Review of Economic Studies (2026) **93**, 1200–1240 https://doi.org/10.1093/restud/rdaf040
© The Author(s) 2025. Published by Oxford University Press on behalf of The Review of Economic Studies Limited.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Advance access publication 12 June 2025

Hiring as Exploration

Danielle Li
Massachusetts Institute of Technology and National Bureau of Economic Research, USA
Lindsey Raymond
Massachusetts Institute of Technology, USA
and
Peter Bergman
The University of Texas at Austin and National Bureau of Economic Research, USA

First version received January 2022*; Editorial decision August* 2024*; Accepted May* 2025 *(Eds.)*

This article views hiring as a contextual bandit problem: to find 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. Instead, we build a resume screening algorithm that values exploration by evaluating candidates according to their statistical upside potential. Using data from professional services recruiting within a Fortune 500 firm, we show that this approach improves the quality (as measured by eventual hiring rates) of candidates selected for an interview, while also increasing demographic diversity, relative to the firm's existing practices. The same is not true for traditional supervised learning-based algorithms, which improve hiring rates but select far fewer Black and Hispanic applicants. Together, our results highlight the importance of incorporating exploration in developing decision-making algorithms that are potentially both more efficient and equitable.

Key words: Hiring, Machine learning, Algorithmic fairness, Contextual bandits, Job search

JEL codes: D80, J20, M15, M51, O33

1. INTRODUCTION

Increasing access to job opportunity for minorities and women is crucial for reducing well-documented race, ethnicity, and gender gaps in the economy. While a proliferation of initiatives related to diversity, equity, and inclusion speak to firms' interest in these issues, a persistent doubt remains: how can firms increase diversity without sacrificing quality?

Concerns about equity–efficiency tradeoffs in hiring are predicated on the assumption that firms are able to perfectly predict the quality of the applicants they encounter. In such a case, any deviation from the predicted ranking—whether to select more minority or majority group members—would result in a decline in worker quality. In practice, however, an extensive literature has documented that firms, and the recruiters that they employ, are often inaccurate or biased in

The editor in charge of this paper was Katrine V. Løken.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

their predictions (Benson *et al.*, 2021; Kline *et al.*, 2022). Given that firms appear to be far from perfect in their ability to forecast quality, there may be significant scope for improved evaluation tools to expand opportunities for a broader range of candidates while maintaining or even improving worker quality.

In this article, we examine the role of algorithms in the hiring process. Resume screening algorithms have become increasingly prevalent in recent years and have been used to assess job candidates across various industries and occupations.[1,2] Amazon, for instance, was widely criticized for using a resume screening algorithm that penalized the presence of the term "women" (*e.g.* "captain of women's crew team") on resumes.[3]

Our article uses data from a large Fortune 500 firm to study the decision to grant first-round interviews for high-skill positions in consulting, financial analysis, and data science—sectors which offer lucrative jobs with opportunities for career advancement, but which have also been criticized for their lack of diversity. We study the impact of two types of algorithmic approaches: a supervised learning model that selects the best candidates as predicted based on its current training data and a contextual bandit model that seeks to expand its training data in order to learn about the best candidates over time. Our findings demonstrate that while both algorithmic approaches improve the quality of applicants selected by the firm, they differ in their ability to select diverse candidates. The traditional supervised learning approach leads to a significant reduction in the number of Black and Hispanic workers receiving interviews compared to human hiring practices. In contrast, the contextual bandit approach increases the representation of underrepresented minorities. To our knowledge, this study provides the first empirical evidence that algorithmic design can lead to Pareto improvements in both representation and worker quality.

Modern hiring algorithms typically model the relationship between applicant covariates and outcomes in a given training dataset and then apply this model to predict outcomes for subsequent applicants. By systematically analysing historical examples, this supervised learning approach can unearth predictive relationships that may be overlooked by human recruiters. Yet because this approach implicitly assumes that past examples extend to future applicants, firms that rely on this approach may favour groups with proven track records, to the detriment of non-traditional applicants. Indeed, because algorithms are most frequently used at the very top of the hiring funnel, this may prevent such applicants from accessing even initial interviews.

We develop and evaluate an alternative algorithm 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 only in exploitation because they are designed to solve static prediction problems. By contrast,

1. Accurate adoption rates are elusive, but a 2020 survey of human resource executives found that 39% reported using predictive analytics in their hiring processes, a significant increase from just 10% in 2016 (Mercer, 2020). Furthermore, a survey of technology companies indicates that 60% plan to invest in AI-powered recruiting software in 2018, and over 75% of recruiters believe that artificial intelligence will revolutionize hiring practices (Bogen and Rieke, 2018). We discuss evidence on algorithmic adoption further in Section 2.2. Throughout this article, we use the terms "hiring algorithm," "hiring ML," and "resume screening algorithm" interchangeably to refer to algorithms that assist in making initial interview recommendations. It remains rare for algorithms to make final hiring decisions (Raghavan *et al.*, 2020).

2. For example, see Yala *et al.* (2019), McKinney *et al.* (2020), Mullainathan and Obermeyer (2021), Schrittwieser *et al.* (2020), Russakovsky *et al.* (2015), and Ajunwa (2019).

3. See https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

bandit models are designed to solve dynamic prediction problems that involve learning from sequential actions: in the case of hiring, these algorithms value exploration because learning improves future choices.

Our supervised learning model (hereafter, "SL") is based on a logit LASSO that is trained to predict an applicant's underlying hiring potential, *e.g.* whether they would receive and accept and offer if interviewed.[4] Our model is dynamic in the sense that we update its training data throughout our analysis period with the offer and hiring outcomes of the applicants it chooses to interview.[5] This updating allows the SL model to learn about the quality of the applicants it selects, but the model remains myopic in the sense that it does not incorporate the value of this learning into its selection decisions ex ante.

Our contextual bandit approach implements an Upper Confidence Bound (hereafter, "UCB") algorithm. In contrast to the SL model, which evaluates candidates based on their point estimates of hiring potential, a UCB contextual bandit selects applicants based on the most optimistic assessment of their hiring potential. That is, among applicants with the same predicted hiring potential, the UCB model would prefer the one for whom its estimate is most uncertain. Once candidates are selected, we incorporate their realized offer and hiring outcomes into the training data and update the algorithm for the next period.

In terms of demographics, we show that a traditional SL model would interview substantially fewer Black and Hispanic applicants relative to the firm's current hiring practices: a reduction from 9.4% under the status quo to 4.2% with the SL model. In contrast, implementing a UCB model would more than double the share of interviewed applicants who are Black or Hispanic, from 9.4% to 24.3%. Both models increase the share of women relative to human recruiters. These results suggest that exploration in the bandit sense—selecting candidates with covariates for which there is more uncertainty—can lead firms to give more opportunities to workers from groups that are under-represented in their training data, even if diversity goals are not a distinct part of the algorithm's mandate.

A key question, however, is what would happen to worker quality. Bandit algorithms may increase demographic representation by exploring, but this exploration could come at the expense of worker quality. To assess this, we must overcome a missing data problem: we only observe hiring outcomes for candidates who were interviewed in reality.[6] We take three complementary approaches, each based on different assumptions, all of which show that algorithms outperform human recruiters in terms of identifying applicants who are more likely to be hired by the firm.

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. We find that, for both ML models, applicants with high scores are much more likely to be hired than those with low scores. In contrast, there is almost no relationship between an applicant's propensity to be selected by a human, and their eventual hiring outcome; if anything, this relationship is negative.

Our second approach uses inverse propensity score weighting to recover an estimate of mean hiring likelihood among applicants selected from our full applicant sample. This approach infers hiring outcomes for applicants who are not interviewed using observed outcomes among interviewed applicants with similar covariates. The propensity score weighting approach is consistent

4. In Section 7.1, we consider models which maximize offer likelihood.

5. In practice, we can only update the model with hiring outcomes for applicants it selects who are also actually interviewed in practice. See Section 5.2 for a more detailed discussion of how this impacts our analysis.

6. This is also referred to as a "selective labels" problem. See, for instance, Lakkaraju *et al.* (2017), Kleinberg *et al.* (2018), and Arnold *et al.* (2021) .

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

as long as there is no selection on unobservables. In our setting, this assumption is realistic because human recruiters have access to largely the same resume information we do prior to making an interview decision and do not interact with applicants. We continue to find that ML models improve hiring yield (*i.e.* average hire rates among interviewed applicants): 32 and 27% of applicants selected by the SL and UCB models are eventually hired, respectively, compared with only 10% among those selected by human recruiters.

Our third approach uses an instrumental variables strategy to address concerns about the potential for selection on unobservables. In our setting, applicants are randomly assigned to initial resume screeners, who vary in their leniency in granting an interview. We show that applicants selected by stringent screeners (*e.g.* those subject to a higher bar) have no better outcomes than those selected by more lax screeners: this suggests that humans are not positively screening candidates based on their unobservables. We use this same variation to identify the returns to following ML recommendations on the margin by looking at instrument compliers. We find that marginal candidates with high UCB scores have better hiring outcomes and are also more likely to be Black or Hispanic. Such a finding suggests that following UCB recommendations on the margin would increase both the hiring yield and the demographic diversity of selected interviewees. In contrast, following SL recommendations on the margin would generate similar increases in hiring yield but decrease minority representation.

We also provide some evidence relating hiring yield to other measures of applicant quality. We observe job performance ratings and promotion outcomes for a small subset of workers hired in our sample. Among this selected group, we show that our ML models (trained to maximize hiring likelihood) appear more positively correlated with on-the-job performance ratings and future promotion outcomes than a model trained to mimic the choices of human recruiters. This provides suggestive evidence that following ML recommendations designed to maximize hiring yield does not come at the expense of on-the-job performance, relative to following human recommendations.

Finally, we also show that our results are broadly robust to focusing on whether an applicant receives an offer, rather than whether they are hired (*e.g.* receive and accept an offer). We repeat much of our initial analysis using models designed to maximize offer likelihood rather than hiring likelihood. Again, we find similar results: relative to human practices, UCB models select a more diverse set of candidates who are also more likely to receive offers. This same is not true for the SL-based offer model which improves offer likelihood but selects fewer minority applicants.

Together, our main findings show that there need not be an equity-efficiency tradeoff when it comes to expanding diversity in the workplace. Specifically, firms' *current* recruiting practices appear to be far from the Pareto frontier, leaving substantial scope for new ML tools to improve both hiring rates and demographic representation. Incorporating exploration in our setting would lead our firm to interview twice as many under-represented minorities while more than doubling its predicted hiring yield. This logic is consistent with a growing number of studies showing that firms may hold persistently inaccurate beliefs about the quality of minority applicants, and may benefit from nudges (algorithmic or otherwise) that generate additional signals of their quality.[7]

At the same time, our SL model leads to similar increases in hiring yield, but at the cost of drastically reducing the number of Black and Hispanic applicants who are interviewed. This divergence in demographic representation between our SL and UCB results demonstrates the importance of algorithmic design for shaping access to labour market opportunities.

7. For instance, see Miller (2017), Bohren *et al.* (2019, 2025), and Lepage (2020a, 2020b).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

In extensions, we consider several alternative screening policies. We show that blinding our algorithms to race, ethnicity, and gender variables still generates increases in the share of Black, Hispanic, and female applicants who are selected relative to human hiring. The main difference between our blinded and unblinded UCB models is that the share of selected Asian applicants increases while the share of selected White applicants decreases. We also show that our UCB model performs better on quality when compared to a supervised learning model in which we implement group-specific quotas. Our model further has the advantage of achieving increases in diversity without requiring firms to explicitly specify interview slots by sensitive categories such as race, ethnicity, and gender, a practice that often faces legal challenges or requires information that firms may not always be able to collect.

2. BIAS IN HIRING PRACTICES

2.1. *Human hiring*

To be successful, firms must identify and hire the right workers. In most firms, this task falls to human workers, who screen initial applications for further consideration, conduct interviews, and make final hiring decisions. Because resumes, interviews, and other assessment tools are limited in their ability to reveal an applicant's potential, firms ultimately have to rely on the personal judgment of their recruiters.

A longstanding social sciences literature shows that human evaluators perform their jobs imperfectly. Human decision-makers may be simultaneously cognitively limited in their ability to process data (Treisman and Gelade, 1980; Benjamin, 2019; Gabaix, 2019), overconfident in their assessments (Fischhoff *et al.*, 1977; Svenson, 1981; Kausel *et al.*, 2016), and update both too little and too much in response to feedback (Möbius *et al.*, 2022). In addition to these behavioural biases, evaluators may have social preferences for particular applicants. For example, in an ethnographic study, Rivera (2012) documents how recruiters at elite professional services firms favour applicants who share the same hobbies ("She plays squash. Anyone who plays squash I love"). Reviewers' biases may further be exacerbated by time pressure, which may lead them to lean more heavily on unreliable heuristics. In a study of hiring for software engineering roles, Lerner and Bergman (2024) found that recruiters spent a median of 31 seconds per resume.

Such behaviours may contribute to already well-documented race, ethnicity, and gender gaps in the labour market (Bertrand and Duflo, 2017; Blau and Kahn, 2017; Pager and Shepherd, 2008). For example, role congruity theory suggests that managers may find it more difficult to imagine women succeeding in high-level roles because of a mismatch between the qualities stereotypically associated with effective leaders and with women (Eagly and Karau, 2002). Benson *et al.* (2021) find, indeed, that managers incorrectly assess women as having lower "potential" within the firm. In a large scale correspondence study, Kline *et al.* (2022) find evidence that recruiters discriminate against Black applicants across a range of firms and industries. In a study in Eastern Europe, Bartos *et al.* (2016) show that discrimination in outcomes may be presaged by discrimination in attention: hiring managers pay less attention when evaluating resumes with Roma-sounding names. Recent work by Bohren *et al.* (2019) suggests that some of these differences may be due to managers having incorrect biased beliefs.

A variety of studies consider ways to mitigate these biases, with mixed results. For example, a common suggestion is that women and minorities may benefit by being evaluated by other women and minorities. This solution, however, is often not supported in the data: Bagues and Esteve-Volart (2010), for example, finds that the presence of women on recruiting committees

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

can, in fact, hurt female applicants. Another suggestion is to require decision-makers to undertake anti-bias trainings. Yet, while lab studies have shown that de-biasing exercises (perspective taking, counter-stereotyping) can reduce biases, there is less evidence about their efficacy in real organizations (Paluck and Green, 2009). Rather, evidence on durable changes in attitudes seems to come from prolonged cross-group exposure (*e.g.* shared living, schooling, or service) that is difficult for firms to implement as a policy (Rao, 2019; Bagues and Roth, 2023). Finally, affirmative-action approaches to redressing discrimination face increasing legal scrutiny (United States Court of Appeals for the First Circuit, 2020).

Rather than mitigating the biases that evaluators may hold, another strand of research considers the impact of limiting their ability to exercise unconstrained judgment. Hastie and Dawes (2001) surveys studies examining the predictive accuracy of human evaluators across a range of settings and concludes that "expert judgments are rarely impressively accurate and virtually never better than a mechanical judgment rule." Less is known about how constraining human judgment may impact diversity outcomes. Proponents of holistic review have argued that minority groups can benefit from an evaluator's ability to account for assessments of adversity. At the same time, allowing for discretion may introduce opportunities for decisions to be clouded by an evaluator's implicit biases or personal preference (Prendergast and Topel, 1993; Bertrand *et al.*, 2005).

Rules-based assessments, in essence, suggest that decisions can be improved if humans behaved more like machines. Our article takes this idea to its conclusion and examines how the growing adoption of algorithms may impact both the quality and equity of firms' hiring practices.

2.2. *Algorithmic approaches*

Firms are increasingly turning toward data-driven tools to improve their hiring practices. The most ubiquitous hiring technology is an Applicant Tracking System (ATS). While the baseline versions of these tools simply keep track of applicants, they often offer additional functionality, such as allowing recruiters to filter applicants based on whether their resumes meet the requirements listed on-the-job application. A 2021 survey of employers in the U.S., U.K., and Germany found that 94% of respondents used some automated tools to filter or rank candidates (Fuller *et al.*, 2021).

