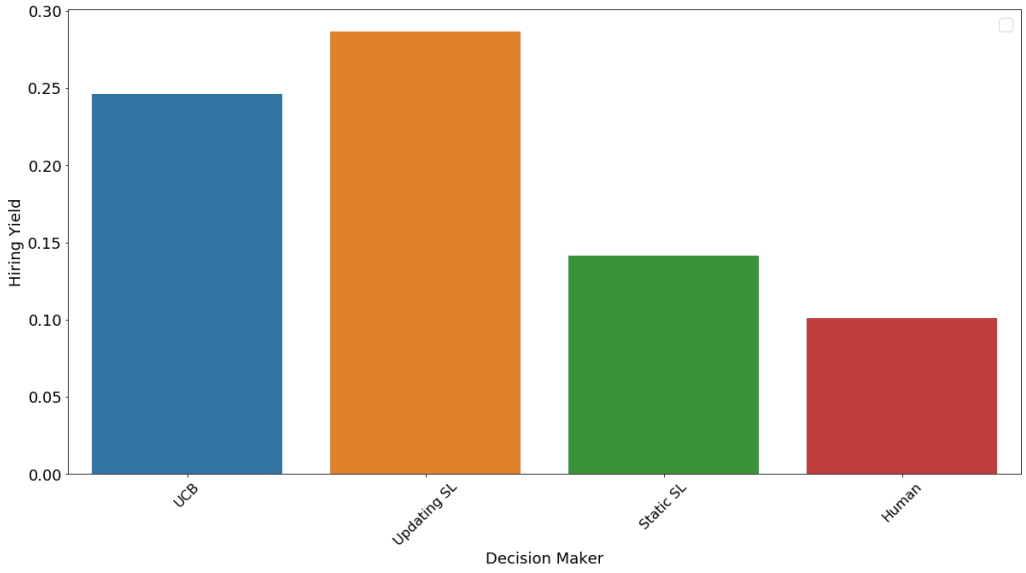
Notes: This figure shows the gender composition of applicants to professional-services positions at the firm from 2018-2019. Panel A shows the gender composition of applicants selected for an interview by the firm. Panel B shows the gender composition of applicants selected for an interview by the static 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



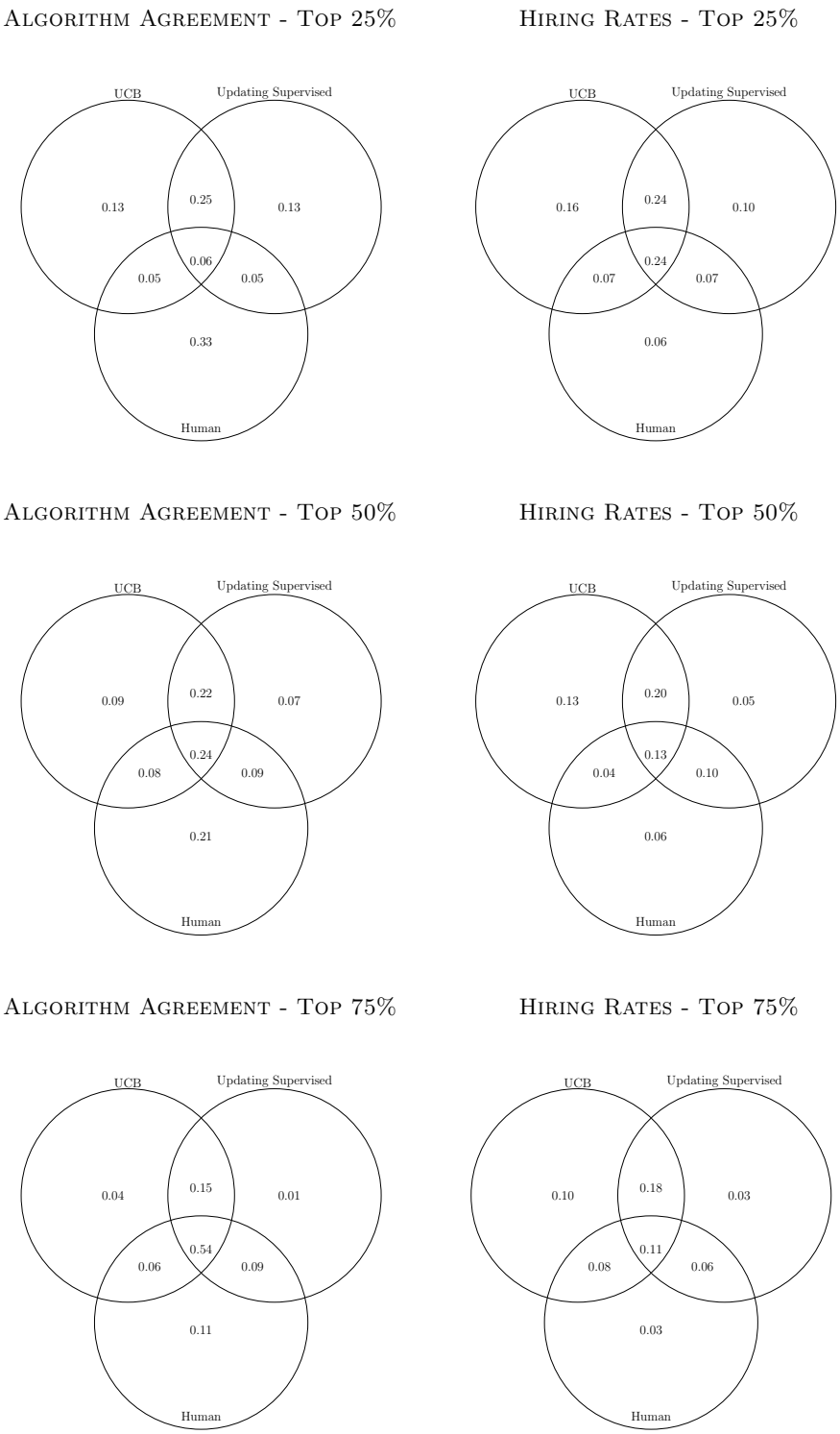
NOTES: Each panel of this figure plots algorithm selection scores on the x-axis and the likelihood of an applicant being hired if interviewed on the y-axis. Panel A shows the selection scores from an algorithm that predicts the firm's actual selection of which applicants to interview. Panel B shows the 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



