Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening

Bo Cowgill*

Columbia University

March 21, 2020

Abstract

Where should better learning technology improve decisions? I develop a formal model of decision-making in which better learning technology is complementary with experimentation. Noisy, inconsistent decision-making by humans introduces quasi-experimental variation into training datasets, which complements learning. The model makes heterogeneous predictions about when machine learning algorithms can improve human biases. These algorithms will can remove human biases exhibited in historical training data, but only if the human training decisions are sufficiently noisy; otherwise the algorithms will codify or exacerbate existing biases. I then test these predictions in a field experiment hiring workers for white-collar jobs. The introduction of machine learning technology yields candidates that are a) +14% more likely to pass interviews and receive a job offer, b) +18% more likely to accept job offers when extended, and c) 0.2σ-0.4σ more productive once hired as employees. They are also 12% less likely to show evidence of competing job offers during salary negotiations. These results were driven by candidates who were evaluated in a noisy, biased way in historical data used for training. These candidates are broadly non-traditional, particularly candidates who graduated from non-elite colleges, who lack job referrals, who lack prior experience, whose credentials are atypical and who have strong non-cognitive soft-skills.

*The author thanks seminar participants at NBER Summer Institute (Labor), NBER Economics of Digitization, the Institutions and Innovation Conference at Harvard Business School, the Tinbergen Institute, Universidad Carlos III de Madrid, the Kauffman Emerging Scholars Conference, Columbia Business School, the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford, the Summer Institute for Competitive Strategy at Berkeley, the Wharton People and Organizations Conference, the NYU/Stern Creativity and Innovation Seminar and the University of Chicago Advances in Field Experiments Conference, as well as Jason Abaluck, Ajay Agrawal, Susan Athey, Thomas Barrios, Wouter Diesen, Laura Gee, Dan Gross, Linnea Gandhi, John Horton, Daniel Kahneman, Danielle Li, Christos Makridis, Stephan Meier, John Morgan, Harikesh Nair, Paul Oyer, Olivier Sibony, Tim Simcoe and Jann Spiess.

1 Introduction

Where should better learning technology improve decisions? In many theoretical and empirical settings, better use of empirical data improves productivity and reduces bias. However, predictive algorithms are also at the center of growing controversy about algorithmic bias. Scholars concerned about algorithmic bias have pointed to a number of troubling examples in which algorithms trained using historical data appear to codify and amplify historical bias. Examples appear in judicial decision-making (Angwin et al., 2016), to hiring (Datta et al., 2015; Lambrecht and Tucker, 2016) to targeted advertising (Sweeney, 2013). Policymakers ranging from German chancellor Angela Merkel[1] to the US Equal Employment Opportunity Commission[2] have reacted with public statements and policy guidance. The European Union has adopted sweeping regulations targeting algorithmic bias.[3]

Counterfactual comparisons between algorithms and other decision-making methods are rare. Where they exist, machine judgement often appears to less biased than human judgement, even when trained on historical data (Kleinberg et al., 2017; this paper). How can algorithms trained on biased historical data ultimately decrease bias, rather than prolong it? Where in the economy will machine learning and data create better decision-making and resulting productivity benefits?

In this paper, I model the relationship between learning technology and decision-making. The key feature of the model is that learning technology and experimentation are complementary. However, even if human decision-makers refuse to experiment and are biased towards certain behaviors, their pattern of choices can nonetheless provide useful exploration. The model endogenizes algorithmic bias, showing that its magnitude depends on the noisiness of the underlying data-generating process. The behavioral noisiness of human decision-making effectively substitutes for deliberate exploration, and provides the random variation that is complementary with learning technology.

The model shows the relationship between the magnitudes of noise and bias. As a bias becomes increasingly large, a progressively smaller amount of noise is needed for de-biasing. The model suggest that tasks and sectors featuring noisy, biased human decision-makers are most ripe for productivity enhancements from machine learning. With sufficient noise, superior learning technology can overcome not only taste-based biases against certain choices, but also biases in how outcomes are graded. However, the requirements for completely eliminating bias are extreme. A more plausible scenario is that algorithms using this approach will reduce, rather than fully eliminate, bias.