In recent years, a growing number of firms have begun offering more powerful ML-based tools that predict a candidate's suitability for a role based on historical data on hiring and performance in that role.[8] These types of algorithms are most commonly used at the "top of the funnel," to prioritize applicants for initial interviews. Recruiters at this stage often face the task of sifting through thousands of applications for just a handful of open positions.[9] Organizations surveyed about their use of such algorithms frequently express the hope that these tools will enable them to efficiently identify qualified candidates and fill vacancies more quickly (Bogen and Rieke, 2018). A 2020 industry survey found that 55% of U.S. firms use predictive analytics at some point in their human resource decision-making process, while 41% use algorithms to make predictions about worker fit (Mercer, 2020).[10] Algorithms are also commonly used in

8. Raghavan *et al.* (2020) provide an overview of such vendors.

9. Fuller *et al.* (2021) report that the average job opening for a corporate position posted in 2020 receives 200 applications, up from 100 in 2010.

10. Many well-known companies, such as Intel, Johnson and Johnson, Dominos, JP Morgan, United Parcel Service, Mastercard, LinkedIn, Unilever, and Accenture, have openly acknowledged using algorithmic hiring tools for a variety of job roles (Todd, 2019). Indeed, a recent 2023 study found that 60% of Fortune 100 firms work with a single hiring analytics company, HireVue (Nawrat, 2023).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

the public sector; Nawrat (2023) reports that eight of the ten largest U.S. federal agencies use algorithmic screening for some roles.

The use of algorithms, moreover, is not restricted to organizations who haven specifically chosen to buy a customized algorithmic solution. Employers, including small employers, often post their job openings on third party job search platforms such as LinkedIn or ZipRecruiter, all of which use ML-based tools to decide which applicants to recommend for an open position. As a result, algorithms play a role in screening applicants even for organizations that do not actively choose to employ algorithmic tools. Further, as AI tools become increasingly integrated with widely used ATS systems, the use of algorithms is likely to grow.[11]

Algorithms may not suffer from some of the key limitations that human recruiters face. Whereas individual recruiters are likely to base their judgments on their own narrow experience, algorithms are trained on much larger datasets of applicants. For any given applicant, algorithms are able to form predictions using many variables, while human attention is more limited.[12] Algorithms also assess applicants instantaneously, consistently, and without fatigue, in contrast with research showing that human evaluators are inconsistent and suffer from cognitive fatigue (Gabaix, 2019; Hirshleifer *et al.*, 2019).

Consistent with these advantages, existing evidence suggests that algorithms may improve the quality of hiring decisions (Hoffman *et al.*, 2017; Cowgill, 2020). Crucially, however, a growing literature has raised questions about how the increasing adoption of algorithms may impact equity and access to job opportunities.[13] A key concern is that algorithms may be trained on data that reflects historical inequities and, in turn, replicate these biases (Lambrecht and Tucker, 2019; Obermeyer *et al.*, 2019) Anecdotal accounts of algorithmic bias in hiring have also been widely reported in the popular press: an audit of one resume screening model, for instance, found that the two variables it most strongly favoured were being named "Jared" and playing high school lacrosse.[14]

Much of this criticism has implicitly focused on algorithms based on supervised learning. Supervised learning relies on the existence of labelled datasets to train models to predict a given outcome. In the context of hiring, these datasets tend to be based on applicants that a firm has seen and hired in the past. A supervised learning model may then favour applicants who play lacrosse because socioeconomic status or cultural fit has historically been predictive of success in the hiring process. To the best of our knowledge, most commercially available hiring algorithms are based on this type of approach.[15]

In this article, we highlight an alternative class of algorithms that have thus far not been applied or studied in the context of hiring: contextual bandit algorithms. Whereas supervised

11. For example, the ATS vendor Workday recently acquired the hiring prediction firm HiredScore in order to build more AI tools into their platform: acquisition announcement here.

12. Mullainathan and Obermeyer (2021), for instance, provide evidence that the optimal number of variables that predict patient outcomes is greater than the number that doctors can attend to.

13. For surveys of algorithmic fairness, see Bakalar *et al.* (2021), Barocas and Selbst (2016), Corbett-Davies *et al.* (2023), and Cowgill and Tucker (2019). For a discussion of broader notions of algorithmic fairness, see Kasy and Abebe (2020) and Kleinberg *et al.* (2017).

14. See https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased.

15. In general, most firms do not provide information on the specifics of their proprietary algorithms. However, several industry sources have indicated that this is true of their own algorithms. Further, most discussions of hiring ML implicitly assume that this is the case. For example, in a survey of firm approaches, Raghavan *et al.* (2020) discuss many different ways in which firms may implement supervised learning approaches (*e.g.* what outcomes to train on or what historical data to use), but there is no discussion of any alternative algorithmic approaches that firms may take. We were also unable to find any reports of firms using bandit approaches in our review of various industry surveys, *e.g.* Mercer (2020) and Bogen and Rieke (2018).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

learning models focus solely on selecting applicants with high predicted quality, bandit algorithms also seek out candidates in order to learn about their quality. A small empirical literature has shown that firms can benefit from non-algorithmic policies that push them to adopt more exploratory practices: Miller (2017) shows that temporary affirmative action policies can generate persistent gains in minority representation, while Whatley (1990) documents a similar finding by examining the racial integration of firms following World War I. We take this idea and ask whether algorithms can implement exploration in a more efficient way.

These ideas have broader implications for various settings in which decision-makers need to assess the quality of applicants. While we focus on hiring, these same selection problems arise in promotion, credit scoring, loan approval, university admissions, investing, and allocating research funding. In all these cases, decision-makers strive to gain insights into an applicant's quality, often in situations where historical data records may be incomplete or biased. In such scenarios, supervised learning algorithms designed to solve static prediction problems may not be the most suitable tools.[16] Instead, our article proposes that these problems could be reframed as dynamic learning problems, for which exploration-based algorithms can be particularly useful. By embracing a more adaptive and exploratory approach, decision-makers can potentially overcome the limitations of historical data and make more informed, forward-looking assessments of applicant quality.

3. OUR SETTING

We focus on recruiting for high-skilled, professional services positions, a sector that has seen substantial wage and employment growth in the past two decades (BLS, 2019). At the same time, this sector has attracted criticism for its perceived lack of diversity: female, Black, and Hispanic applicants are substantially under-represented relative to their overall shares of the workforce (Pew Research Center, 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 the recruiting, retention, and promotion of those from low-income and historically under-represented groups.[17] However, despite these efforts, organizations routinely struggle to expand the demographic diversity of their workforce—and to retain and promote those workers—particularly in technical positions (Athey *et al.*, 2000; Castilla, 2008; Jackson, 2023).

The hiring process at our firm works as follows. Applicants submit their resumes to posted jobs using the firm's online portal. On average, our firm receives approximately 200 applications per opening, which is in line with a 2021 industry study showing that corporate job postings receive an average of 250 applications (Fuller *et al.*, 2021). Applicants are first screened by human recruiters, who are full-time HR professionals employed by the firm. Recruiters do not meet or interact with applicants and their job is to decide which applicants to interview based only on the information submitted via the online portal.

16. In many settings, such as college admissions, there has been less algorithmic adoption compared to hiring. However, we note that the absence of explicit algorithms does not necessarily mean the absence of (potentially biased) algorithm-like thinking. For example, academics frequently rely on institutional affiliation when assessing the quality of a piece of research.

17. 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.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

Once a candidate passes the initial resume screening, they are invited to participate in an initial interview. This is often a case-style interview designed to assess job skills.[18] In our data, we will use the term "interviewed" to refer to candidates who reach this initial interview stage. Candidates who do well in this initial interview are then invited for a visit in which they complete additional interviews, this time conducted by employees whom the candidate is likely to be working under if hired. These interviews include additional case interviews, as well as assessments of job fit, problem solving, and leadership. Finally, after the interviews, the hiring team meets to discuss each candidate's strengths and weaknesses, grading them on a pre-defined rubric that includes assessments of their technical skills, communication skills, and cultural fit. The group then ultimately votes on whether or not to extend an offer.

We focus on the resume review stage. Due to the need to divert current employees from other productive tasks to conduct interviews, firms are highly selective when choosing which applicants to interview: our firm rejects 95% of applicants during its initial resume review. Because of the volume of applicants, recruiters must eliminate many applications quickly, based on limited information. For instance, Lerner (2024) found that the median time recruiters spent reviewing a resume was only 31 s. Mistakes in screening not only impact firm productivity but may also perpetuate inequalities if recruiters inadvertently rely on heuristics that disadvantage qualified individuals who do not fit traditional models of success (Rivera, 2012; Friedman and Laurison, 2019). In light of these issues, we believe that it is particularly important to understand whether algorithmic tools can be used to improve decisions at the critical initial screening stage.

3.1. *Data*

Our data come from a Fortune 500 company in the United States that hires workers in several job families spanning business and data analytics (Anonymous Firm, 2019). 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 underrepresented groups. We have data on 88,666 job applications from January 2016 to April 2019, as described in Table 1. Most applicants in our data are male (68%), Asian (58%), or White (29%). Black and Hispanic candidates comprise 13% of all applications, but under 5% of hires. Women, meanwhile, make up 33% of applicants and 34% of hires. We describe our sample and variables in more detail in our Data Appendix, Sections A.1 and A.2.

3.1.1. Applicant covariates. We have information on applicants' educational background, work experience, referral status, basic demographics, as well as the type of position to which they applied. Appendix Table A.I provides a list of these raw variables, as well as some summary statistics. We have self-reported race/ethnicity (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 as 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) U.S. or non-U.S. educational institution. We record the geographic location of education experience at an aggregated level (*e.g.* India, China, Europe). We also track the job family each

18. For example, if the applicant was applying for a business consultant role, the interviewer may describe a hypothetical business problem and ask how the applicant would approach it. If the candidate were applying for a data science role, the candidate may be asked to demonstrate knowledge of statistical principles and describe the types of analyses they might run to answer a particular question.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

TABLE 1
Applicant summary statistics

Variable	Mean training	Mean test	Mean overall
Black	0.087	0.087	0.087
Hispanic	0.040	0.043	0.042
Asian	0.573	0.591	0.581
White	0.300	0.279	0.290
Male	0.677	0.658	0.668
Female	0.323	0.342	0.332
Referred	0.140	0.114	0.129
B.A. Degree	0.232	0.242	0.237
Associate Degree	0.005	0.005	0.005
Master's Degree	0.612	0.643	0.626
Ph.D.	0.065	0.074	0.069
Attended a U.S. College	0.747	0.804	0.772
Attended Elite U.S. College	0.128	0.143	0.134
Interviewed	0.054	0.053	0.054
Offered	0.012	0.010	0.011
Hired	0.006	0.005	0.006
Observations	48,719	39,947	88,666

Notes: This table shows applicants' demographic characteristics, education histories, and work experience. The sample in Column 1 consists of all applicants who applied to a position during our training period (2016 and 2017). Column 2 consists of applicants who applied during the analysis period (2018 to Q1 2019). Column 3 presents summary statistics for the full pooled sample.
Source: All data come from the firm's application and hiring records.

candidate applied to, the number of applications submitted, and the time between the first and most recent application.

3.1.2. Quality measures. A key challenge our firm faces is being able to hire qualified workers to meet its labour demands; even after rejecting 95% of candidates in deciding whom to interview, 90% of interviews do not result in a hire. These interviews are costly because they divert high-skill current employees from other productive tasks (Kuhn and Yu, 2021). In our article, we therefore measure an applicant's quality as their likelihood of actually being hired by the firm. By this definition, a high quality applicant is one that meets the firm's own hiring criteria (whatever that may be) and who accepts the firm's offer of employment. In Section 7.1, we consider quality defined only by whether the firm chooses to extend an offer.

Of course, in deciding whom to interview, firms may also care about other objectives: they may look for applicants who have the potential to become superstars—either as individuals, or in their ability to manage and work in teams—or they may avoid applicants who are more likely to become toxic employees (Reagans and Zuckerman, 2001; Housman and Minor, 2015; Deming, 2017; Benson *et al.*, 2019; Schumann *et al.*, 2019). Unfortunately, we observe little information on applicants' post-hire performance. For the small set of workers for which we observe this data, we provide noisy evidence that ML models trained to maximize hiring rates are also positively related to performance ratings and promotion rates (see Section 7.2).

4. EMPIRICAL STRATEGY

The goal of our article is to understand how implementing an exploration-based resume screening algorithm would impact firms' interview outcomes, relative to it's existing practices, and relative to traditional supervised learning approaches. An ideal comparison would involve

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

randomizing screening technologies (human, supervised ML, or exploration ML) through an experiment.

Our analysis, however, relies on archival data. While we observe demographics for all applicants regardless of whether they are interviewed, we observe quality measures—hiring and offer outcomes—when an applicant is interviewed. This means that we face a "selective labels" problem: if an algorithm selects a candidate who is not interviewed in practice, we will not observe that candidate's interview outcome.

In this section, we describe our framework for addressing this inference challenge. For this discussion, it is sufficient to consider a generic ML-based interview policy. In Section 5, we will describe the details of the specific algorithms we implement.

4.1. *Baseline framework*

We consider a firm that makes interview decisions over time. In each period t, the firm sees a set of job applicants indexed by i, and must choose which candidates to interview $I_{it} \in \{0, 1\}$. The firm would only like to interview candidates that meet its hiring criterion, so a measure of an applicant's quality is her "hiring likelihood": $Y_{it} \in \{0, 1\}$. Y_{it} should be thought of as a potential outcome: would applicant i applying at time t be hired by the firm *if* they were granted an interview? Empirically, Y_{it} is an indicator for whether an applicant receives an offer from the firm or is actually hired (receives and accepts an offer). Regardless of outcomes, the firm pays a cost, c_t, per interview, which can vary exogenously with time to reflect the number of interview slots or other constraints in a given period.

The firm's payoff for interviewing worker i is given by:

$$\text{Payoff}_{it} = \begin{cases} Y_{it} - c_t & \text{if } I_{it} = 1 \\ 0 & \text{if } I_{it} = 0 \end{cases}$$

For each applicant i in period t, the firm also observes a vector of demographic, education, and work history information, denoted by X'_{it}. These variables provide "context" that can inform the expected returns to interviewing a candidate. We write $E[Y_{it}|X'_{it}] = \mu(X'_{it}\theta^*_t)$, where $\mu : \mathbb{R} \to \mathbb{R}$ is a link function and θ^*_t is an unobserved vector describing the true predictive relationship between covariates X'_{it} and hiring potential Y_{it}.[19] We allow X'_{it} to include components that are observed and unobserved by the econometrician. After each period t, the firm observes the payoffs associated with its chosen actions.

Given this information, we can think of a firm's interview decision for applicant i at time t as given by:

$$I_{it} = \mathbb{I}(s_t(X'_{it}) > c_t). \tag{4.1}$$

Here, $s_t(X'_{it})$ can be thought of as a score measuring the value the firm places on a candidate with covariates X'_{it} at time t. This score is indexed by t to reflect the fact that the value of a given applicant can change over time if the firm's beliefs about their quality change or if the firm's priorities do. The firm's goal is to identify a scoring function $s_t(X'_{it})$ that leads it to identify and interview applicants with $Y_{it} = 1$ as often as possible.

19. In practice, when estimating contextual bandit models, most algorithms make functional form assumptions about the underlying true relationships. Of course, in practice firms do not know the true relationship, preventing firms from implementing the ideal decision rule.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

Our model mirrors a standard contextual multi-arm bandit (MAB) problem.[20] Leaving aside the optimal choice of scoring function (which we discuss later in Section 5), we can think of *any* interview policy as being described by some scoring function $s_t(\cdot)$ and its associated interview decision I_{it}. In particular, we write s_t^H and I_{it}^H to refer to the (H)uman interview policy that is used by the firm and s_t^{ML} and I_{it}^{ML} to refer to any counterfactual machine-learning (ML) based interview policy. For notational simplicity, we suppress the subscripts for applicant i at time t for the remainder of the article, unless we are discussing specific regressions or details of algorithm construction.

4.2. *Addressing sample selection*

We are interested in understanding how the quality and demographics of the interviewed candidates change under different interview policies. We see demographics for all applicants and therefore do not face sample selection problems in comparing the demographics of counterfactual interview policies. However, we would also like to examine quality by comparing $E[Y|I^H = 1]$ versus $E[Y|I^{\text{ML}} = 1]$ for traditional and exploration-based ML approaches. $E[Y|I^H = 1]$ is readily computable because we directly observe the hiring potential Y for all workers chosen to be interviewed by human recruiters. $E[Y|I^{\text{ML}} = 1]$, however, is only partially observable because we only see hiring outcomes for the subset of ML-selected applicants who are actually interviewed (*e.g.* selected by human recruiters): $E[Y|I^{\text{ML}} = 1 \cap I^H = 1]$. In our analysis, we address potential biases in assessing the counterfactuals associated with sample selection in three complementary ways. In Section 6, we show that we obtain similar results under each approach.