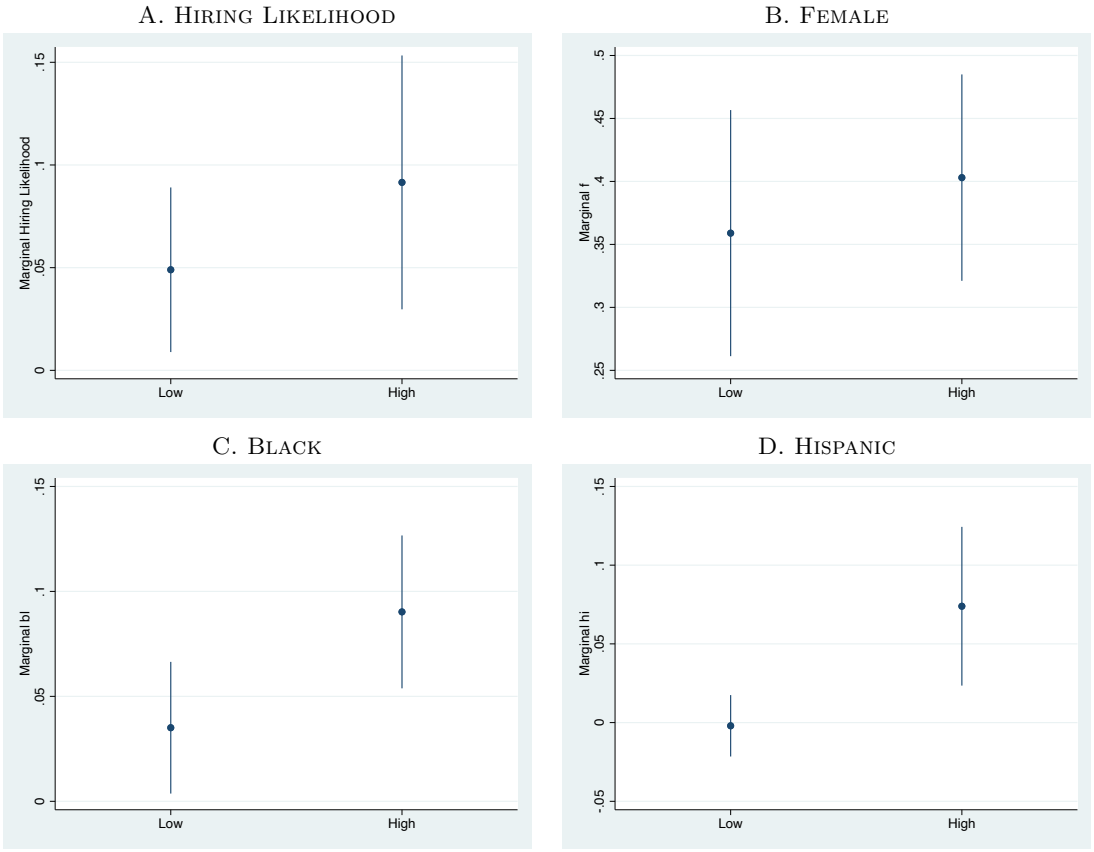
NOTES: This figure shows the unconditional average hiring yield of each unblinded algorithmic selection strategy and the human recruiters. The first bar shows average hiring yield for applicants selected by the contextual bandit UCB algorithm, the second by the updating supervised learning algorithm 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: Algorithm Agreement