I then test the predictions of this model in a field experiment in hiring for full-time, white collar office jobs (software engineers). Before the experiment, a large company trained an algorithm to predict which candidates would pass its interviews. In the experiment, this algorithm randomly overrides the choices of experienced human screeners (the status quo at the firm) in deciding who

[1] In October 2016, German chancellor Angela Merkel told an audience that "Algorithms, when they are not transparent, can lead to a distortion of our perception." https://www.theguardian.com/world/2016/oct/27/angela-merkel-internet-search-engines-are-distorting-our-perception

[2] In October 2016, the US EEOC held a symposium on the implications of "Big Data" for Equal Employment Opportunity law. https://www.eeoc.gov/eeoc/newsroom/release/10-13-16.cfm

[3] See the EU General Data Protection Regulation https://www.eugdpr.org/, adopted in 2016 and enforceable as of May 25, 2018.

is extended an interview.

The field experiment yields three main results. First, the machine candidates outperform human screeners on nearly all dimensions of productivity. I find that the marginal candidate picked by the machine (but not by the human) is +14% more likely to pass a double-blind face-to-face interview with incumbent workers and receive a job offer offer, compared to candidates who the machine and human both select. These "marginal machine candidates" are also +18% more likely to accept job offers when extended by the employer, and 12% less likely to show evidence of competing job offers during salary negotiations. They are 0.2σ-0.4σ more productive once hired as employees. The increased quality of hires is achieved while increasing the volume of employees hired.

Second, the algorithm increases hiring of non-traditional candidates. In addition, the productivity benefits come from these candidates. This includes women, racial minorities, candidates without a job referral, graduates from non-elite colleges, candidates with no prior work experience, candidates who did *not* work for competitors.

Lastly, I find that the machine advantage comes partly from selecting candidates with superior non-cognitive soft-skills such as leadership and cultural fit, and *not* from finding candidates with better cognitive skills. The computer's advantage appear precisely the soft dimensions of employee performance which some prior literature suggests that humans – and not machines – have innately superior judgement. Given the findings of psychologists about the noisiness and bias of assessments of cultural fit, this is also consistent with the theoretical model.

I also find three results about the mechanisms behind these effects. First, I show evidence for the key feature in the of the theoretical model: The noisiness and inconsistency of human recruiters provides exploration of non-traditional candidates' performance. The strongest effects come through candidates who were relatively *unlikely* to be selected by human recruiters, but who were also evaluated in a noisy, inconsistent way. This provided performance observations on these candidates, despite their disadvantages in the process.

Second, human underperformance is driven by poor calibration on a relatively small number of variables. Between 70% and 90% of the productivity benefit from the algorithm's can be recovered from a low-dimensional OLS approximation. However, recovering the optimal parameters of the simpler model is not straightforward. The parameters may be impossible to recover without first creating the higher-dimensional model, and then approximating it with a simpler model. The machine's advantage in processing higher number of variables than a human can may be indirectly useful. They may help the machine learn a small number of optimal weights, even if most of the variables are ultimately can have effectively zero weight in evaluations.

Third, tests of *combining* human and algorithmic judgement fare poorly for human judgement. Regressions of interview performance and job acceptance on both human and machine assessments puts nearly all weight on the machine signal. I found no sub-group of candidates for whom human judgement is more efficient. When human screeners are informed of the machine's judgment, they elect to defer to the machine.

In the final section of the paper, I compare heterogeneous treatment effects to the "weights" inside the algorithm's model. Policy entrepreneurs often seek transparency in algorithms (publishing an algorithm's code and numerical weights) as a way of evaluating their bias and impact. However, my experiment shows how this could be highly misleading. Even if an algorithm (say)

penalizes inexperienced candidates with negative weights, it might help such candidates if the weights in the counterfactual method are worse. This experiment shows that the weights are not only different magnitudes as the treatment effects – they are also often not even the same sign.