4.2.1. Interviewed sample only. Our first approach examines the predictive relationship between algorithm scores s^{ML} and applicant quality among the subset of applicants who are actually interviewed, *e.g.* those for whom we observe realized hiring potential. To compare our ML model's preferences to that of human recruiters, we construct a measure of s^H, the implicit "score" that humans assign to applicants by training a model to predict an applicant's likelihood of being selected for an interview $E[I^H|X]$, as described in Appendix A.3.4. We then compare the predictive power of $\hat{s}^H$ with that of s^{ML} among the interviewed sample. Our findings, discussed in Section 6.2.1, will show that the average quality among candidates preferred by ML-based screening approaches is higher than that of candidates preferred by human recruiters.

4.2.2. Full sample, assuming no selection on unobservables. A concern with our above approach is that human recruiters may add value by screening out particularly poor candidates so that they are never observed in the interview sample to begin with. In this case, there may be no correlation between human preferences and hiring potential among those who are interviewed, even if human preferences are highly predictive of quality in the full sample.

Our next approach addresses this by estimating the average quality of *all* ML-selected applicants, $E[Y|I^{\text{ML}} = 1]$. We infer hiring likelihoods for ML-selected applicants who were not interviewed using observed hiring outcomes from applicants with similar covariates who were interviewed, assuming no selection on unobservables: $E[Y|I^{\text{ML}} = 1, X] = E[Y|I^{\text{ML}} = 1, I = 1, X]$.

20. In a generic contextual MAB, an agent receives information on "context" before deciding which bandit "arm" to pull in order to receive different "rewards." In our case, the context information is an applicant's resume and demographics X_{it}; the arms are the decision of whether to interview or not $I_{it} \in \{0, 1\}$; and the rewards are the associated payoffs $Y_{it} - c_t$ if I_{it} if interview or 0 if not.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

In our setting, this is a plausible assumption because recruiters have very little additional information relative to what we also observe. Screeners make interview decisions on the basis of applicant resumes and never interact with applicants. Because the hiring software used by our data firm further standardizes this information into a fixed set of variables, they generally do not observe cover letters or even resume formatting. Given this, the types of applicant information that are observable to recruiters but not to the econometrician are predominately related to resume information that we do not code into our feature set. For example, we convert education information into indicator variables for college major designations, institutional ranks, and types of degrees. A recruiter, by contrast, will see whether someone attended the University of Wisconsin or the University of Michigan.[21] In addition to worker characteristics, our models also include characteristics of the job search itself to account for factors that influence hiring demand independent of applicant characteristics. For more details on the construction of our key variables, see Appendix A.

Following Hirano *et al.* (2003) and assuming no selection on unobservables, we can write the inverse propensity weighted estimate of ML-selected workers' hiring likelihood as:

$$E[Y|I^{\text{ML}}=1]=\frac{p(I=1)}{p(I^{\text{ML}}=1)}E\left[\frac{p(I^{\text{ML}}=1|X)}{p(I=1|X)}Y|I=1\right]. \tag{4.2}$$

Equation (4.2) says that we can recover the mean quality of ML-selected applicants by reweighting outcomes among the human-selected interview sample, using the ratio of ML and human-interview propensity scores.[22] The ML decision rule is a deterministic function of covariates X, meaning that the term $p(I^{\text{ML}}=1|X)$ is an indicator function equal to one if the ML rule would interview the applicant, and zero if not. The term $p(I^{H}=1|X)$ is just the human selection propensity which we estimate as $\hat{s}^{H}$, described previously. Finally, because we always select the same number of applicants as are actually interviewed in practice, the term $\frac{p(I=1)}{p(I^{\text{ML}}=1)}$ is equal to one by construction.

4.2.3. Marginally promoted sample, IV analysis. We continue to be concerned about the possibility of selection on unobservables, particularly if human recruiters screen out unobservably bad applicants. In that scenario, our previous approach would overstate hiring outcomes for ML-selected applicants who do not receive an interview.

To address this, our next approach compares the quality of human and ML selected candidates, using exogenous variation in interview propensity arising from the random assignment of applicants to resume screeners, following the methodology pioneered by Kling (2006). For each applicant, we form the jackknife mean interview rate of their assigned screener and use the leniency of the screener as an instrument, Z, for whether the applicant is interviewed.

Appendix Figure A.XII plots the distribution of jackknife interview pass rates in our data, restricting to the 54 screeners (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 2 shows that this variation is predictive of whether a given applicant is interviewed (Column 1) but is not related to any of the applicant's covariates (Columns 2–8). In Appendix Figure A.XIII, we plot the relationship between ML scores (both UCB and SL) and interview likelihood separately for applicants assigned to strict or lenient screeners. We show that those assigned to lenient screeners are more likely to be interviewed across all ML scores,

21. Adding additional granularity in terms of our existing variables into our model does not improve its AUC.
22. See Appendix C.2 for the full derivation of equation (4.2).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

TABLE 2
Instrument validity

	Interviewed (1)	Black (2)	Hispanic (3)	Asian (4)	White (5)	Female (6)	Ref. (7)	MA (8)
JK interview rate	0.0898***	0.000158	−0.000433	0.00899	−0.00716	−0.00448	−0.0113	0.00910
	(0.00832)	(0.00470)	(0.00189)	(0.0122)	(0.00972)	(0.00557)	(0.0126)	(0.00961)
Observations	37,662	37,662	37,662	37,662	37,662	37,662	37,662	37,662

Notes: This table shows the results of regressing applicant characteristics on interviewer leniency, defined as the jack-knife mean-interview rate for the recruiter assigned to an applicant, controlling for fixed effects for job family, management level, application year, and location of the job opening. The leave-out mean is defined as the share of interviews the assigned recruiter grants, excluding the focal applicant. This leave-out mean is standardized to be mean zero and standard deviation one. The outcome in Column 1 is an indicator variable for being interviewed. The outcomes in Columns (2)–(8) are indicators for baseline characteristics of the applicant. The sample is restricted to recruiters who screened at least 50 applicants. Standard errors are clustered at the recruiter level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.
Source: All data come from the firm's application and hiring records.

and that this advantage is essentially constant. In Appendix D, we show that strict and lenient screeners appear to have similar preference orderings of applicants, reducing concerns about violations of monotonicity.

Given this instrument, the goal of our analysis in this section is to ask whether firms can improve both hiring yield by following ML recommendations "on the margin," that is, in cases when the human recruiter appears to be on-the-fence about whether or not to interview a candidate. Specifically, consider the following counterfactual interview policy, given our recruiter leniency instrument Z and algorithmic score s^{ML}:

$$\tilde{I} = \begin{cases} I^{Z=1} & \text{if } s^{\mathrm{ML}} \geq \tau, \\ I^{Z=0} & \text{if } s^{\mathrm{ML}} < \tau. \end{cases}$$

$\tilde{I}$ takes the firm's existing interview policy, I, and modifies it at the margin. The new policy $\tilde{I}$ favours applicants with high ML scores by asking the firm to make interview decisions I as if these applicants were randomly assigned to a generous initial screener ($Z = 1$).[23] That is, $I^{Z=1}$ refers to the counterfactual interview outcome that would be obtained, if an applicant were evaluated by a lenient screener. Similarly, $\tilde{I}$ penalizes applicants with low ML scores by making interview decisions for them as though they were assigned to a stringent screener ($Z = 0$).

By construction, the interview policy $\tilde{I}$ differs from the status quo policy I only in its treatment of instrument compliers. Compliers with high ML scores will be selected under $\tilde{I}$ because they are treated as if they are assigned to lenient recruiters. Conversely, compliers with low ML scores are rejected because they are treated as if they are assigned to strict reviewers. As such, $\tilde{I}$ provides a concrete example of an alternative policy that improves hiring yield by following ML recommendations more if compliers with high ML scores have greater hiring potential than compliers with low ML scores, $E[Y|I^{Z=1} > I^{Z=0}, s^{\mathrm{ML}} \geq \tau]$ versus $E[Y|I^{Z=1} > I^{Z=0}, s^{\mathrm{ML}} < \tau]$.

23. For simplicity in exposition, we let Z be a binary instrument in this example (whether an applicant is assigned to an above or below median stringency screener) although in practice we will use a continuous variable.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

5. ALGORITHM DESIGN

Having discussed our general empirical strategy, we now provide an overview of the specific algorithms we consider. In Appendix A, we provide additional information regarding our sample and feature construction, as well as model training, fitting, and updating procedures.

5.1. *Preliminaries*

We begin by clarifying some relevant issues for all models we consider.

We divide our data into two periods, the first consisting of the 48,719 applicants arriving before 2018 (2,617 of whom are interviewed), and the second consisting of the 39,947 applications that arrive in 2018–9 (2,275 of whom are interviewed). We think of the 2016–7 period as our "training" data and the 2018–9 period as our "analysis" period. This approach to defining a training dataset (rather than taking a random sample of our entire data) most closely approximates a real world setting in which firms would likely use historical data to train a model that is then applied prospectively. We also continue to update our models during the 2018–9 analysis sample by adding the outcomes of the applicants it selects back into its training data.

Our goal is to predict applicants' hiring potential, Y, as defined in Section 4.1. Because hiring potential is only directly observed for applicants who are interviewed, we train our models using data from interviewed applicants only. We note that not all screening algorithms are trained in this way. One common approach is to predict the human recruiter's interview decision, I, rather than the applicant's hiring potential Y. In this case, vendors are able to train their data on all applicants. Another common approach is to focus on predicting hiring likelihood Y, but to set $Y = 0$ for all applicants who are not interviewed. This essentially assumes that candidates who were not interviewed had low hiring potential. We choose not to follow either of these approaches because they conflate a recruiter's decision I with the quality of that decision Y. Rambachan and Roth (2020), in addition, show that such approaches tend to be more biased against racial minorities.

Next, we acknowledge that our measures of quality—hire and offer outcomes—are based on the discretion of managers and potentially subject to various types of evaluation biases (Castilla, 2011; Quadlin, 2018). Indeed, many on-the-job performance metrics, such as performance evaluations, would also be subject to this concern. Without a truly "objective" measure of quality, we interpret our results as asking whether ML tools can improve firm decisions, as measured by its own revealed preference metrics.

Finally, our models may generate inaccurate predictions if the relationship between covariates X and hiring likelihood Y differs between the full applicant sample and the interviewed subset. While there are a growing set of advanced ML tools that seek to correct for training-sample selection,[24] testing these approaches is outside of the scope of this article and we are not aware of any commercially available algorithms that employ sample selection correction. We provide evidence that selection on unobservables does not appear to be a large concern in Section 6.2.2.

In general, we emphasize that the ML models we build should not be thought of as "optimal" in either their design or performance, but as an example of what could be feasibly achieved by most firms that are able to organize their administrative records into a modest training dataset, with a standard set of resume-level input features, using a technically accessible ML toolkit.

24. See, for example, Dimakopoulou *et al.* (2018, 2019) which discuss doubly robust estimators to remove sample selection and Si *et al.* (2023).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

5.1.1. Supervised learning. Our first model uses a standard supervised learning approach to predict an applicant's likelihood of being hired, conditional on being interviewed. We begin with an initial training dataset, D_0 and use it to form an estimate of applicant quality $\hat{E}[Y_{it}|X'_{it}; D_0]$ using a L1-regularized logistic regression (LASSO). Appendix Figure A.I plots the receiver operating characteristic (ROC) curve and its associated AUC, or area under the curve. This model has an AUC of 0.64, meaning that it will rank an interviewed applicant who is hired ahead of an interviewed but not hired applicant 64% of the time. We also plot the confusion matrix in Appendix Figure A.II, which provides more information on the model's classification performance.

Having trained this initial model on 2016–7 data, we use it to make interview decisions for future applicants. It is common for firms to train once and then continue to use the same static model. In our article, however, we estimate a dynamic SL model that updates the firm's training data with the outcomes of applicants it selects later on. Specifically, we divide our analysis sample (2018–9) into "rounds" of 100 applicants. After each round, we take the applicants the model has selected and update its training data. We then retrain the model and use its updated predictions to make selection decisions in the next round. At any given point t, the SL model's interview policy is as follows, based on equation (4.1) of our conceptual framework:

$$I_{it}^{\text{SL}} = \mathbb{I}(s_t^{\text{SL}}(X'_{it}) > c_t), \quad \text{where } s_t^{\text{SL}}(X'_{it}) = \hat{E}[Y_{it}|X'_{it}; D_t^{\text{SL}}]. \tag{5.3}$$

Here, D_t^{SL} 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, $D_{t+1}^{\text{SL}} = D_t^{\text{SL}} \cup (I_t^{\text{SL}} \cap I_t)$. 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 actual human recruiters. This may impact the degree to which the SL model can learn about the quality of the applicants selected, relative to a world in which hiring potential is fully observed for all applicants. We discuss this in more detail shortly, in Section 5.2.

5.1.2. Upper confidence bound. While there is, in general, no generic optimal strategy for the contextual bandit model described in Section 4.1, it is widely known that exploitation-only approaches—such as the SL model described above—are inefficient solutions because they do not factor the ex post value of learning into their ex ante selection decisions (Dimakopoulou *et al.*, 2018; Bastani *et al.*, 2021). An emerging literature in computer science has therefore focused on developing a range of computationally tractable algorithms that incorporate exploration.[25]

The particular exploration-based implementation we use is an Upper Confidence Bound Generalized Linear Model (UCB-GLM) described in Li *et al.* (2017). We choose this approach because it best fits our setting. UCB-GLM works well when the relationship between "context"

25. The best choice of algorithm for a given situation will depend on the number of possible actions and contexts, as well as on assumptions regarding the parametric form relating context to reward. For example, recently proposed contextual bandit algorithms include UCB (Auer, 2002), Thompson Sampling (Agrawal and Goyal, 2013), and LinUCB (Li *et al.*, 2010). In addition, see Agrawal and Goyal (2013) and Bastani and Bayati (2020). Furthermore, the existing literature has provided regret bounds—*e.g.* the general bounds of Russo and Roy (2016), as well as the bounds of Rigollet and Zeevi (2010) and Slivkins (2014) in the case of non-parametric function of arm rewards—and has demonstrated several successful applications areas of application—*e.g.* news article recommendations (Li *et al.*, 2010) or mobile health (Lei *et al.*, 2022). For more general scenarios with partially observed feedback, see Rejwan and Mansour (2020) and Bechavod *et al.* (2019).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

variables (covariates X_{it}) and "reward" (hiring potential, Y_{it}) follows a generalized linear functional form ($E[Y_{it}|X'_{it}] = \mu(X'\theta^*_t)$): we measure Y_{it} as a binary hiring outcome and estimate $E[Y_{it}|X'_{it}]$ using a logistic regression. Under these circumstances, Li *et al.* (2017) provides the algorithm implementation we follow and shows that it is asymptotically regret-minimizing.[26]

Specifically, our UCB algorithm scores applicant i in period t as follows:

$$I_{it}^{\text{UCB}} = \mathbb{I}(s_t^{\text{UCB}}(X'_{it}) > c_t), \quad \text{where } s_t^{\text{UCB}}(X'_{it}) = \hat{E}[Y_{it}|X'_{it}; D_t^{\text{UCB}}] + \alpha B(X'_{it}; D_t^{\text{UCB}}). \quad (5.4)$$

In equation (5.4), the scoring function $s_t^{\text{UCB}}(X'_{it})$ is a combination of the algorithm's expectations of an applicant's quality based on its training data and an "exploration bonus" given by:

$$B(X'_{it}; D_t^{\text{UCB}}) = \sqrt{(X_{it} - \bar{X}_t)' V_t^{-1} (X_{it} - \bar{X}_t)}, \quad \text{where } V_t = \sum_{j \in D_t^{\text{UCB}}} (X_{jt} - \bar{X}_t)(X_{jt} - \bar{X}_t)'. \quad (5.5)$$

Intuitively, equation (5.4) breaks down the value of an action into an exploitation component and an exploration component. In any given period, a strategy that purely focuses on exploitation would choose to interview a candidate on the basis of her expected hiring potential: this is encapsulated in the first term, $\hat{E}[Y_{it}|X'_{it}; D_t^{\text{UCB}}]$. Indeed, this is essentially the scoring function for the SL model, described in equation (5.3). Meanwhile, a strategy that purely focuses on exploration would choose to interview a candidate on the basis of the distinctiveness of her covariates: this is encapsulated in the second term, $B(X'_{it}; D_t^{\text{UCB}})$, which shows that applicants receive higher bonuses if their covariates deviate from the mean in the population $(X_{it} - \bar{X}_t)$, especially for variables X'_{it} that generally have little variance in the training data (*e.g.* weighted by the precision matrix V_t^{-1}). To balance exploitation and exploration, equation (5.4) combines these two terms. As a result, candidates are judged on their mean expected quality *plus* their distinctiveness from the existing training data. The term α captures the weight that we put on the exploration component relative to the exploitation component. Taken together, Li *et al.* (2017) show that this provides an upper bound on the confidence interval associated with an applicant's true quality, given the training data D_t^{UCB}, hence the term UCB. In our model, we follow the approach described in Li *et al.* (2017) and choose an α of 1.96 so that we are, in fact, using the upper 95th percentile bound. In essence, UCB approaches are based on the principle of "optimism in the face of uncertainty," favouring candidates with the highest statistical upside potential.[27]

At time $t = 0$ of our analysis sample, our UCB and SL models share the same predicted quality estimate, which is based on the baseline model trained on the 2016–7 sample. As with the SL model, we update the UCB model's training data with the outcomes of applicants it has selected during the 2018–9 analysis period. Based on these new training data, the UCB algorithm updates both its beliefs about hiring potential and the bonuses it assigns. As was the case with the SL model, we can only add applicants who are selected by the model and also interviewed in practice.