Notes: This figure shows the agreement rates and hiring rates for the human, updating supervised learning, and upper-confidence bound (UCB) algorithms. The diagrams in the left column show the share of agreement across the algorithms and those the right column show the hiring rates when the algorithms agree or disagree. The top panel shows these results when selecting from the top 25% of candidates of any algorithm; the middle panel from the top 50% of candidates of any algorithm; and the lower panel from the top 75% of candidates. See text for the exact specification of each algorithm. All data come from the firm's application and hiring records.

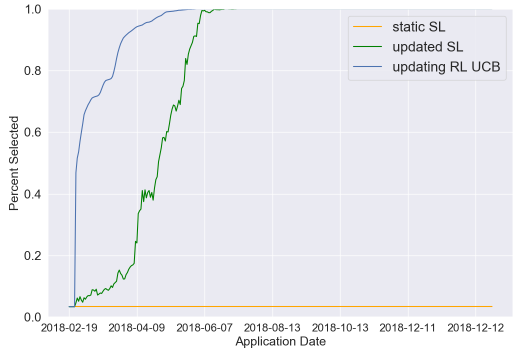
FIGURE 6: CHARACTERISTICS OF MARGINAL INTERVIEWEES, BY UCB SCORE



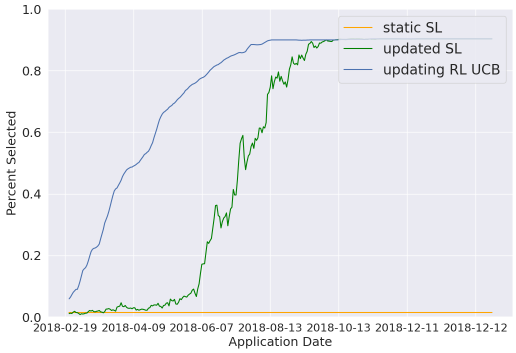
NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics are estimated separately for applicants in the top tercile of the supervised-learning algorithm's score and for applicants in the lowest tercile 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 7: DYNAMIC UPDATING, INCREASED QUALITY