While my data comes from a stylized setting, these results show evidence of productivity gains from IT adoption (). Limiting or eliminating human discretion improves both the rate of false positives (candidates selected for interviews who fail) as well as false negatives (candidates who were denied an interview, but would have passed if selected). These benefits come exclusively through re-weighting information on the resume – not by introducing new information (such as human-designed job-tests or survey questions) or by constraining the message space for representing candidates.

Section 2 outlines a theoretical framework for comparing human and algorithmic judgment. Section 3 discusses the empirical setting and experimental design, and section 4 describes the econometric framework for evaluating the experiment. Section 5 contains results. Section 6 concludes with discussion of some reasons labor markets may reward "soft skills" even if they can be effectively automated, and the effect of integrating machine learning into production processes.

1.1 Related Literature

The model in this paper is related to the emerging fairness literature in computer science (Friedler and Wilson, eds, 2018), and particularly to the usefulness for randomness in learning. Within the fairness literature, several papers explore the application multi-armed bandidts, active- and online learning (Joseph et al., 2016; Dimakopoulou et al., 2017). These papers emphasize the benefit of deliberate, targeted exploration through randomization.

However, some settings give researchers the bandit-like benefits of random exploration for free because of noise in the environment (particularly noise in human decision-making). This may be particularly useful when multi-armed bandits aren't allowed or feasible.[4] However, the experiments arising from environmental noise (described in this paper) are inefficient and poorly targeted.

Methods from the bandit literature are far more statistically efficient because they utilize noise more effectively than human psychology's behavioral quirks. In addition, many bandit-methods eventually (asymptotically) converge to unbiasedness. However as I discuss in Proposition 4, the approach in this paper may not ever converge if the environment isn't sufficiently noisy.

This paper also builds on an early formal models of the effects of machine learning and algorithmic in decision-making, particularly in a strategic environment. This theory model is related to Mullainathan and Obermeyer (2017); Chouldechova and G'Sell (2017); Hardt et al. (2016); Kleinberg et al. (2016). In addition, Hoffman et al. (2016) contains a theoretical model of decision-making by humans and algorithms and evaluates differences.

This paper is related to Agrawal et al. (2017), which models the economic consequences of improved prediction. The paper concludes by raising "the interesting question of whether improved machine prediction can counter such biases or might possibly end up exacerbating them."

[4] Algorithms requiring deliberate randomization are sometimes viewed as taboo or unethical.

This paper aims to advance this question by characterizing and decomposing the nature of prediction improvements. Prediction improvements may come about from improvements in bias or variance, which may have differing economic effects. As Agrawal et al. (2017) allude, some changes that superficially resemble "prediction improvements" may in fact reinforce deeply held biases.

The model in this paper separately integrates prediction errors from bias and variance – and the possibility of each improving – into a single model that makes heterogeneous predictions about the effects of AI. It also develops microfoundations for how these changes arise endogenously – from the creation of training data through its use by machine learning engineers.

How would this effect the overall quality of candidates? This approach has the advantage of permitting real-world empirical verification of the models. Researchers can organize trials – field experiments and A/B tests – to test the policy by modifying screening policy. By contrast, researchers cannot easily randomly alter candidate characteristics in the real world. The idea of using model features (weights, coefficients, or derivatives of an algorithm) to measure the impact of an algorithm – which is implicit in Kusner et al. (2017) and related papers – is formally analyzed in Proposition 9 of this paper.

2 Theoretical Framework

In this section, I examine a simple labor market search problem in which an employer evaluates candidates. The employer's selections, along with outcomes for selected candidates, are codified into a dataset. This dataset is then utilized by a machine learning engineer to develop and optimize a predictive algorithm.

The model endogenizes the level of algorithmic bias. Algorithmic bias arises from sample selection problems in the engineer's training data. These sample selection problems arise endogenously from the incentives, institutions, preferences and technology of human decision-making.

The framework is motivated by hiring, and many of the modeling details are inspired by the empirical section later in the paper. However, the ideas in the model can be applied to decision-making in other settings. The goal is to characterize settings where algorithmic decision-making will improve decision-making more generally.