5.1.3. Model comparisons. A large theoretical literature shows that exploration-based model such as UCB will outperform exploitation-only based approaches in the long run (Dimakopoulou *et al.*, 2018). Li *et al.* (2017) prove that the specific model we adopt, UCB-GLM

26. See Equation 6 and Theorem 2 of their paper.

27. The basic UCB approach for non-contextual bandits was introduced by Lai and Robbins (1985) and, since then, various versions of this approach have been developed for different types of contextual bandit settings, and shown to be regret minimizing.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

will asymptotically minimize regret via more efficient learning: that is, it will select applicants with greater hiring potential. Yet while a UCB based approach is expected to out-perform SL models in the long run, the quality differences we would observe in practice capture both the long term benefits of learning and the short-term costs of exploration. This tradeoff will also depend on the specifics of our empirical setting. In particular, if quality is not evolving and there is relatively rich initial training data, SL models may perform as well as, if not better than, UCB models because the value of exploration will be limited. If, however, the training data were sparse or if the predictive relation between context and rewards evolves over time, then the value of exploration is likely to be greater.

In terms of diversity, our UCB algorithm favours candidates with distinctive covariates because this helps the algorithm learn more about the relationship between applicant covariates and hiring outcomes. This suggests that a UCB model would, at least in the short run, select more applicants from demographic groups that are under-represented in its training data, relative to an SL model.[28] Over time, however, exploration bonuses will decline as the model receives more information about applicants of all types. As a result, long run differences between SL and UCB models will be primarily driven by differences in their beliefs about applicant quality. Gains in diversity driven by exploration bonuses will not be sustainable if minority applicants are actually weaker.

Our main results will come from a 16-month period from January 2018 to April 2019. Because most organizations cannot afford to care only about the long run impacts of new hiring policies, it becomes important to empirically examine how exploration-based algorithms behave over medium-run time scales.

5.2. *Feasible versus live model implementation*

In a live implementation, each algorithm would select which applicants to interview, and the model would be updated with the outcomes of all selected candidates. In our retrospective analysis, we are only able to update our models with outcomes for ML selected candidates who were actually interviewed in practice. Here, we discuss how the actual implementation of our models—which we term "feasible" SL or UCB—may differ from a live implementation.

For concreteness, suppose that the UCB model wants to select 50 theatre majors but, in practice, only five such applicants were actually interviewed. In our feasible implementation, we would only be able to update the UCB's training data with the outcomes of these five applicants, whereas in a live implementation, we would be able to update with outcomes for all 50 UCB-selected candidates.

If humans are not selecting on unobservables, the feasible UCB's estimate of the quality of theatre majors would be the same as the live UCB's estimate but, because it observes five rather than 50 instances, its estimates would be less precise. This would impact the exploration in the next period: even though it has the same beliefs about quality, the feasible UCB would select more theatre majors in the next period because its uncertainty about these applicants is higher. This would slow down the learning of the feasible UCB model relative to the live model but, with a large enough sample, both should converge to the same beliefs and actions.[29]

28. We note, however, that we calculate bonuses over all covariates. If White and Asian applicants are more heterogeneous along other dimensions such as education and work history, then they may nonetheless receive high exploration bonuses.

29. Formally, the distinction between the feasible and live versions of our ML models is related to regression in which outcomes are missing at random conditional on unobservables. Under the assumption of no selection on unobservables, common support, and well-specification of the regression function (in our case, the logit), the feasible and live

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

This analysis changes if human screeners select based on unobservables. Suppose the five theatre majors who are interviewed are unobservably better than the 45 theatre majors who are not interviewed. A feasible UCB model would then be too optimistic about the quality of this population relative to a live UCB model that correctly learns the quality of all 50 applicants. In the next period, the feasible UCB model would select more theatre majors both because uncertainty remains higher and because positive selection on unobservables induces upwardly biased beliefs. This latter bias can lead our approach to select too many applicants from groups whose weaker members are screened out of the model's training data by human recruiters. In Section 6.2.2, we provide IV-based evidence that human recruiters do not appear to be selecting on unobservables. In addition, Section 9.2 shows simulation results that more closely approximate a live implementation in which we can update outcomes for all selected candidates.

6. MAIN RESULTS

6.1. *UCB and SL versus human recruiters: diversity of selected applicants*

We begin by assessing the impact of each policy on the diversity of candidates selected for an interview in our analysis sample. This is done by comparing $E[X|I = 1]$, $E[X|I^{\text{SL}} = 1]$, and $E[X|I^{\text{UCB}} = 1]$, for various demographic measures X, where we choose to interview the same number of people as the actual recruiters in a given year-month. We observe demographic covariates such as race or ethnicity and gender for all applicants, regardless of their interview status, and do not face a selective labels problem for comparisons of demographics.

We focus on the racial and ethnic composition of selected applicants. Figure 1A shows that, at baseline, 58% of applicants in our analysis sample are Asian, 29% are White, 9% are Black, and 4% are Hispanic. Figure 1B shows that human recruiters select a similar proportion of Asian and Hispanic applicants (57% and 5%, respectively), but relatively more White and fewer Black applicants (34% and 5%, respectively). In Figure 1C, we show that the SL model reduces the share of Black and Hispanic applicants from 10% to under 5%, White representation increases more modestly from 34% to 40%, and Asian representation stays largely constant. In contrast, Figure 1D shows that the UCB model increases the Black share of selected applicants from 5% to 16%, and the Hispanic share from 4% to 9%. The White share stays constant, while the Asian share falls from 57% to 43%.

Appendix Figure A.VII plots the same set of results for gender. Panel B shows that 66% of interviewed applicants are men and 34% are women; this is largely similar to the gender composition of the overall applicant pool. Unlike the case of race or ethnicity, both our ML models are aligned in selecting more women than human recruiters, increasing their representation to 41% (SL) or 38% (UCB).

Next, we explore why our UCB model selects more Black and Hispanic applicants. Appendix Figure A.VIII shows that Black and Hispanic applicants receive slightly larger exploration bonuses on average. This reflects both direct differences in population size by race or ethnicity, as well as indirect differences arising from the correlation between race or ethnicity and other variables that also factor into bonus calculations.

A crucial question raised by this analysis is whether these differences in diversity are associated with differences in applicant quality. We will discuss this extensively in the next section and provide evidence that, despite their demographic differences, hiring outcomes for applicants

versions of our models should both be consistent estimators of the underlying parameter θ^* linking covariates with hiring outcomes: $E[Y_{it}|X'_{it}] = \mu(X'_{it}\theta^*_t)$ (Robins *et al.*, 1995; Wang *et al.*, 2010). In a finite sample, of course, the point estimates of the feasible and live models may differ.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

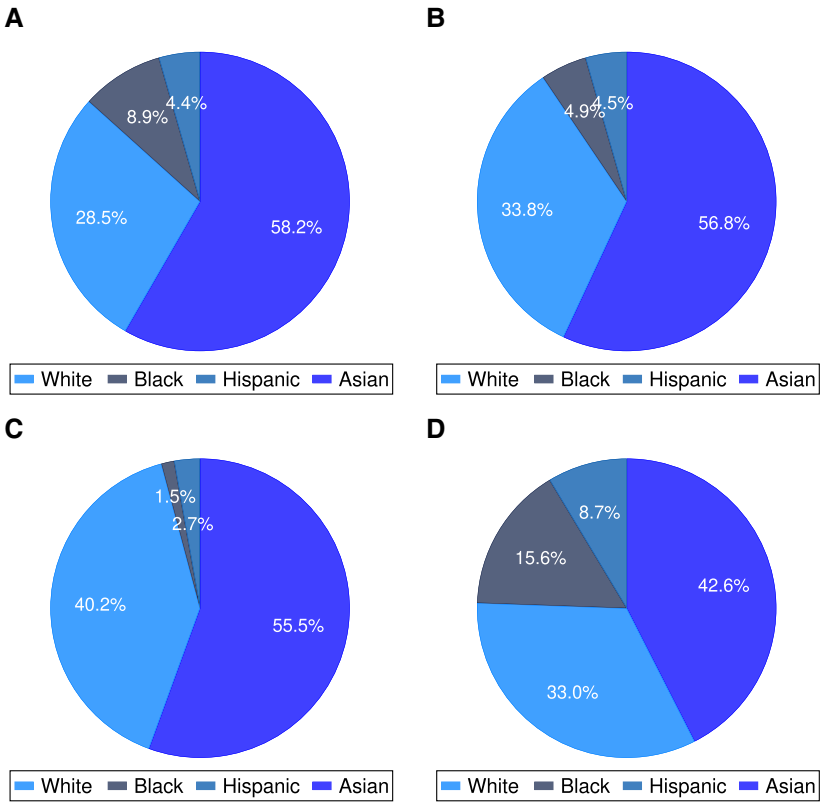


FIGURE 1
Racial composition A. Applicant Pool. B. Actual Interview. C. SL Model. D. UCB Model

Notes: (A) The race/ethnicity composition of applicants in our data is shown. (B) The composition of applicants actually selected for an interview by the firm is shown. (C) The racial composition of applicants who would be selected if chosen by the supervised learning algorithm described in equation (5.3) predicting hiring potential is shown. Finally, (D) shows the composition of applicants who would be selected for an interview by the UCB algorithm described in equation (5.4) predicting hiring potential. By construction, all methods are constrained to match the number of applicants interviewed by human recruiters. Applicants' demographic information is collected by our firm during the application process.

Source: All data come from the firm's application and hiring records.

selected by our SL and UCB models are comparable to each other, and much better than those selected by human recruiters.

6.2. *UCB and SL versus human recruiters: quality of selected applicants*

While we observe demographics for all applicants, we only observe hiring potential H for applicants who are actually interviewed. We therefore cannot directly observe hiring potential for applicants selected by an algorithm, but not by the human reviewer. To address this, we take three complementary approaches, described previously in Section 4. Across all three approaches, we find evidence that both SL and UCB models would select applicants with greater hiring potential, relative to human screening.

6.2.1. Interviewed sample. Our first approach restricts to the sample of applicants who are interviewed, for whom we directly observe hiring outcomes. Among this set, we directly observe

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

ML scores s^{SL} and s^{UCB}. We do not, however, directly observe the implicit score that human recruiters give each candidate. Before continuing, we therefore need to generate an estimate of "s^H," an applicant's propensity to be selected for an interview by a human recruiter. To do this, we simply generate a model of $E[I|X]$, where $I \in \{0, 1\}$ are realized human interview outcomes, using same logistic LASSO approach described in Section 5.1.1.[30] We describe this model construction and the training and procedure in Appendix Section A.3.4. Appendix Figure A.III 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 77% of the time.[31]

Figure 2 plots a binned scatterplot depicting the relationship between algorithm scores and hiring outcomes among the set of interviewed applicants; each dot represents the average hiring outcome for applicants in a given scoring bucket. Among those who are interviewed, applicants' human scores are uninformative about their hiring likelihood; if anything this relationship is slightly negative.[32]

In contrast, all ML scores have a statistically significant, positive relation between algorithmic priority selection scores and an applicant's (out of sample) likelihood of being hired.[33]

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

For consistency, Appendix Figure A.IX revisits our analysis of diversity using the same type of selection rule described in this section: specifically, picking the top 50% of candidates among the set of interviewed. Again, we find that UCB selects a substantially more diverse set of candidates than the SL model.

6.2.2. Full sample. A concern with our analysis of the $I = 1$ sample is that human recruiters may add value by screening out particularly poor candidates so that they are never observed in the interview sample to begin with. In this case, then we may see little relation

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

31. Although a "good" AUC number is heavily context specific, a general rule of thumb is that models with an AUC in the range of 0.75–0.85 have acceptable discriminative properties, depending on the specific context and shape of the curve (Fischer *et al.*, 2013).

32. This weak relation between human preferences and outcomes is consistent with existing work documenting that humans often have incorrect perceptions of worker quality. For instance, Hoffman *et al.* (2017) find that firms see worse hiring outcomes when humans make exceptions to algorithmic suggestions. In a study of personnel assessment, Yu and Kuncel (2020) find that the scores of expert human resource managers were weakly related to on-the-job performance. Similarly, in a study of recruiters for software engineering positions, Lerner (2024) found a weak correlation between recruiter ranking and eventual hiring likelihood.

33. Appendix Table A.II shows these results as regressions to test whether the relationships are statistically significant.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

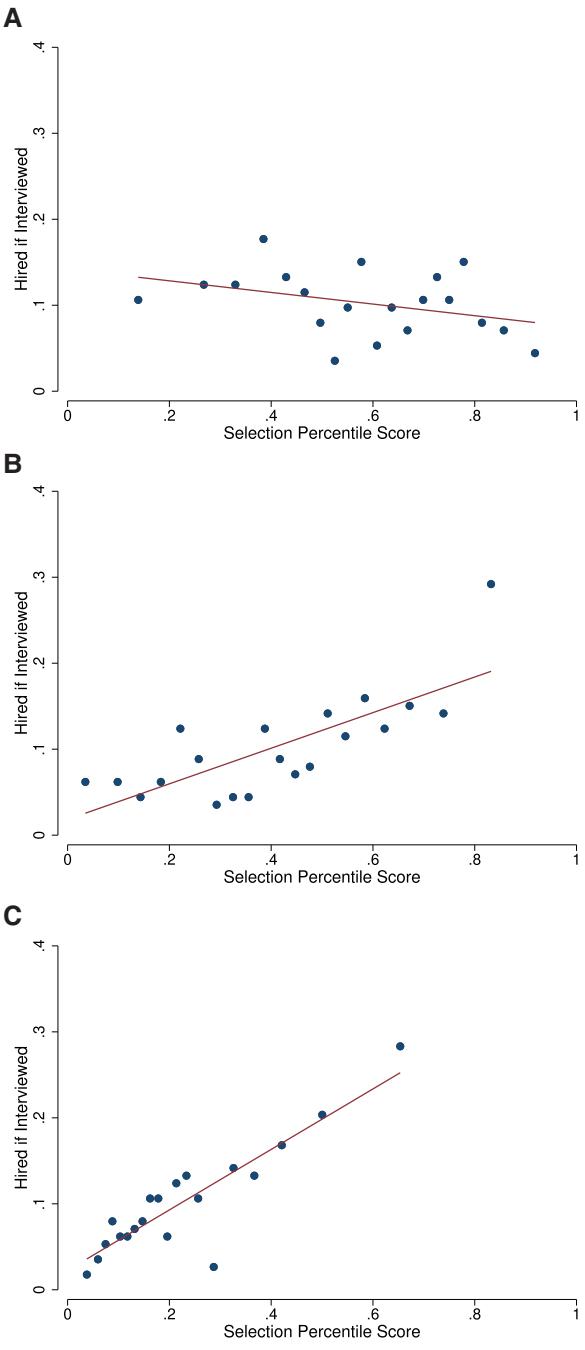


FIGURE 2

Correlations between algorithm scores and hiring likelihood A. Human. B. SL Model and C. UCB Model

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. (A) The selection scores from an algorithm that predicts human recruiters interview selection choices are shown. (B) The selection scores from the supervised learning algorithm described by equation (5.3) are shown. (C) The selection scores from the UCB algorithm described in equation (5.4) are shown.