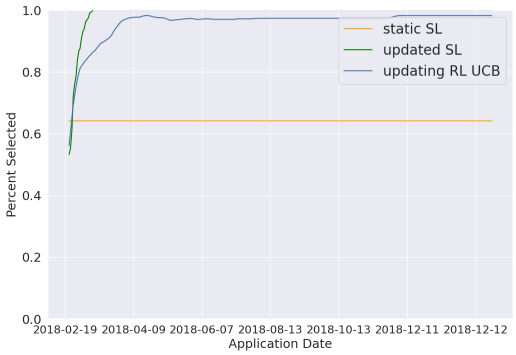
A. BLACK



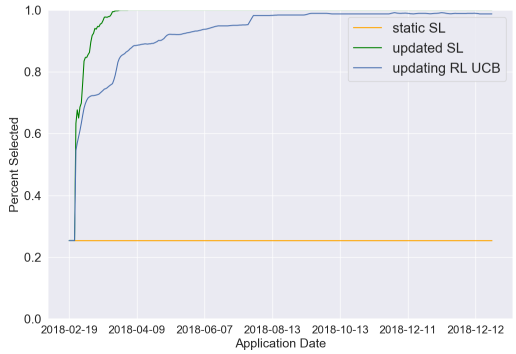
B. HISPANIC



C. ASIAN

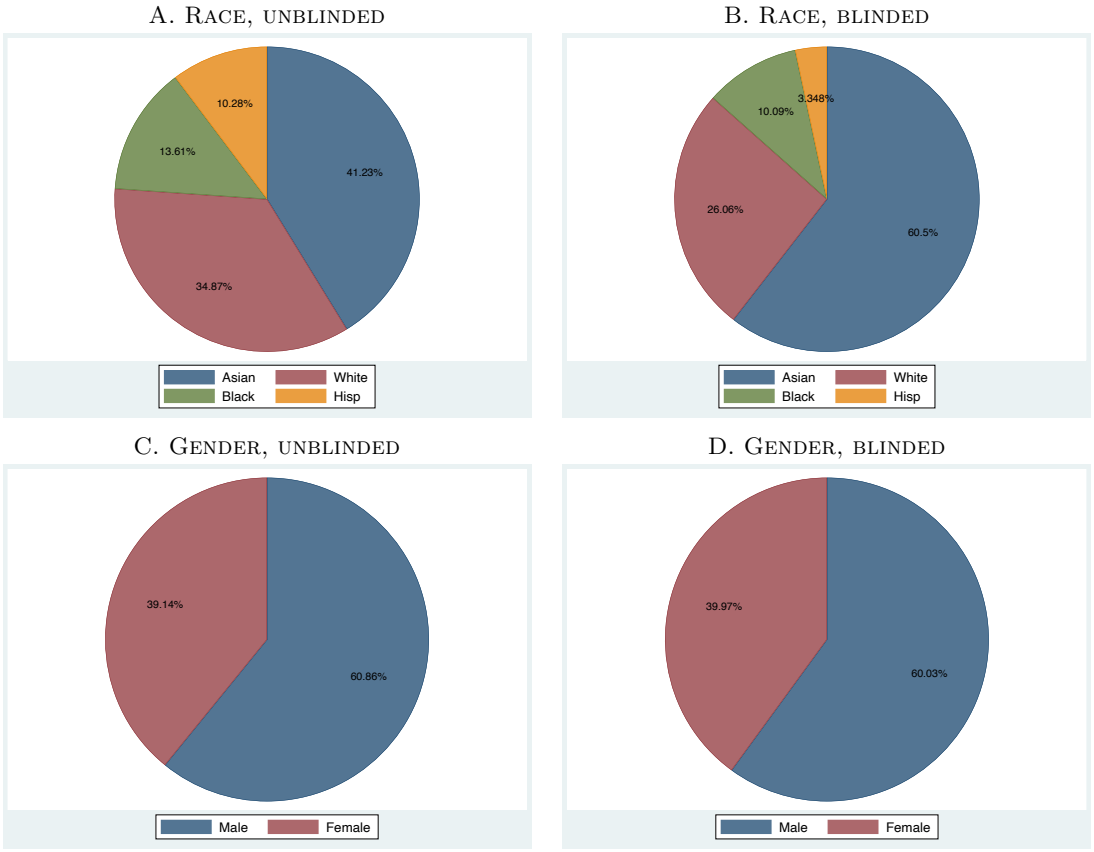


D. WHITE



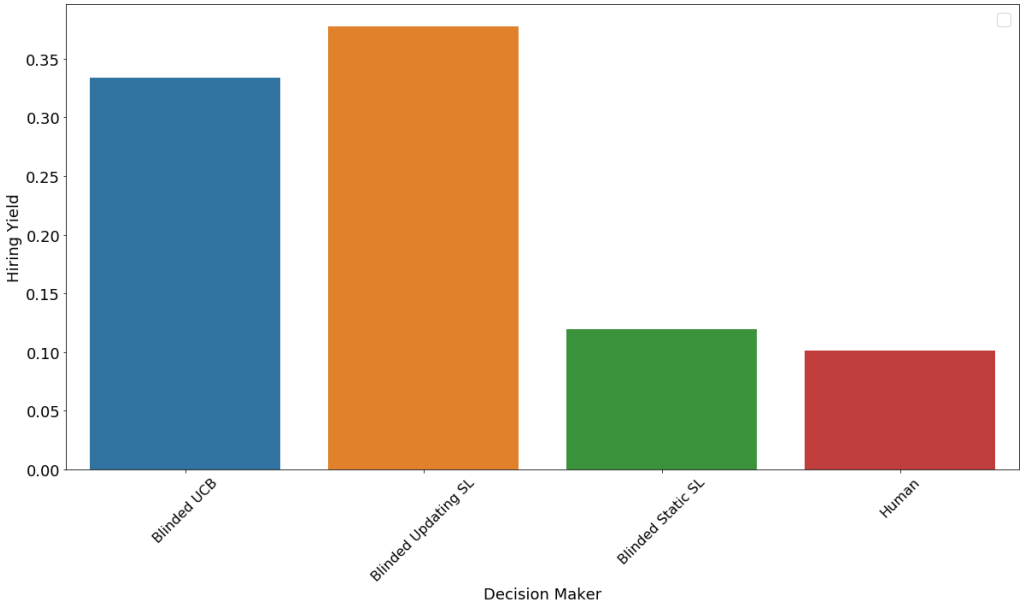
NOTES: This figure shows the racial composition of applicants recommended for interviews under by the UCB algorithm, updating supervised learning and a static version of supervised learning, when quality of a given race dramatically improves. The graph plots each algorithm's belief about quality for 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 8: 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 9: 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 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
Observations	54,243	43,997	98,240

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: APPLICANT FUNNEL BY DEMOGRAPHICS

Variable	Observations	Share of Previous Rd.
Black Applicants		
Applications	7,741	
Interviewed	218	0.03
Hired	8	0.04
Hispanic Applicants		
Applications	3,712	
Interviewed	210	0.06
Hired	16	0.08
Asian Applicants		
Applications	51,678	
Interviewed	2,768	0.05
Hired	308	0.11
White Applicants		
Applications	25,846	
Interviewed	1,561	0.06