2.1 Setup: Human Decision-maker

A recruiter is employed to select candidates for a job test or interview. The recruiter faces discrete choice problem: Because interviewing is costly, he can select only one candidate and must choose the candidate most likely to pass. If a selected candidate passes the interview, the recruiter is paid a utility bonus of $r \geq 0$. r comes from a principal who wants to encourage the screener to find candidates who pass. If the selected candidate does not pass, the recruiter is paid zero bonus and is not able to re-interview the rejected candidate.

For ease of explanation, suppose there are two candidates represented by $\theta = 1$ and $\theta = 0$. The interviewer can see k characteristics about each type. The recruiter uses these k characteristics

to estimate probabilities p_0 and p_1 that either candidate will pass the interview if selected. We will assume that the recruiters' p estimates are accurate (later, we will relax this assumption). Job candidates not strategic players in the model and either pass (or not) randomly based on their true ps. Type 1 is more likely to pass ($p_1 > p_0$).

The recruiter's decisions exhibit bias. The recruiter a hidden taste payoff $b \geq 0$ for choosing Type 0, compelling taste-based discrimination. In addition, the recruiter also receives random net utility shocks $\eta \sim F$ for picking Type 1. The η utility shocks add random noise and inconsistency to the recruiter's judgement. They are motivated by the psychology and behavioral economics literature, showing the influence of random extraneous factors in human decision-making. For example, the noise shocks may come from exogenous factors such as weather (Schwarz and Clore, 1983; Rind, 1996; Hirshleifer and Shumway, 2003; Busse et al., 2015), sports victories (Edmans et al., 2007; Card and Dahl, 2011), stock prices (Cowgill and Zitzewitz, 2008; Engelberg and Parsons, 2016), or other sources of environmental variance that affect decision-makers' mindset or mood, but are unrelated to the focal decision.[5] At a recent NBER conference on economics and AI, (Kahneman, 2017) stated "We have too much emphasis on bias and not enough emphasis on random noise [...] most of the errors people make are better viewed as random noise [rather than bias]."

This formulation of noise – a utility function featuring a random component – is used in other models and settings, beginning as as early as (Marschak, 1959) and in more recent discrete choice research. Noise has been a feature of the contest- and tournament- literature since at least Lazear and Rosen (1981) and continuing into the present day (Corchón et al., 2018), and is mostly typically as measurement error (they can be here as well).

Suppose F is continuous, symmetric, and has continuous and infinite support. F could be a normal distribution (which may be plausible based on the central limit theorem) but can assume other shapes as well. The mean of F is zero. If there are average non-zero payoffs to the screener to picking either type, this would expressed in the bias term b.

2.2 Recruiter's Optimal Choices

Before turning to the machine learning engineer, I'll briefly characterize the recruiter's optimal choices. Given these payoffs, a risk-neutral human screener will make the "right" decision (Type 1) if $rp_1 + \eta > rp_0 + b$. In other words, the screener makes the right decision if the random utility shocks are enough to offset the taste-based bias (b) favoring Type 0. Let $\underline{\eta} = r(p_0 - p_1) + b$ be the minimum η necessary to offset the bias, given the other rewards involved. Such an η (or greater) happens with probability of $\Pr(\eta > r(p_0 - p_1) + b) = 1 - F(r(p_0 - p_1) + b)) = q$.

Because this paper is motivated by employer bias, we will restrict attention to the set of distributions F for which $q \in [0, \frac{1}{2}]$. In other words, there will be variation in how often the screener chooses the right decision, but she does not make the right decision in a majority of cases.

The probability q of picking the right candidate changes as a function of the other parameters of this model. The simple partial derivatives of q are the basis for Proposition 1 and the comparative statics of the human screener selecting Type 1.

[5] Note that these exogenous factors may alter the payoffs for picking both Type 0 and Type 1 candidates; F is the distribution of the *net* payoff for picking Type 1.

Proposition 1. *The screener's probability of picking Type 1 candidates (q) is decreasing in b, increasing in r, increasing in the quality difference in Type 1 and Type 0 ($p_1 - p_0$), and increasing in the variance of F.* Proof*: See Appendix A.1.*

Proposition 1 makes four statements that can be interpreted as follows. First, as the bias b is greater, the shock necessary to offset this bias must be larger. If F is held constant, these will be more rare.