Source: All data come from the firm's application and hiring records.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

TABLE 3
Predictive accuracy of Human versus ML models, among Interviewed Applicants

Selectivity (Top X%)	Overlap % (1)	Both (2)	Human only (3)	SL only (4)
A. Human versus updating SL				
25	13.33	18.52	6.83	17.78
50	37.22	10.99	7.47	16.67
75	64.93	10.31	4.67	18.68
Selectivity (Top X%)	Overlap % (1)	Both (2)	Human Only (3)	UCB Only (4)
B. Human versus UCB				
25	15.72	17.95	6.46	20.33
50	36.00	12.09	6.33	16.57
75	61.28	10.76	3.89	16.30
Selectivity (Top X%)	Overlap % (1)	Both (2)	SL Only (3)	UCB Only (4)
C. Updating SL versus UCB				
25	42.43	23.39	9.91	14.22
50	60.59	15.33	8.21	10.71
75	74.43	13.14	5.98	5.98

Notes: This table shows the hiring rates of each algorithm when they make the same recommendation or differing recommendations. The top panel compares the human versus SL algorithm, the middle panel compares the human versus the UCB algorithm, and the lower panel compares the SL versus the UCB algorithm. "Human" refers to our model of human recruiter interview propensity introduced in Section 4.2. Each row of a given panel conditions on selecting either the top 25%, 50%, and 75% of applicants according to each of the models. For the two algorithms being compared in a given panel, Column 1 shows the percent of selected applicants that both algorithms agree on. Column 2 shows the share of applicants hired when both algorithms recommend an applicant, and Columns 3 and 4 show the share hired when applicants are selected by only one of two algorithms being compared.
Source: All data come from the firm's application and hiring records.

between human preferences and hiring potential among those who are interviewed, even though human preferences are highly predictive of quality in the full sample.

To address this, we compute estimates of the hiring likelihood of all ML selected applicants, using the inverse propensity weighting (IPW) approach described earlier in Section 4.2. Our results are presented in Figure 3. Among those selected by human recruiters, the average observed hiring likelihood is 10%. In contrast, our calculations show that ML models select applicants with almost three times higher predicted hiring potential. In particular, the average expected hiring likelihood for applicants selected by the UCB model is 27% and 32% for the SL model. The slightly weaker performance of the UCB model may be explained by the fact that an emphasis on exploration means that the UCB algorithm may select weaker candidates, particularly in earlier periods. Together, this set of results is consistent with our findings from the interviewed-only subsample: the hiring yield of ML algorithms are similar to each other and, in all cases, better than the human decision-maker. We find no evidence that the gains in diversity that we document in Section 6.1 come at the cost of substantially reducing hiring rates among selected applicants.

6.2.3. Testing IPW assumptions. This analysis relies on two assumptions—no selection on unobservables and a common support between ML and human preferences.

First, we test for the presence of selection on unobservables using variation from random assignment to lenient and strict reviewers, as described in Section 4.2.3. The logic is

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

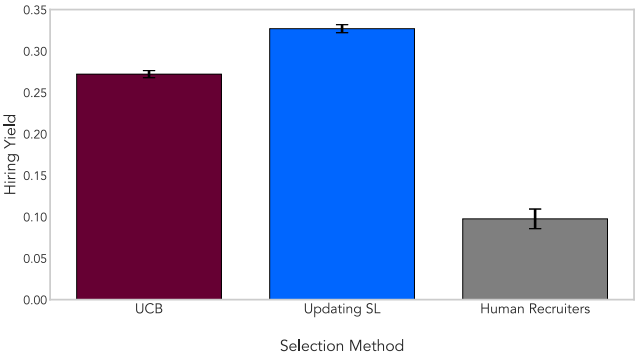


FIGURE 3
Average hiring likelihood

Notes: This figure shows our inverse propensity weighting estimates of $E[Y|I^{\text{ML}} = 1]$ for each algorithmic selection strategy (SL or UCB), alongside actual hiring yields from human recruiter decisions. Our inverse propensity weighting estimation method is described in Section 4.2. We also plot the 95% confidence intervals around each estimate of hiring yield.

Source: All data come from the firm's application and hiring records.

as follows: if humans are, on average, positively selecting candidates on observables, then it should be the case that applicants selected by more stringent reviewers—*e.g.* those who are subjected to a higher human threshold—should be more likely to be hired conditional on being interviewed than those selected by more lenient reviewers. That is, if there is a positive (negative) relationship between human selection propensities and hiring outcomes, then going further down the distribution by selecting more candidates should decrease (increase) average quality.

Figure 4 plots the relationship between screener leniency and hiring outcomes. Figure 4A does not control for applicant observables while Figure 4B does. In both cases, there appears to be little relationship between leniency and hiring outcomes, suggesting that strict reviewers do not appear to select a stronger set of candidates. If anything, there appears to be a slightly positive relation, suggesting that lenient reviewers may weakly select a stronger group of stronger candidates. The possibility that human recruiters are actively bad at selecting candidates is consistent with our results in Figure 2A, which shows that workers preferred by humans have somewhat lower hiring rates among those who were interviewed. We note that both these figures include fixed effects for job family, job level, work location, and application year so that this slightly positive association is not being driven by confounding differences in hiring demand across positions or times. Interestingly, we also note that the pattern is similar regardless of whether we control for applicant covariates: this suggests that strict screeners are not better at selecting applicants based on either observed or unobserved covariates.

Next, we test for common support among human and ML preferences. Intuitively, the IPW approach infers the quality of ML-selected candidates using actual hiring outcomes from candidates with similar covariates who were actually selected to be interviewed by a human recruiter. We therefore require that candidates selected by the ML algorithm have some non-zero probability of also being selected by human recruiters.

Appendix Figure A.X plots the distribution of a candidate's estimated propensity to be selected by a human recruiter for the set of applicants chosen by our SL and UCB models. In both cases, we find that all ML-selected applicants have a human selection propensity strictly between 0 and 1; we see no mass at or near zero.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

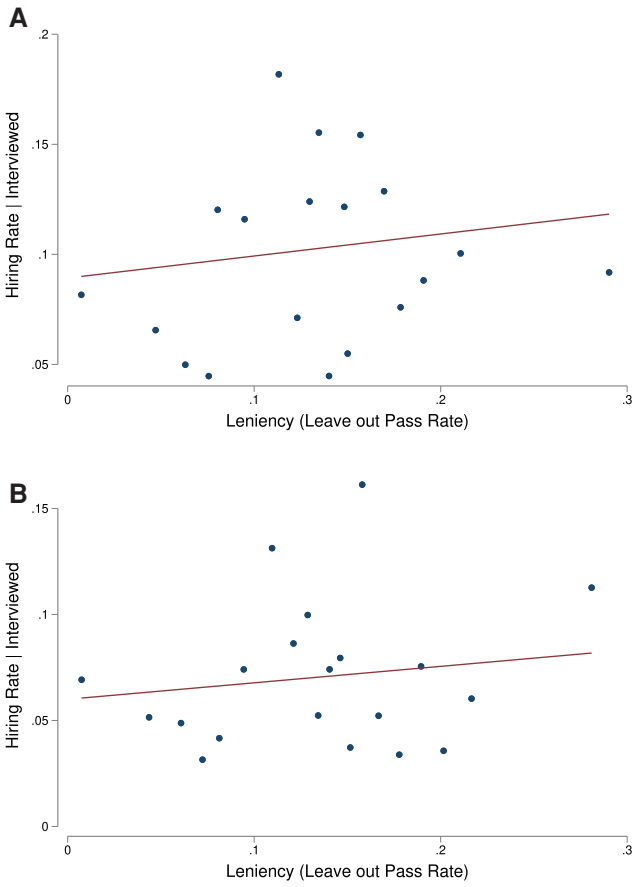


FIGURE 4
Testing for positive selection A. Full Sample, No Controls and B. Full Sample, Controls

Notes: These binned scatterplots show the relationship between the leniency of randomly assigned screeners and the hiring outcomes of the applicants they select to be interviewed. (A) This relationship, controlling for job level characteristics: type of job, seniority level, work location, and application year is plotted. (B) This relationship after adding additional controls for applicant characteristics: education, work history, and demographics is plotted.

Source: All data come from the firm's application and hiring records.

6.2.4. Marginally interviewed sample. Our final approach asks the firm to modify its current human interview decisions by following ML recommendations when evaluating "marginal" candidates. Section 4.2.3 shows that we can assess the impact of this alternative interview policy $\tilde{I}$ by comparing the characteristics of instrument compliers with high and low ML scores. Compliers can be thought of as "marginal" in that they are interviewed only because they were randomly assigned to a lenient recruiter.

Figure 5 presents our results. Figure 5A, C, and E focus on applicants who are marginally selected based on SL model scores while Figure 5B, D, and F focus on marginal applicants as defined by UCB scores. In Figure 5A and B, we see that compliers with high SL and UCB scores are both more likely to be hired than those with low scores. This indicates that, on the margin, nudging human interview decisions toward either UCB or SL preferences would increase the expected hiring yield.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

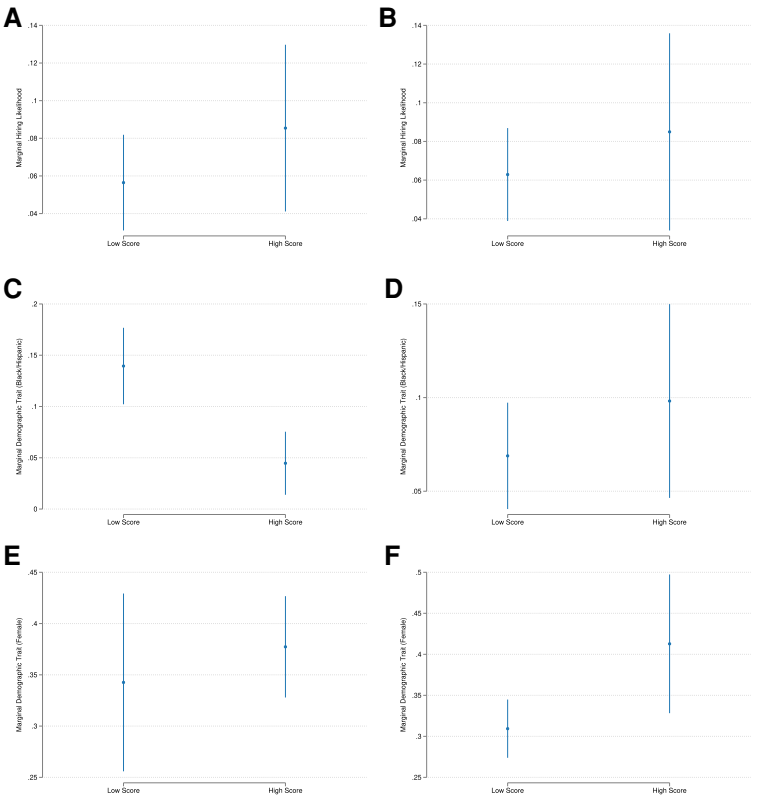


FIGURE 5

Characteristics of marginal interviewees A. Hiring Likelihood, SL. B. Hiring Likelihood, UCB. C. Black/Hispanic, SL. D. Black/Hispanic, UCB. E. Female, SL. F. Female, UCB

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 and bottom half of the UCB algorithm's score. (A, C, and E) Marginal applicants as defined by SL model scores are considered. (B, D, and F) Marginal applicants as defined by UCB model scores are considered. In (A and B), the *y*-axis is the average hiring likelihood of marginally interviewed candidates; (C and D) focus on the share of selected applicants who are Black or Hispanic; (E and F) focus on the share of selected applicants who are female. The confidence intervals shown in each panel are derived from robust standard errors clustered at the recruiter level.

Source: All data come from the firm's application and hiring records.

In the remaining panels, we consider how following ML recommendations on the margin would change the demographics of selected candidates. In Figure 5C, we see that marginally selected applicants with high SL scores are substantially less likely to identify as Black or Hispanic. As such, nudging toward SL scores would tend to decrease the racial and ethnic diversity of selected applicants, relative to existing human decisions. In contrast, Figure 5D shows the opposite for the UCB model. Here, we find that compliers with high UCB scores are more likely to be Black or Hispanic. As such, the interview policy defined by $\tilde{I}$ would increase quality and diversity on the margin, relative to the firm's current practices. In Panels E and F, we show that both the SL and UCB models would tend to increase the representation of women. Panels A, C, and D of Table 4 show the corresponding regression results.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

TABLE 4
Impacts of following ML recommendations, IV analysis

	Low UCB (1)	High UCB (2)	Low SL (3)	High SL (4)
A. Hire rates				
Marginally selected	0.0629***	0.0849***	0.0564***	0.0854***
	(0.0123)	(0.0260)	(0.0130)	(0.0226)
Observations	18,710	18,956	18,862	18,804
B. Offer rates				
Interviewed	0.108***	0.168***	0.111***	0.183***
	(0.0220)	(0.0341)	(0.0248)	(0.0326)
Observations	18,538	19,128	18,417	19,249
C. Share Black or Hispanic				
Marginally selected	0.0689***	0.0982***	0.139***	0.0447***
	(0.0145)	(0.0264)	(0.0190)	(0.0157)
Observations	18,710	18,956	18,862	18,804
D. Share female				
Marginally selected	0.309***	0.413***	0.343***	0.377***
	(0.0181)	(0.0431)	(0.0442)	(0.0252)
Observations	18,710	18,956	18,862	18,804

Notes: This table examines the characteristics of marginally interviewed applicants according to our IV strategy described in the text. Specifically, each number represents the result of the regressions outlined in equation (3). The reported coefficients are the IV estimates of the coefficient on whether an applicant is interviewed and can be interpreted as the average outcome variable among treatment compliers. For example, the coefficients in Panel A Columns 1 and 2 represent the estimated average hiring rates of IV compliers with low and high UCB scores, respectively. Applicants receiving high or low scores are those who are above or below the median, respectively. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.
Source: All data come from the firm's application and hiring records.

These results are again consistent with our earlier results. In both cases, following UCB recommendations can increase hiring yield and diversity relative to the firm's present policies, while following traditional SL recommendations increases quality but decreases racial and ethnic diversity.

7. ALTERNATIVE MEASURES OF QUALITY

7.1. *Maximizing offer rates*

In our main analyses, we focus on screening models that are designed to maximize hiring yield. This is our preferred specification as it captures the key reason why firms turn to algorithms in the first place: the desire to fill vacancies with qualified workers.

Hiring requires that a worker both receive and accept a job offer. To isolate an algorithm's ability to identify applicants a firm would *like* to hire, we build an alternative set of UCB and SL models that maximize the likelihood that an applicant is extended an offer, regardless of whether they accept. These models are trained in the same way as our main models, except using receiving an offer as the outcome variable of interest. Appendix Figure A.V shows that we correctly predict offer outcomes in our baseline training data approximately 68% of the time. Appendix Sections A.3.3 and A.3.2 provide additional details on the training and out-of-sample accuracy of our offer model.