Hired	190	0.12
Female Applicants		
Applications	31,341	
Interviewed	1,631	0.05
Hired	191	0.12
Male Applicants		
Applications	63,302	
Interviewed	3,418	0.05
Hired	359	0.11

NOTES: This table shows the number and share of applicants who make it through each round of the interview process—the application round, the interview round and hiring—by demographic group. The initial sample is all applicants who applied to a business analyst, data analyst, data scientist, or financial analyst position from the year 2016 through March of 2019. The first three rows application, interview and hiring rates for Black applicants, and the remaining rows show these rates for Hispanic, Asian, White, female and male applicants. The "Share of Previous Rd." column shows the share of applicants who make it to a particular stage of the process relative to the number in the previous round. All data come from the firm's application and hiring records.

Table 3: 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 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.

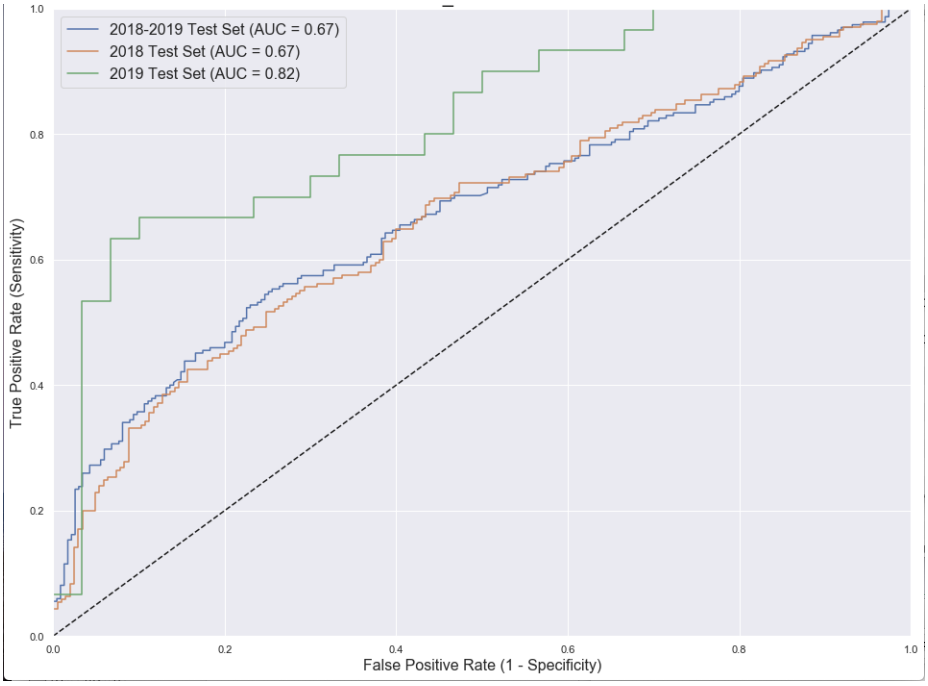
TABLE 5: INSTRUMENT VALIDITY

	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.