Second, as the reward for successful decisions r increases, the human screener is equally (or more) likely to make the right decision to pick Type 1. This is because the rewards benefit from picking Type 1 will increasingly outweigh his/her taste-based bias. The ηs necessary to offset this bias are smaller and more common.

Third: Proposition 1 states that as the difference between Type 1 and Type 0 ($p_1 - p_0$) is larger, the screener is more likely to choose Type 1 despite her bias. This is because the taste-based bias against Type 1 is offset by a greater possibility of earning the reward r. The minimum η necessary for the Type 1 candidate to be hired is thus smaller and more probable.

Finally, q can be higher or lower depending on the characteristics of F, the random utility shocks function with mean of zero. For any b and r, I will refer to the *default* decision as the type the screener would choose without any noise. Given this default, F is "noisier" if increases the probability mass necessary to flip the decision from the default. This is similar to the screener "trembling" (Selten, 1975) and picking a different type than she would without noise.

Where Type 0 is the default, a *default* F will place greater probability mass above $\underline{\eta}$. This corresponds to a greater η realizations above $\underline{\eta}$ favoring Type 1 candidates. In these situations, q is increasing in the level of noise in F. For a continuous, symmetric distribution such as the normal distribution, greater variance in F places is noisier regardless of r and b, since it increases the probability of a η that flips the decision.

2.3 Setup: Machine Learning Engineer

Recruiters' utility comes entirely from r, b and η. After interviewing is completed, each candidate has an interview outcome y, a binary variable representing whether the candidate was given an offer. Candidates who have an offer feature $y = 1$, candidates who are not interviewed or don't pass have $y = 0$. After every recruiter decision, y and the k characteristics (including θ) are recorded into a *training dataset*. The choice to interview is not recorded; only those who are given an offer or not.[6]

After many rounds of recruiter decisions, training dataset is given to an algorithm developer. The developer is tasked with creating an algorithm to select candidates for interviews. Like the recruiter, the engineer is paid r for candidates who pass the interview. The developer can view θ and y for each candidate, but cannot see the values p_1, p_0, q, b or the η realizations.

The humans' η noise realizations are hidden, and this complicates the measurement of bias in the training dataset. Even if the engineer knows about the existence of the ηs (and other variables

[6] I discuss this assumption later in Section **??**.

observable to the human and not the engineer), she may not know that these variables are noise and uncorrelated with performance.

The machine learning engineers thus face a limited ability to infer information about candidate qualities. A given training dataset could plausibly be generated by a wide variety of p_1, p_0, q, b or the η realizations. This is similar to labor economists' observation that differential hiring rates are not (alone) evidence of discrimination or bias.

The engineers may have wide priors about which of these is more likely, and little clues from the data alone. They may believe a variety of stories are equally likely. Furthermore, because the ηs were not recorded, the engineers cannot exploit the ηs for econometric identification. The engineers lack the signals necessary to isolate "marginal" candidates that are the topic of economic studies of bias.

The engineers predicament in this model is realistic and should be familiar to empirical economists. Econometric identification is difficult for many topics, particularly those around bias and discrimination. Researchers perpetually search for convincing natural experiments. Like the engineers in this model, they often fail to discover them.

The engineers' flat priors is also realistic. Machine learning engineers often approach problems with an extensive prediction toolkit, but without subject-area expertise. Even if they did, improving priors may be difficult because of the identification issue above.

Although the above issues frustrate the engineers' estimation p, the engineers can clearly attempt to estimate $y \in \{0, 1\}$ (the variable representing whether the candidate was extended an offer). y is a composite variable combining both choice to be interviewed and the performance of the interview. $E[y|\theta]$ will clearly be misleading estimate of $p|\theta$. However as I describe below, many machine learning practitioners proceed to estimate y in this scenario, particularly when the issues above preclude a better strategy. The next section of results outlines when $E[y|\theta]$ will be practically useful alternative to human decision-making, even if it is a misleading estimate of p.