In Figure 6, we show that offer-based SL and UCB models behave similarly to our hire-based models. Figure 6A and B compare the demographics of applicants selected under SL and UCB models. Similar to our main results in Figure 1, we find that the SL model dramatically

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

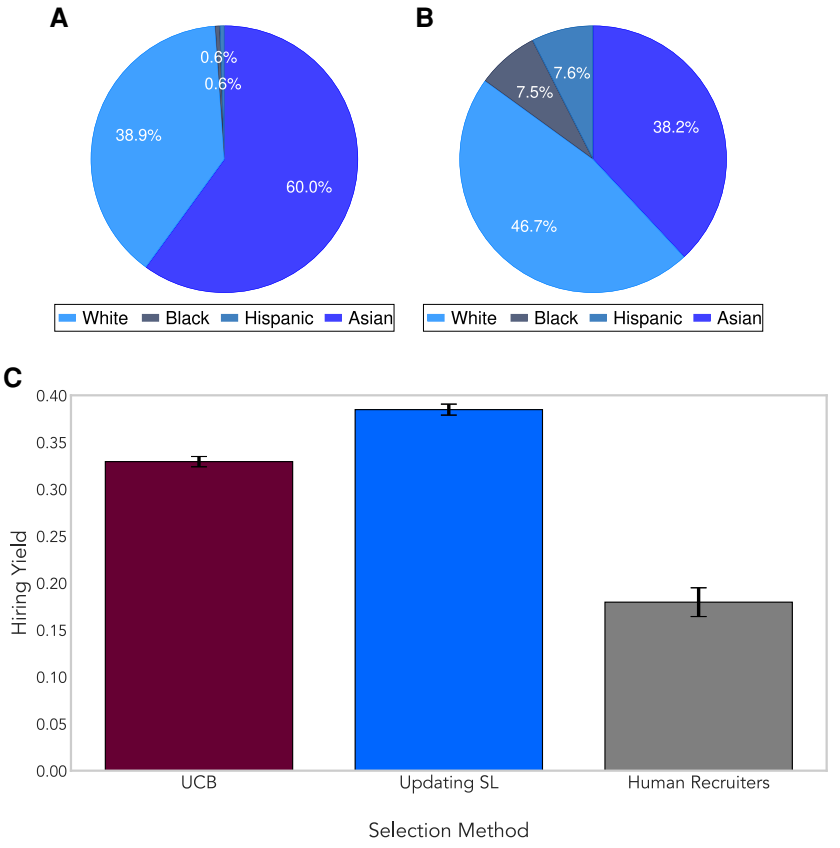


FIGURE 6

Racial composition—offer model A. SL Offer Model. B. UCB Offer Model. C. Average Offer Rate

Notes: (A) The race/ethnicity composition of an interview pool selected by a supervised learning algorithm described in equation (5.3) predicting offer potential is shown. (B) The composition of applicants who would be selected for an interview by the UCB algorithm described in equation (5.4) predicting offer is shown. By construction, all methods are constrained to match the number of applicants interviewed by human recruiters. Applicants' demographic information is collected by our firm during the application process. (C) The quality (measured as the percentage of selected applicants who receive an offer using inverse propensity reweighting method described in Section 4.2) for the two ML models, as well as the true offer yield from human interview decisions is compared.

Source: All data come from the firm's application and hiring records.

reduces the share of Black and Hispanic applicants who are selected for an interview (to less than 2% from a human recruiter baseline of just under 10%) while the UCB model increases this share to approximately 15%). In Figure 6C, we compare the average offer rate of UCB, SL and human selected applicants, using our inverse propensity weighting estimates discussed in Section 6.2.2.[34] Consistent with Figure 3, we find that both UCB and SL models outperform human recruiters, with the SL model somewhat outperforming UCB over the 18 months of our analysis period. Panel B of Table 4 shows the corresponding regression results.

Appendix Figure A.XV plots the correlation between UCB and SL scores and offer rates, among the set of applicants who are interviewed, analogous to the results presented for the

34. Appendix Figure A.XI shows that all offer model selected applicants have a human selection propensity strictly between 0 and 1 with no mass at or near zero.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

hire model in Figure 2. Similarly, Appendix Figure A.XVI repeats our marginal sample IV analysis for the offered models. Our results are very similar. Appendix Figure A.XV shows that candidates with higher UCB or SL scores are more likely to receive an offer, whereas applicants preferred by human recruiters tend to have, if anything, worse offer outcomes. Appendix Figure A.XVI shows that firms can improve offer rates by following the recommendations of either ML model on the margin, but that SL recommendations decrease the share of under-represented minorities while UCB recommendations increase representation.

7.2. *On-the-job performance*

A concern with our analysis is that both hiring and offer outcomes may not be the measure of quality that firms are seeking to maximize. If firms ultimately care about on-the-job performance metrics, then they may prefer that it's 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: of the nearly 49,000 applicants in our training data, only 296 have data on job performance ratings, making it difficult to accurately build such a model. As a result, we take an alternative approach and correlate measures of on-the-job performance with ML and human preferences. If humans were trading off hiring likelihood for job performance, then our human SL score, s^H, should be positively predictive of job performance relative to s^{SL} and s^{UCB}.

Table 5 presents results using two measures: mid-year performance ratings and whether a worker was promoted. Performance ratings are given on a scale of 1 (below), 2 (at), or 3 (above) average performance and 13% of workers receive an above average rating. Eight percent of hires in our sample are promoted during the analysis period. Panel A examines the correlation between an applicant's likelihood of being selected by a human recruiter and their likelihood of receiving a top performance rating (Column 1) and or a promotion (Column 2). In both 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. In contrast, Panels B and C conduct the same exercise for each of our ML models. For our SL hired model, these correlations are positively signed but statistically insignificant. For the SL offered model, we see a positive and statistically significant correlation between scores and top performance ratings, and a zero correlation for promotions. We find a similar pattern for the UCB scores: we see a positive and sometimes statistically significant relationships between the UCB hired model score and on-the-job performance. For the offered model, we again see a positive and statistically significant correlation between scores and top performance ratings, and no correlation with promotions.

We caution that these data are potentially subject to strong sample selection due to the small proportion of workers for whom we have 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.

8. ALTERNATIVE POLICIES

So far, we have given our algorithms access to applicant's demographics and have made no restrictions on which applicants it can select. Here, we consider two alternative approaches that treat demographic information differently. The first regulates algorithmic *inputs*: it restricts the

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

TABLE 5
Correlations between human scores and on-the-job performance

	Top rating (1)	Promoted (2)
A. Human scores		
Human SL score	−0.282**	−0.0961
	(0.116)	(0.0782)
Observations	180	233

	Top rating		Promoted	
	(1)	(2)	(3)	(4)
B. SL scores				
SL hired	0.0791		0.0816	
	(0.103)		(0.0641)	
SL Offered		0.168**		−0.0170
		(0.0800)		(0.0537)
Observations	180	180	233	233
	Top rating		Promoted	
	(1)	(2)	(3)	(4)
C. UCB scores				
UCB hired	0.0377		0.161***	
	(0.106)		(0.0619)	
UCB Offered		0.163*		−0.0245
		(0.0850)		(0.0576)
Observations	180	180	233	233

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 performance data. "Top rating" refers to receiving a 3 on a scale of 1–3 in a mid-year evaluation. "Promoted" indicates an applicant receives a promotion. Robust standard errors are shown in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.
Source: All data come from the firm's application and hiring records.

model's ability to access information on race, gender, and ethnicity. The second approach regulates algorithmic *outputs*: it maintains access to demographic information, but imposes a quota on which applicants the model can select.

These policies correspond to principles of discrimination law. In the U.S., the Equal Opportunity Employment Commission (EEOC) looks for "disparate treatment" (treating applicants differently on the basis of their demographics) or "disparate outcomes" (success rates that are substantially different by demographic group). This is similar to the European Union's Equal Treatment Directive, which prohibits "direct discrimination" and "indirect discrimination." Blinding algorithms are a way of preventing disparate treatment by regulating algorithmic inputs while quotas are a way of preventing disparate outcomes by regulating algorithmic outputs.

8.1. *Demographics blinding*

Our main algorithms are trained on a variety of applicant characteristics, including explicit information on race, ethnicity, and gender. As a result, these models can treat applicants differently on the basis of protected categories, a legal area (Kleinberg *et al.*, 2019). It therefore natural to ask how our results would change if we eliminated the use of race, ethnicity, and gender as model

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

inputs.[35] Demographics-blind algorithms can also be useful in settings where firms do not have access to these data, either because applicants choose not to provide it or where collecting data on demographics is restricted.

The impact of blinding is difficult to predict because demographic information enters the UCB model in two ways: as features of the model that are used to predict quality and as inputs in calculating exploration bonuses. Eliminating this information can therefore shifts the model's predictive abilities as well as its exploration behaviour. To examine what occurs in our setting, we re-estimate the UCB model without applicants' race, gender, and ethnicity in either prediction or bonus provision. As a practical matter, we continue to allow the inclusion of other variables, such as geography, which may be correlated with race and ethnicity.

Figure 7 shows how this blinding impacts diversity.Figure 7A reproduces the composition of applicants selected by the unblinded UCB model and Figure 7B displays the blinded results. Blinding reduces the share of selected applicants who are Black or Hispanic, from 24% to 14%, although there is still greater representation relative to human hiring (10%). The most stark differences come in the treatment of White and Asian applicants. In the non-blinded model, White and Asian applicants make up a similar share of interviewed applicants (33% and 43%, respectively), even though there are substantially more Asian applicants in the overall pool. When the algorithm is blinded, however, many more Asian applicants are selected relative to White applicants (63% versus 23%, recalling that Asian and White applicants make up 57% and 30% of the applicant pool at large, respectively).

Appendix Figure A.XVII provides additional analysis of how blinding impacts UCB model scores. Appendix Figure A.XVIIA shows that blinding decreases exploration bonuses for Black and Hispanic applicants while increasing bonuses for Asian applicants.

In the demographics-aware model, Asian applicants received smaller bonuses because they share a covariate—being Asian—that is very common in the sample. When the algorithm is no longer able to observe this common trait, Asian applicants appear more distinctive because of their less common work and educational backgrounds. Appendix Figure A.XVIIB plots the correlation between an important covariate—attending a highly ranked school—and UCB beliefs about an applicant's quality. Under the demographics-aware model, there is substantial heterogeneity in the "returns" to school rank across demographic groups, with Asian applicants being the least rewarded for having attended a top school. Blinding, however, prevents the model from assigning race-specific returns; in our data, this increases the relative returns to elite education among Asian applicants. Taken together, these findings provide intuition for the large increase in Asian representation under blinding.

Figure 7C examines the accuracy of blinded versus unblinded UCB, using the reweighting approach described in Section 4.2. We find that blinding leads to a small, modest decline in the quality of algorithmically selected candidates; both models continue to substantially outperform human evaluators. In our setting, the small difference in outcomes between the blinded and unblinded UCB models likely combines two distinct impacts. First, blinding reduces the predictive ability of our models. At the same time, Asian applicants tend to have relatively higher hire rates in our data so that, in our case, blinding shifts exploration toward a higher yield group.

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

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

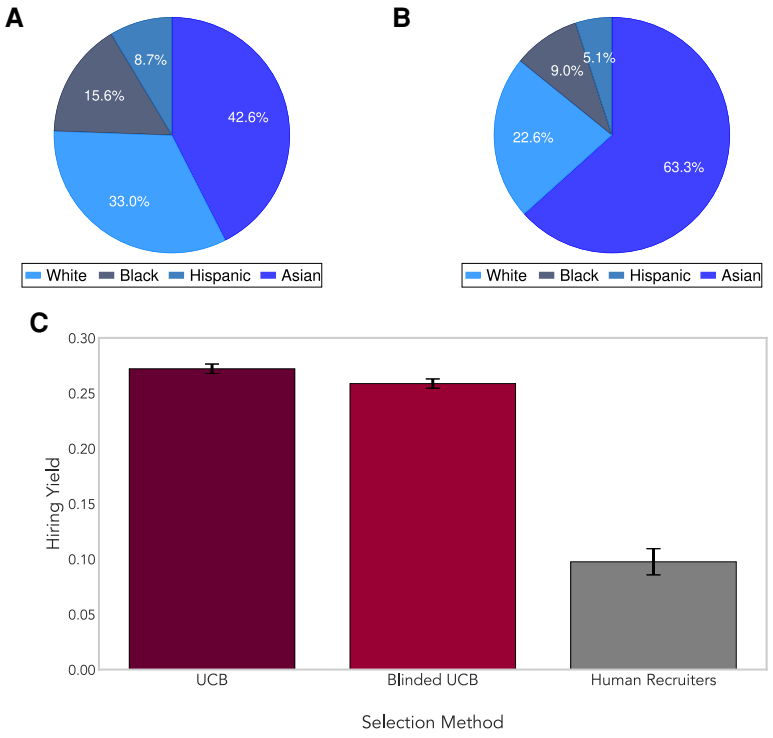


FIGURE 7

Demographics blinding A. Race/Ethnicity, UCB. B. Race/Ethnicity, Blinded UCB. C. Average Hiring Likelihood

Notes: (A) The race/ethnicity and gender composition of applicants recommended for interviews by the UCB algorithm when this algorithm explicitly incorporates race/ethnicity and gender in estimation ("UCB") is shown. (B) The composition of applicants recommended for interviews when the UCB is blinded to race/ethnicity and gender ("Blinded UCB") is shown. (C) Our inverse propensity weighting estimates of $E[Y|I^{\mathrm{ML}} = 1]$ for the demographically aware UCB, blinded UCB, and actual hiring yields from human selection decisions is shown.

Source: All data come from the firm's application and hiring records.

8.2. *Supervised learning with quota*

An alternative approach to achieving greater representation is to introduce diversity as an explicit constraint. In this section, we consider a policy in which applicants are scored by a supervised learning model with access to demographic information, but where the composition of selected applicants must reflect that of the applicant pool.

Figure 8A and B compares the demographics of candidates selected under our baseline UCB model to those selected with our SL with quota policy. By construction, the composition of applicants selected under our quota model is similar to that of the overall applicant pool (Figure 1A). We note that our percentages are not exact because we are working with small discrete numbers so it is not always possible for the share of selected applicants to equal the population share. In Figure 8C, we show that the quality of workers hired under the quota model is *substantially* worse: about 10% of selected applicants are predicted to be hired, compared with close to 30% for the unconstrained SL and UCB models.

We believe that this is due to the fact that a quota model substantially constrains the firm in terms of *when* it must select minority candidates. In settings where applicants are selected as part of a defined cohort (say, college admissions), it is straightforward to define the applicant pool

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

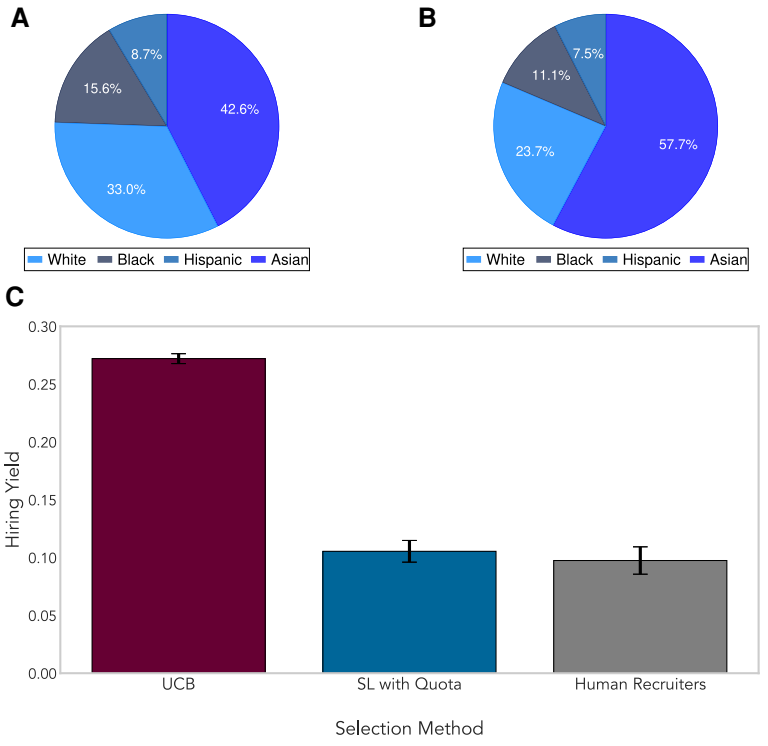


FIGURE 8
Supervised learning with quota A. Race/Ethnicity, UCB. B. Race/Ethnicity, SL with Quota. C. Average Hiring Likelihood

Notes: (A) The race/ethnicity and gender composition of applicants recommended for interviews by the UCB algorithm predicting hiring likelihood is shown. (B) The alternative of using an SL model that is constrained to select applicants in proportion to their representation in the applicant pool is considered. (C) Our inverse propensity weighting estimates of $E[Y|I^{\text{ML}} = 1]$ for the UCB versus the SL with quota, alongside actual hiring yields from human selection decisions is shown.

Source: All data come from the firm's application and hiring records.

over which the quota must be enforced. In most hiring settings, however, applicants are selected on a rolling basis, making it conceptually challenging to specify how many members of each group to select over a given period. Any ex ante constraint could reduce quality by selecting too many minority candidates when their quality is low, and too few when their quality is high.

Appendix Figure A.XVIII shows that this is a real constraint in our setting. Specifically, we compare the average number of Black or Hispanic applicants who are selected by our UCB model (dotted blue line) and the SL with quota over our analysis period (solid red line). The UCB model selects more Black or Hispanic applicants on average but varies significantly in the number it selects each period. By contrast, the quota model is restricted to selecting, on average, one such applicant each period—no more, no less. These results come from using a window of 100 applicants over which to define our quota, but the nature of this challenge is general.