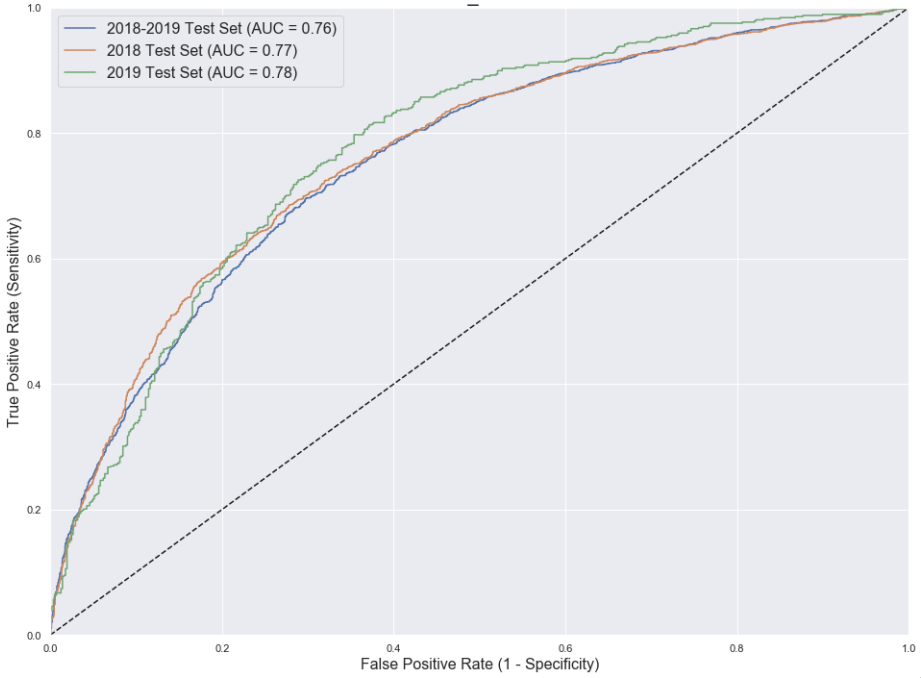
8 Appendix

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



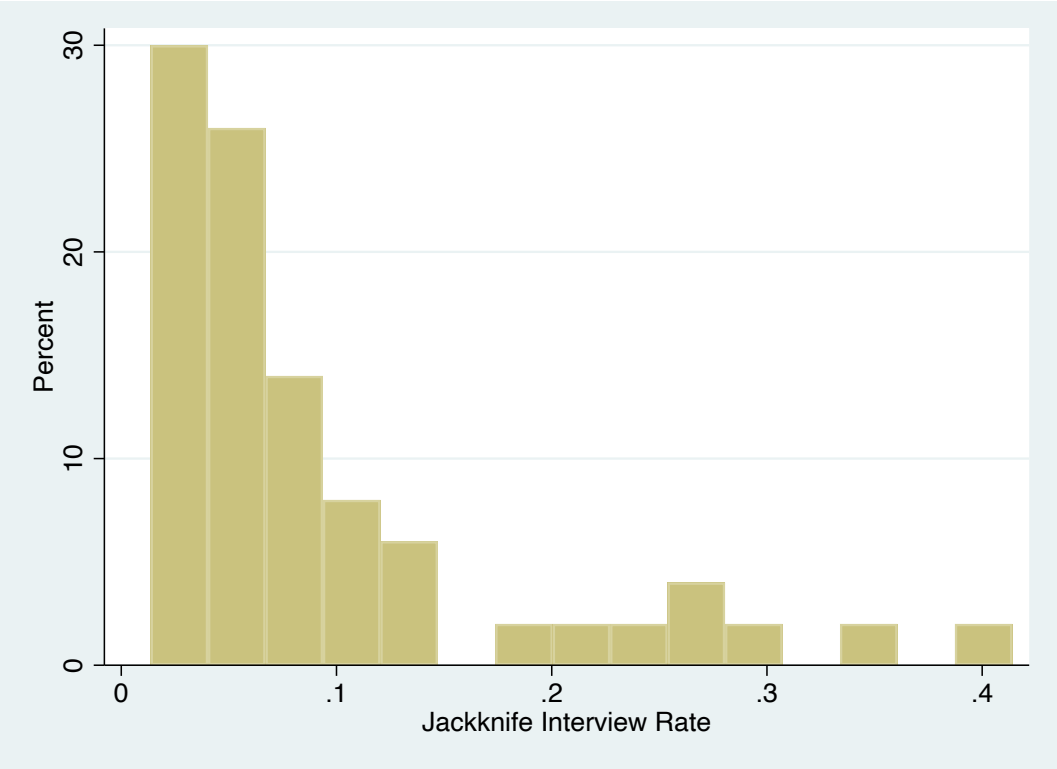
NOTES: This figure shows Receiver-Operating Characteristic (ROC) Curve for the baseline static supervised learning model, which predicts hiring potential. The ROC Curve plots the false positive rate on the x-axis and the true positive rate on the y axis. For each model, we plot this curve for different 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: MODEL PERFORMANCE: PREDICTING INTERVIEW SELECTION