2.4 Modeling Choices

Why not knowing who interviewed? This paper will study an algorithm in which knowing why candidates were interviewed – or whether they were interviewed at all – is not necessary. I will assume that all the ML engineers can see is θ and an outcome variable y for each candidate. y will equal 1 if the candidate was tested and passed and equal zero otherwise.

In the next section, I analyze the equilibrium behavior for the setup above, beginning with a discussion of a few modeling choices. Although the setup may apply to many real world settings, there are a few limitations of the model worth discussing.

First, although human screeners are able to observe and react to the η realizations, they do not recognize them as noise and thus do not learn from the experimentation they induce. This assumption naturally fits settings featuring taste-based discrimination, as I modeled above. In Section 2.7.1, I discuss alternative microfoundations for the model, including statistical discrimination. From the perspective of this theory, the most important feature of the screeners' bias is that it is stubborn and is *not* self-correcting through learning. Insofar as agents are statistical discrimina-

tors, the experimentation is not deliberate and they do not learn from the exogenous variation generated by the noise.

Second, the human screeners and machine learning engineers do not strategically interact in the above model. For example, the human screeners do not attempt to avoid job displacement by feeding the algorithm deliberately sabotaged training data. This may happen if the screeners' direct immediate costs and rewards from picking candidates outweigh the possible effects of displacement costs in the future of automation (perhaps because of present bias).

In addition, there is no role for "unobservables" in this model besides noise. In other words, the only variables privately observed by the human decision-maker (and not in the training data) are noise realizations η. These noise realizations are not predictive of the candidate's underlying quality, and serve only to facilitate accidental experimentation and exploration of the candidate space. By contrast, in other models (Hoffman et al., 2016), humans are able to see predictive variables that the ML algorithm cannot, and this can be the source of comparative advantage for the humans, depending on how predictive the variable is.

For the theory in this paper to apply, the noise realizations η must be truly random – uncorrelated with other observed or unobserved variables, as well as the final productivity outcome. If these conditions are violated, the algorithm may nonetheless have a positive effect on reducing bias. However, this would have to come about through a different mechanism than outlined in the proofs below.

Lastly, this paper makes assumptions about the asymptotic properties of algorithmic predictions. In particular, I assume that the algorithm convergest to $E[Y|\theta]$, but without specifying a functional form. This is similar to Bajari et al.'s 2018 "agnostic empirical specification." The convergence property is met by a variety of prediction algorithms, including OLS. However, asymptotic properties of many machine learning algorithms are often still unknown. Wager and Athey (2017) shows that the predictions of random forests are asymptotically unbiased. I do not directly model the convergence or its speed. The paper is motivated by applications of "big data," in which sample sizes are large. However, it is possible that for some machine learning algorithms, convergence to this mean may be either slow or nonexistant, even when trained on large amounts of data.

Note that in this setup, the human labeling process is both the source of training data for machine learning, as well as the counterfactual benchmark against which the machine learning is assessed.

2.5 ML Engineer's Choices

As previously discussed in Section **??**, this paper examines a set of algorithms in which the engineer is asked to predict y (passing the test) from θ by approximating $E[y|\theta]$. For Type 0 candidates, this converges to $(1-q)p_0$. For Type 1 candidates, this convergees to qp_1.

The ML engineers then use the algorithm to pick the type with a higher $E[Y|\theta]$. It then implements this decision consistently, without any noise. I will now compare the performance of the algorithm's selected candidate to that of the human decision process.

2.6 Effects of Shift from Human Screener to Algorithm

Proposition 2. *If screeners exhibit bias but zero noise, the algorithm will perfectly codify the humans' historical bias. The algorithm's perfomance will precisely equal that of the biased screeners and exhibit high goodness-of-fit measures on historical human decision data.* Proof*: See Appendix A.2.*

Proposition 2 formalizes a notion of algorithmic bias. In the setting above – featuring biased screeners $b > 0$ with no noise – there is no difference in the decision outcomes. The candidates approved (or rejected) by the humans would face the same outcomes in the machine learning algorithm.