9. ADDITIONAL RESULTS: TIME DYNAMICS AND LEARNING

Our main results show that our UCB algorithm increases the hiring yield of selected applicants, while also increasing demographic diversity. A key question relates to how these patterns

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

evolve over time: are gains in diversity transient, and does exploration generate greater losses in efficiency in the short run? In this section, we explore how the hiring yield and demographics associated with selected candidates evolve over time.

9.1. *Time dynamics in analysis data*

Figure 9 shows how the quality (Figure 9A) and race/ethnicity (Figure 9B) of selected applicants evolve over our analysis sample. In Figure 9A, we compute the expected quality of ML-selected applicants using the inverse propensity weighting approach discussed in Section 4.2. Our estimates are in general somewhat noisy and we are unable to observe any statistically significant differences in estimated hiring yield between applicants selected by the SL and UCB models, though both ML models select applicants who are more likely to be hired than those selected by human recruiters. However, taking the point estimates seriously, the quality of the SL model appears to decline over time, while the quality of UCB choices is more stable. At the end of our sample, the quality of applicants selected by both models is essentially identical.

In Figure 9B, we find no discernible downward trend in the proportion of Black and Hispanic candidates selected by our UCB model over time. This suggests that, in our sample, hiring outcomes for minority applicants are high enough that our models do not update downward upon selecting them. As discussed in Section 5.2, one may be concerned that the stability of our demographic results represents a failure to learn due to biases arising from sample selection. Our tests for selection on unobservables, described in Section 6.2.2 suggests that this possibility is not driving our results here.

9.2. *Learning in simulated applicant data*

We do not see evidence of either the UCB or SL model learning during our analysis sample. This may be because our analysis above is unfortunately limited by sample size and timing: our analysis period spans just under 1.5 years, and we only observe hiring outcomes among candidates interviewed during this period. Combined, this gives us limited opportunities to observe how our models may evolve over longer periods, or respond to more substantial changes in applicant quality.

To further explore how our UCB and SL models behave, we conduct simulations in which we change the quality of applicants who enter our analysis sample, starting in 2018. We provide details of how we implement this in Appendix G.2 but, essentially, we imagine that the quality of one demographic group begins to increase during the analysis period so that, by the end of the period, all applicants from that group having a hiring yield of 1. In the meantime, we hold the quality of applicants from all other groups constant at their true 2018 mean. Given this stark set up, an efficient model is one that can detect this change in applicant quality and begin interviewing only applicants from that group. To evaluate this, we consider how each of our ML scoring approaches would evaluate the *same* cohort of candidates at different points in time. Specifically, we take the actual set of candidates who applied between January 2019 and April 2019 (hereafter, the "evaluation cohort"), and estimate their ML model scores at different points in 2018. This allows us to isolate changes in the algorithm's scores that arise from differences in learning and exploration over time, rather than from differences in the applicant pool.

For intuition, consider the scores of candidates on 1 January 2018, the first day of the analysis period. In this case, both the SL and UCB algorithms would have the same beliefs about the hiring potential of candidates in the evaluation cohort, because they share the same estimate of $E[Y_{it}|X'_{it}; D_0]$ trained on the initial data D_0. The UCB model, however, may have a different

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

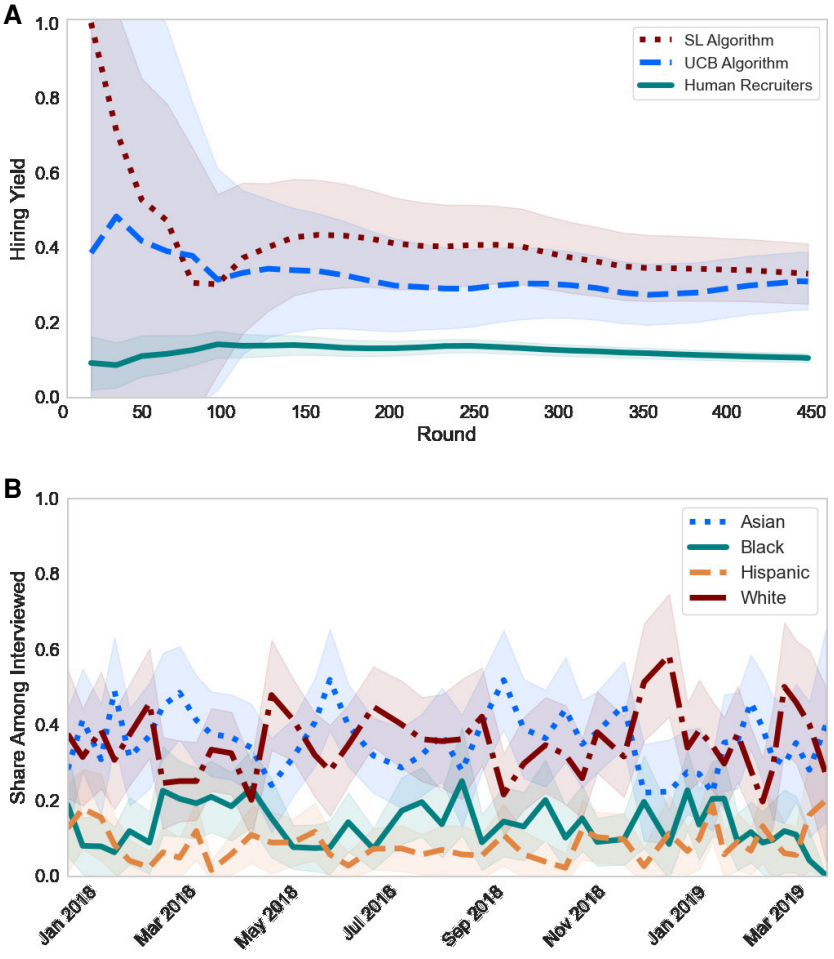


FIGURE 9
Hiring yield over time A. Hiring Yield (IPW estimates) and B. Race/Ethnicity

Notes: Panel A shows the average hiring yield of each interview screening method calculated cumulatively over time. The "SL Algorithm" line plots $E[H|I^{\text{SL}} = 1]$, the "UCB Algorithm" line plots $E[H|I^{\text{UCB}} = 1]$ and the "Human Recruiters" line plots $E[H|I^{H} = 1]$. Each estimate of quality is calculated using the inverse propensity weighting described in Section 4.2.2. Panel B shows the composition of applicants selected to be interviewed by the UCB model at each point during the analysis period. We plot a rolling average across a ten-round window and the 95% confidence intervals around each mean.

Source: All data come from the firm's application and hiring records.

score, because it also factors in its exploration bonus. On 31 December 2018, however, the SL and UCB algorithms would have both different beliefs (based on their potentially different history of selected applicants) and different scores (because the UCB factors in its exploration bonus in addition to expectations of quality). To better understand how the UCB model differs from the SL, we also consider a third variant, which tracks who the UCB model would have selected based on its estimates of $E[Y_{it}|X'_{it}; D_t^{\text{UCB}}]$ alone; this model allows us to track the evolution of the UCB model's beliefs separately from its exploration behaviour.

Figure 10 displays the results of this exercise for the simulation in which we increase the quality of Black applicants. In Appendix G.2, we discuss the results of simulations in which we increase the quality of other demographic groups, as well as the analogous simulations in which

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

we decrease the quality of applicants by demographic group. Figure 10A focuses on the share of Black applicants who are selected. We report the results of three different selection criteria. The blue dashed line reports the selection decisions of the UCB model. The UCB model rapidly increases the share of Black candidates it selects. To better understand why this happens, we plot a green dash-dot-dot line, which tracks the UCB model's *beliefs*: that is, the share of Black applicants it would select if its decisions were driven by the $\hat{E}[Y_{it}|X'_{it}; D_t^{\text{UCB}}]$ component of equation (5.4) only, leaving out the exploration bonus. Initially, the blue dashed line is above the green dash-dot-dot line; this means that the UCB model begins by selecting more Black applicants not because it necessarily believes that they have strong hiring potential, but because it is looking to explore. Over time, the green dash-dot-dot line increases as the models see more successful Black candidates and positively updates it's beliefs. Eventually, the two lines cross: at this point, the UCB model has strong positive beliefs about the hiring potential of Black applicants, but it holds back from selecting more Black candidates because it would still like to explore the quality of other candidates. By the end of the simulation period, however, exploration bonuses have declined enough so that the UCB model's decisions are driven by its beliefs, and it selects almost exclusively Black candidates.

The solid blue line shows this same process using the SL model. While it is eventually able to learn about the simulated increase in the hiring prospects of Black applicants, it does so at a significantly slower rate relative to UCB. Because supervised learning algorithms focus on maximizing current predicted hiring rates, the 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 now higher quality. This is unsurprising considering Figure 1, which shows that SL models are very unlikely to select Black applicants.

Figure 10B plots the analogous change in the quality of selected applicants over time. While the SL model eventually catches up in terms of quality, we see that the UCB model outperforms earlier because it is able to more quickly identify the group with improved quality. In Figure A.XX of Appendix G.2, we show that the gap in performance between the UCB and SL models is highest in simulations where the group whose quality is improving is less likely to be selected at baseline. This is because the UCB model proactively looks for applicants with rare covariates.

In Appendix G.2, we also discuss simulations in which applicant quality decreases. We find that the UCB model will drastically reduce the number of minority applicants it selects once it begins to learn that their quality has fallen. This differs from a quota-based system that sets minimum levels of representation.

10. CONCLUSION

This article advances our understanding of how algorithmic design affects access to job opportunities. While previous work has highlighted potential gains from following algorithmic recommendations, we highlight how algorithm design can shape the impact of these decision tools. We show that exploration-based algorithms can help firms more effectively identify candidates that meet their hiring criteria while simultaneously increasing the representation of minority applicants. This occurs even though our algorithm is not explicitly charged with increasing diversity, and even when it is blinded to demographic inputs.

Our findings shed light on the relationship between efficiency and equity in hiring. In our data, supervised learning algorithms increase hiring yield but decrease diversity, relative to the firm's current practices. A natural interpretation of this finding is that algorithms and human recruiters make different tradeoffs at the Pareto frontier, with humans prioritizing equity over efficiency. Our UCB results, however, show that such explanations may be misleading. By

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

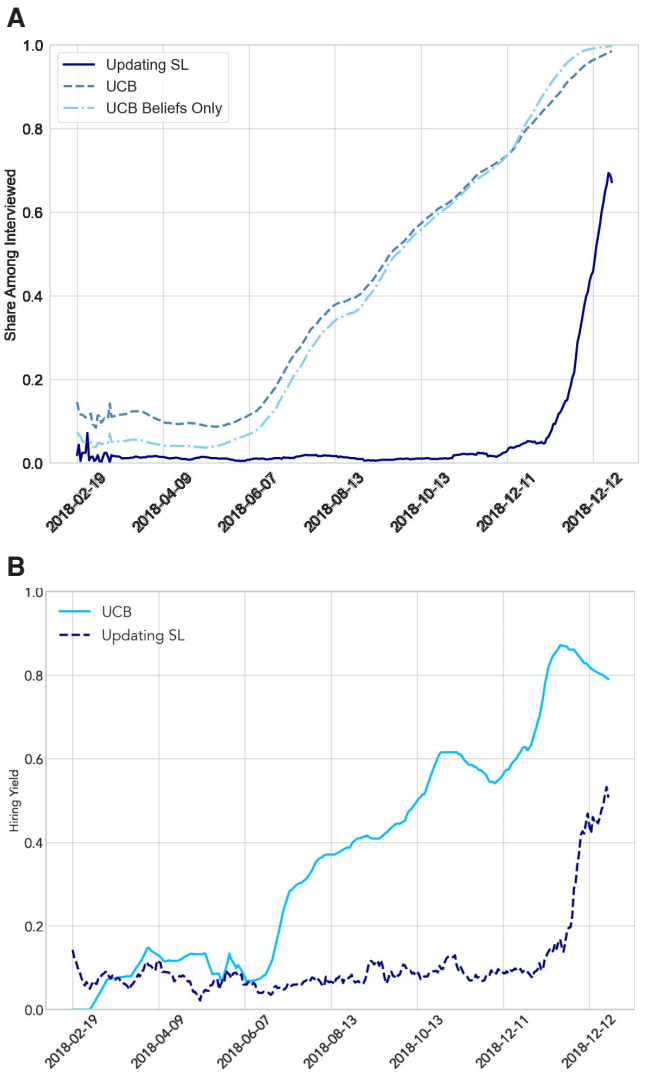


FIGURE 10
Simulations increasing the quality of black applicants. (A) Share of Selected Applicants who are Black. (B) Average Hiring Yield

Notes: This figure shows the results of a simulation in which the quality of Black applicants increases over time, as described in Section 9.2 and Appendix G.2. In (A), the *y*-axis graphs the share of Black applicants in the evaluation cohort who are selected to be interviewed by each ML policy. (B) The overall quality of interview decisions, as measured by hiring yield among interviewed applicants is plotted.

Source: All data come from the firm's application and hiring records.

demonstrating that an algorithmic approach can improve hiring outcomes while expanding representation, we provide evidence that human recruiters operate inside the Pareto frontier: in seeking diversity, they select weaker candidates over stronger ones from the same demographic groups. This leaves room to design more data-driven approaches that better identify strong candidates from under-represented backgrounds.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

Finally, our findings raise important directions for future research. As firms increasingly adopt algorithmic screening tools, it becomes crucial to understand the organizational and general equilibrium effects of such changes in HR practice. For example, there is considerable debate about the impact of diversity on team performance and how changes in the types of employees may impact organizational dynamics.[36] Such changes may also impact the validity of the predictive relation between applicant covariates and outcomes. In addition, when adopted by a single firm, an exploration-focused algorithm may identify strong candidates who are overlooked by other firms using more traditional screening techniques; yet if all firms adopt similar exploration-based algorithms, the ability to hire such workers may be blunted by supply-side constraints or competition from other firms. These equilibrium effects may reduce the potential benefits of algorithmic selection. While there is limited empirical evidence on the equilibrium effects of algorithms, Raymond (2023) shows that the adoption of algorithmic prediction impacts equilibrium prices and investment in the housing market. Such shifts in the aggregate demand for skills may also have long run impacts on the supply of skills in the applicant pool and on the returns to those skills. Both the magnitude and direction of these potentially conflicting effects deserve future scrutiny.

Acknowledgments. We gratefully acknowledge the editor, Katrine Løken, and anonymous referees for many valuable insights and suggestions. We thank David Autor, Pierre Azoulay, Dan Bjorkegren, Emma Brunskill, Max Cytrynbaum, Eleanor Dillon, Alex Frankel, Bob Gibbons, Nathan Hendren, Max Kasy, Pat Kline, Jin Li, Fiona Murray, Anja Sautmann, Scott Stern, John Van Reenen, Kathryn Shaw, and various seminar participants for helpful comments and suggestions. The content is solely the responsibility of the authors and does not necessarily represent the official views of MIT, the University of Texas at Austin, or the NBER.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability

The data and code underlying this article are available on Zenodo at https://doi.org/10.5281/zenodo.14003438.

REFERENCES

AGAN, A. and STARR, S. (2018), "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment", *The Quarterly Journal of Economics*, **133**, 191–235.
AGRAWAL, S. and GOYAL, N. (2013), "Further Optimal Regret Bounds for Thompson Sampling", in Carvalho, C. M. and Ravikumar, P. (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, Volume 31 of *Proceedings of Machine Learning Research*, Scottsdale, Arizona, USA, 99–107.
AJUNWA, I. (2019), "The Paradox of Automation as Anti-Bias Intervention", *Cardozo Law Review*, **41**, 1671.
ALSTON, M. (2019), "The (Perceived) Cost of Being Female: An Experimental Investigation of Strategic Responses to Discrimination" (Working Paper).
Anonymous Firm (2019), "Personnel Records".
ARNOLD, D., DOBBIE, W. and HULL, P. (2021), "Measuring Racial Discrimination in Algorithms", *AEA Papers and Proceedings*, **111**, 49–54.
ÅSLUND, O. and SKANS, O. N. (2012), "Do Anonymous Job Application Procedures Level the Playing Field?" *Industrial & Labor Relations Review*, **65**, 82–107.
ATHEY, S., AVERY, C. and ZEMSKY, P. (2000), "Mentoring and Diversity", *American Economic Review*, **90**, 765–786.
AUER, P. (2002), "Using Confidence Bounds for Exploitation-Exploration Trade-Offs", *Journal of Machine Learning Research*, **3**, 397–422.
BAGUES, M. and ROTH, C. (2023), "Interregional Contact and the Formation of a Shared Identity", *American Economic Journal: Economic Policy*, **15**, 322–350.