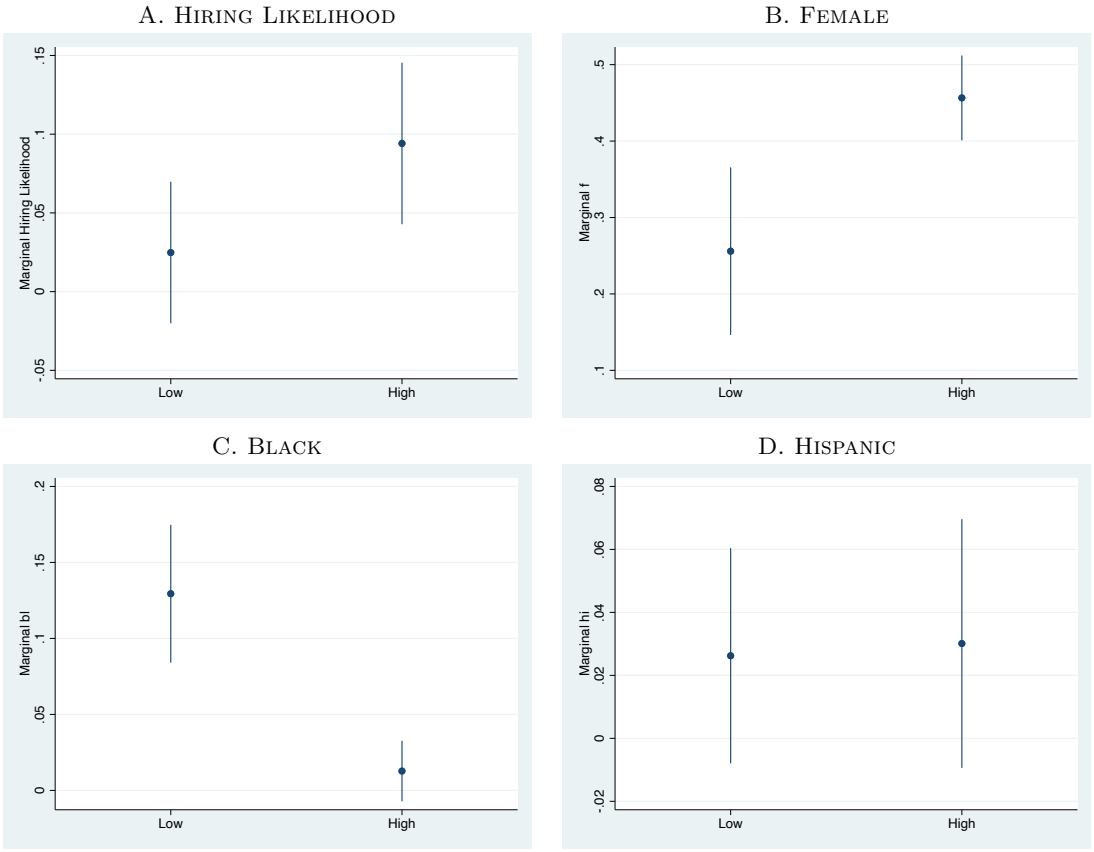
NOTES: This figure shows Receiver-Operating Characteristic (ROC) Curve for the human-decision making model, which predicts who is interviewed. The ROC Curve plots the false positive rate on the x-axis and the true positive rate on the y axis. For each model, we plot this curve for different test data: 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.3: 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.4: CHARACTERISTICS OF MARGINAL INTERVIEWEES, BY UPDATING SUPERVISED SCORE



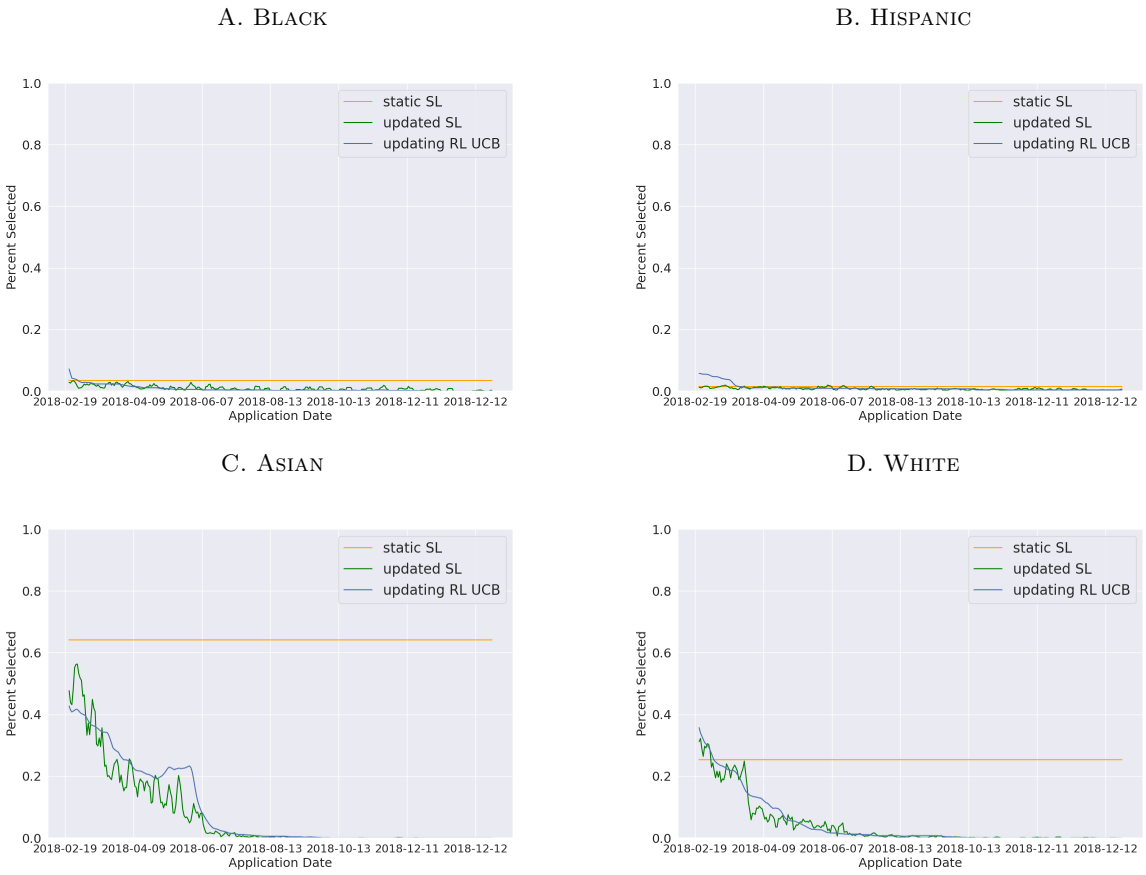
NOTES: Each panel in this figure shows the results of estimating the characteristics of applicants interviewed on the margin. In each panel, these characteristics is estimated separately for applicants in the top half of the supervised-learning algorithm's score and for applicants in the lower half of 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.

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.

FIGURE A.5: DYNAMIC UPDATING, DECREASED QUALITY



NOTES: This figure shows the racial composition of applicants recommended for interviews under by the UCB algorithm, 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.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.