36. For instance, see Reagans and Zuckerman (2001) for a discussion of the role of diversity, and, for instance, Athey *et al.* (2000) and Fernandez *et al.* (2000) for a discussion of how changes in firm composition can shift mentoring, promotion, and future hiring patterns.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

BAGUES, M. F. and ESTEVE-VOLART, B. (2010), "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment", *Review of Economic Studies*, **77**, 1301–1328.
BAKALAR, C., BARRETO, R., BERGMAN, S., *et al.* (2021), "Fairness on the Ground: Applying Algorithmic Fairness Approaches to Production Systems".
BAROCAS, S. and SELBST, A. D. (2016), "Big Data's Disparate Impact", *California Law Review*, **104**, 671–732.
BARTOS, V., BAUER, M., CHYTILOVA, J., *et al.* (2016), "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition", *American Economic Review*, **106**, 1437–1475.
BASTANI, H. and BAYATI, M. (2020), "Online Decision Making with High-Dimensional Covariates", *Operations Research*, **68**, 276–294.
BASTANI, H., BAYATI, M. and KHOSRAVI, K. (2021), "Mostly Exploration-Free Algorithms for Contextual Bandits", *Management Science*, **67**, 1329–1349.
BECHAVOD, Y., LIGETT, K., ROTH, A., *et al.* (2019), *Equal Opportunity in Online Classification with Partial Feedback* (Red Hook, NY: Curran Associates Inc.).
BEHAGHEL, L., CRÉPON, B. and LE BARBANCHON, T. (2015), "Unintended Effects of Anonymous Resumes", *American Economic Journal: Applied Economics*, **7**, 1–27.
BENJAMIN, D. J. (2019), "Chapter 2 - Errors in Probabilistic Reasoning and Judgment Biases", in Bernheim, B. D., DellaVigna, S. and Laibson, D. (eds) *Handbook of Behavioral Economics - Foundations and Applications 2*, Volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1* (Amsterdam: North-Holland) 69–186.
BENSON, A., LI, D. and SHUE, K. (2021), "Potential and the Gender Promotion Gap".
—— —— —— (2019), "Promotions and the Peter Principle", *The Quarterly Journal of Economics*, **134**, 2085–2134.
BERTRAND, M., CHUGH, D. and MULLAINATHAN, S. (2005), "Implicit Discrimination", *American Economic Review*, **95**, 94–98.
BERTRAND, M. and DUFLO, E. (2017), "Field Experiments on Discrimination", in Banerjee, A. V. and Duflo, E. (eds) *Handbook of Field Experiments*, Volume 1 of *Handbook of Economic Field Experiments* (Amsterdam: North-Holland) 309–393.
BLAU, F. D. and KAHN, L. M. (2017), "The Gender Wage Gap: Extent, Trends, and Explanations", *Journal of Economic Literature*, **55**, 789–865.
BLS (2019), "Industries with the Largest Wage and Salary Employment Growth and Declines".
BOGEN, M. and RIEKE, A. (2018), "Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias" (Upturn).
BOHREN, J. A., HAGGAG, K., IMAS, A., *et al.* (2025), "Inaccurate Statistical Discrimination: An Identification Problem", *The Review of Economics and Statistics*, **107**, 605–620. https://doi.org/10.1162/rest_a_01367
BOHREN, J. A., IMAS, A. and ROSENBERG, M. (2019), "The Dynamics of Discrimination: Theory and Evidence", *American Economic Review*, **109**, 3395–3436.
CASTILLA, E. (2011), "Bringing Managers Back in", *American Sociological Review*, **76**, 667–694.
CASTILLA, E. J. (2008), "Gender, Race, and Meritocracy in Organizational Careers", *American Journal of Sociology*, **113**, 1479–1526.
CORBETT-DAVIES, S., GAEBLER, J. D., NILFOROSHAN, H., *et al.* (2023), "The Measure and Mismeasure of Fairness", *Journal of Machine Learning Research: JMLR*, **24**, 14730–14846.
COWGILL, B. (2020), "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening" (Working Paper, Columbia Business School).
COWGILL, B. and TUCKER, C. E. (2019), "Economics Fairness and Algorithmic Bias" (Working Paper National Bureau of Economic Research).
CRAIGIE, T.-A. (2020), "Ban the Box, Convictions, and Public Employment", *Economic Inquiry*, **58**, 425–445.
DEMING, D. J. (2017), "The Growing Importance of Social Skills in the Labor Market", *Quarterly Journal of Economics*, **132**, 1593–1640.
DIMAKOPOULOU, M., ZHOU, Z., ATHEY, S., *et al.* (2018), 'Estimation considerations in contextual bandits', arXiv, arXiv:1711.07077, https://doi.org/10.48550/arXiv.1711.07077, preprint: not peer reviewed.
——— ——— ———, *et al.* (2019), "Balanced Linear Contextual Bandits", in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. (AAAI Press).
DOLEAC, J. L. and HANSEN, B. (2020), "The Unintended Consequences of "Ban the Box": Statistical Discrimination and Employment Outcomes When Criminal Histories are Hidden", *Journal of Labor Economics*, **38**, 321–374.
EAGLY, A. H. and KARAU, S. J. (2002), "Role Congruity Theory of Prejudice Toward Female Leaders", *Psychological Review*, **109**, 573–598.
FERNANDEZ, R. M., CASTILLA, E. J. and MOORE, P. (2000), "Social Capital at Work: Networks and Employment at a Phone Center", *American Journal of Sociology*, **105**, 1288–1356.
FISCHER, C., KUCHENBÄCKER, K., ENGEL, C., *et al.* (2013), "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*, **50**, 360–367.
FISCHHOFF, B., SLOVIC, P. and LICHTENSTEIN, S. (1977), "Knowing with Certainty: The Appropriateness of Extreme Confidence", *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 1124–1131.
FRIEDMAN, S. and LAURISON, D. (2019), *The Class Ceiling: Why It Pays to Be Privileged* (Chicago, IL: University of Chicago Press).

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

FULLER, J. B., RAMAN, M., SAGE-GAVIN, E., *et al.* (2021), "Hidden Workers: Untapped Talent" (Working Paper, Harvard Business School).
GABAIX, X. (2019), "Chapter 4 - Behavioral Inattention", in Bernheim, B. D., DellaVigna, S. and Laibson, D. (eds) *Handbook of Behavioral Economics - Foundations and Applications 2*, Volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1* (Amsterdam: North-Holland) 261–343.
GOLDIN, C. and ROUSE, C. (2000), "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians", *American Economic Review*, **90**, 715–741.
HASTIE, R. and DAWES, R. M. (2001), *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making* (Los Angeles, CA: Sage Publications).
HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica: Journal of the Econometric Society*, **71**, 1161–1189.
HIRSHLEIFER, D., LEVI, Y., LOURIE, B., *et al.* (2019), "Decision Fatigue and Heuristic Analyst Forecasts", *Journal of Financial Economics*, **133**, 83–98.
HOFFMAN, M., KAHN, L. B. and LI, D. (2017), "Discretion in Hiring", *The Quarterly Journal of Economics*, **133**, 765–800.
HOUSMAN, M. and MINOR, D. (2015), "Toxic Workers" (Working Paper 16-057, Harvard Business School).
JACKSON, S. R. (2023), "(Not) Paying for Diversity: Repugnant Market Concerns Associated with Transactional Approaches to Diversity Recruitment", *Administrative Science Quarterly*, **68**, 824–866.
KASY, M. and ABEBE, R. (2020), "Fairness, Equality, and Power in Algorithmic Decision Making".
KAUSEL, E. E., CULBERTSON, S. S. and MADRID, H. P. (2016), "Overconfidence in Personnel Selection: When and Why Unstructured Interview Information Can Hurt Hiring Decisions", *Organizational Behavior and Human Decision Processes*, **137**, 27–44.
KLEINBERG, J., LAKKARAJU, H., LESKOVEC, J., *et al.* (2018), "Human Decisions and Machine Predictions", *The Quarterly Journal of Economics*, **133**, 237–293.
KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S., *et al.* (2019), "Discrimination in the Age of Algorithms", *Journal of Legal Analysis*, **10**, 113–174.
KLEINBERG, J., MULLAINATHAN, S. and RAGHAVAN, M. (2017), "Inherent Trade-Offs in the Fair Determination of Risk Scores", in Papadimitriou, C. H. (ed) *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)* (Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik) 43:1–43:23.
KLINE, P., ROSE, E. K. and WALTERS, C. R. (2022), "Systemic Discrimination Among Large U.S. Employers*", *The Quarterly Journal of Economics*, **137**, 1963–2036. _eprint: https://academic.oup.com/qje/article-pdf/137/4/1963/51053979/qjac024.pdf.
KLING, J. R. (2006), "Incarceration Length, Employment, and Earnings", *American Economic Review*, **96**, 863–876.
KOLEV, J., FUENTES-MEDEL, Y. and MURRAY, F. (2019), "Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation" (Working Paper 25759, National Bureau of Economic Research).
KUHN, P. and YU, L. (2021), "How Costly is Turnover? Evidence from Retail", *Journal of Labor Economics*, **39**, 461–496.
LAI, T. L. and ROBBINS, H. (1985), "Asymptotically Efficient Adaptive Allocation Rules", *Advances in Applied Mathematics*, **6**, 4–22.
LAKKARAJU, H., KLEINBERG, J., LESKOVEC, J., *et al.* (2017), "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* 275–284.
LAMBRECHT, A. and TUCKER, C. (2019), "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of Stem Career ADS", *Management Science*, **65**, 2966–2981.
LEI, H., LU, Y., TEWARI, A., *et al.* (2022), "An Actor-Critic Contextual Bandit Algorithm for Personalized Mobile Health Interventions".
LEPAGE, L.-P. (2020a), "Endogenous Learning, Persistent Employer Biases, and Discrimination" (Mimeo, University of Michigan).
—— (2020b), "Experimental Evidence on Endogenous Belief Formation in Hiring and Discrimination" (Mimeo, University of Michigan).
LERNER, A. (2024), "Resumes Suck" Here's the Data.
LERNER, A. and BERGMAN, P. (2024), "Are Recruiters Better than a Coin Flip at Judging Resumes? Here's the Data".
LI, L., CHU, W., LANGFORD, J., *et al.* (2010), "A Contextual-Bandit Approach to Personalized News Article Recommendation", in *Proceedings of the 19th international conference on World wide web - WWW '10* (Vol. 661). arXiv: 1003.0146.
LI, L., LU, Y. and ZHOU, D. (2017), "Provably Optimal Algorithms for Generalized Linear Contextual Bandits", in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2071–2080. JMLR.org.
MCKINNEY, S. M., SIENIEK, M., GODBOLE, V., *et al.* (2020), "International Evaluation of an AI System for Breast Cancer Screening", *Nature*, **577**, 89–94.
Mercer (2020), "Global Talent Trends" *Mercer*.
MILLER, C. (2017), "The Persistent Effect of Temporary Affirmative Action", *American Economic Journal: Applied Economics*, **9**, 152–190.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026

MÖBIUS, M. M., NIEDERLE, M., NIEHAUS, P., *et al.* (2022), "Managing Self-Confidence: Theory and Experimental Evidence", *Management Science*, **68**, 7793–7817.
MULLAINATHAN, S. and OBERMEYER, Z. (2021), "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care*", *The Quarterly Journal of Economics*, **137**, 679–727.
NAWRAT, A. (2023), "Inside HireVue's Acquisition of Modern Hire".
OBERMEYER, Z., POWERS, B., VOGELI, C., *et al.* (2019), "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations", *Science*, **366**, 447–453.
PAGER, D. and SHEPHERD, H. (2008), "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets", *Annual Review of Sociology*, **34**, 181–209.
PALUCK, E. L. and GREEN, D. P. (2009), "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice", *Annual Review of Psychology*, **60**, 339–367.
Pew Research Center (2018), *Women and Men in STEM Often at Odds Over Workplace Equity*. Pew Research Center.
PRENDERGAST, C. and TOPEL, R. (1993), "Discretion and Bias in Performance Evaluation", *European Economic Review*, **37**, 355–365.
QUADLIN, N. (2018), "The Mark of a Woman's Record: Gender and Academic Performance in Hiring", *American Sociological Review*, **83**, 331–360.
RAGHAVAN, M., BAROCAS, S., KLEINBERG, J., *et al.* (2020), "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices", in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (New York, NY, USA: Association for Computing Machinery) 469–481.
RAMBACHAN, A. and ROTH, J. (2020), "Bias In, Bias Out? Evaluating the Folk Wisdom", in Roth, A. (ed) *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, Volume 156 of *Leibniz International Proceedings in Informatics (LIPIcs)* (Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik) 6:1–6:15.
RAO, G. (2019), "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools", *American Economic Review*, **109**, 774–809.
RAYMOND, L. (2023), "The Market Effects of Algorithms" (Working Paper).
REAGANS, R. and ZUCKERMAN, E. W. (2001), "Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams", *Organization Science*, **12**, 502–517.
REJWAN, I. and MANSOUR, Y. (2020), "Top-k Combinatorial Bandits with Full-Bandit Feedback", in Kontorovich, A. and Neu G. (eds) *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, Volume 117 of *Proceedings of Machine Learning Research* (PMLR) 752–776.
RIGOLLET, P. and ZEEVI, A. (2010), "Nonparametric Bandits with Covariates".
RIVERA, L. A. (2012), "Hiring as Cultural Matching: The Case of Elite Professional Service Firms", *American Sociological Review*, **77**, 999–1022.
ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data", *Journal of the American Statistical Association*, **90**, 106–121.
RUSSAKOVSKY, O., DENG, J., SU, H., *et al.* (2015), "Imagenet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, **115**, 211–252.
RUSSO, D. and ROY, B. V. (2016), "An Information-Theoretic Analysis of Thompson Sampling", *Journal of Machine Learning Research*, **17**, 1–30.
SCHRITTWIESER, J., ANTONOGLOU, I., HUBERT, T., *et al.* (2020), "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model", *Nature*, **588**, 604–609.
SCHUMANN, C., LANG, Z., FOSTER, J., *et al.* (2019), "Making the Cut: A Bandit-Based Approach to Tiered Interviewing", in Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. D., Fox, E. and Garnett, R. (eds) *Advances in Neural Information Processing Systems* (Vol. 32). (Curran Associates, Inc).
SI, N., ZHANG, F., ZHOU, Z., *et al.* (2023), "Distributionally Robust Batch Contextual Bandits", *Management Science*, **69**, 5772–5793.
SLIVKINS, A. (2014), "Contextual Bandits with Similarity Information", *Journal of Machine Learning Research*, **15**, 2533–2568.
SVENSON, O. (1981), "Are We All Less Risky and More Skillful Than Our Fellow Drivers?" *Acta Psychologica*, **47**, 143–148.
TODD, S. (2019), "People are Terrible Judges of Talent" Can Algorithms do Better? *Quartz*.
TREISMAN, A. M. and GELADE, G. (1980), "A Feature-Integration Theory of Attention", *Cognitive Psychology*, **12**, 97–136.
United States Court of Appeals for the First Circuit (2020), "Students for Fair Admissions V. President and Fellows of Harvard College. Cert. Granted, 142 S. Ct. 895 (2022)".
WANG, L., ROTNITZKY, A. and LIN, X. (2010), "Nonparametric Regression With Missing Outcomes Using Weighted Kernel Estimating Equations", *Journal of the American Statistical Association*, **105**, 1135–1146.
WHATLEY, W. C. (1990), "Getting a Foot in the Door: "Learning," State Dependence, and the Racial Integration of Firms", *The Journal of Economic History*, **50**, 43–66.
YALA, A., LEHMAN, C., SCHUSTER, T., *et al.* (2019), "A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction", *Radiology*, **292**, 60–66.
YU, M. and KUNCEL, N. R. (2020), "Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments", *Personnel Assessment and Decisions*, **6**, 2.

Downloaded from academic.oup.com/restud/article/93/2/1200/8160842 by MIT Lincoln Laboratory user on 09 November 2026