The intuition behind Proposition 2 is that machine learning cannot learn to improve upon the existing historical process without a source of variation and outcomes. Without a source of clean variation – exposing alternative outcomes under different choices – the algorithms will simply repeat what has happened in the past rather than improve upon it.

Because the model will perfectly replicate historical bias, it will exhibit strong goodness-of-fit measures on the training data. The problems with this algorithm will not be apparent from cross-validation, or from additional observations from the data generating process.

Thus there are no decision-making benefits to using the algorithm. However it is possible that the decision-maker receives other benefits, such as lower costs. Using an algorithm to make a decision may be cheaper than employing a human to make the same decision.

Proposition 3. *If screeners exhibit zero bias but non-zero amounts of noise, the algorithm will improve upon the performance of the screeners by removing noise. The amount of performance improvement is increasing in the amount of noise and the quality difference between Type 1 and Type 0 candidates.* Proof*: See Appendix A.3.*

Proposition 3 shows that performance improvements from the algorithm can partly come from improving consistency. Even when human decisions are not biased, noise may be a source of their poor performance. Although noise is useful in some settings for learning – which is the main theme of this paper – the noise harms performance if the decision process is already free of bias.

Proposition 4. *If biased screeners are NOT sufficiently noisy, the algorithm will codify the human bias. The reduction in noise will actually make outcomes worse.* Proof*: See Appendix A.4.*

Proposition 4 describes a setting in which screeners are biased and noisy. This generates some observations about Type 1's superior productivity, but not enough for the algorithm to correct for the bias. In the proof for Proposition 4 in Appendix A.4, I formalize the threshold level of noise below which the algorithm is biased.

Beneath this threshold, the algorithm ends up codifying the bias, similarly to Proposition 2 (which featured bias, but no noise). However, the adoption of machine learning actually worsens decisions in the setting of Proposition 4 (whereas it simply made no difference in the setting of Proposition 2). In a biased human regime, any amount of noise actually helps the right candidates gain employment.

The adoption of the machine learning removes this noise by implementing the decision consistently. Without sufficient experimentation in the underlying human process, this algorithm cannot correct the bias. The reduction in noise in this setting actually makes outcomes worse than if we trusted the biased, slightly noisy humans.

Proposition 5. *If screeners are biased and sufficiently noisy, the algorithm will reduce the human bias.* Proof*: See Appendix A.5.*

Proposition 5 shows the value of noise for debasing – one of the main results of the paper. If the level of noise is above the threshold in the previous Proposition 4, then the resulting algorithm will feature lower bias than the original screeners' data. This is because the random variation in the human process has acted as a randomized controlled trial, randomly exposing the learning algorithm to Type 1's quality, so that this productivity can be fully incorporated into the algorithm.

In this sense, experimentation and machine learning are compliments. The greater experimentation, the greater ability the machine learning to remove bias. However, this experimentation does not need to be deliberate. Random, accidental noise in decision-making is enough to induce the debiasing if the noise is a large enough influence on decision-making.

Taken together, Propositions 4 and 5 have implications for the way that expertise interacts with machine learning. A variety of research suggests that the benefit of expertise is lower noise and/or variance, and that experts are actually *more* biased than nonexperts (they are biased toward their area of expertise, Li, 2017).

If this is true, then Proposition 4 suggests that using expert-provided labels for training data in machine learning will codify bias. Furthermore, the performance improvement coming from lower noise will be small, because counterfactual expert was already consistent. Even if experts' evaluations are (on average) better than nonexperts, experts' historical data are not necessarily more useful for training machine learning (if the experts fail to explore).

Proposition 6. *If the algorithms' human data contain non-zero bias, then "algorithmic bias" cannot be reduced to zero unless the humans in the training data were perfectly noisy (i.e., picking at random).* Proof*: See Appendix A.6.*

Even if screeners are sufficiently noisy to reduce bias (as in Proposition 5), the algorithm's predictions still underestimate the advantage of Type 1 above Type 0.

In particular, the algorithm predicts a y of qp_1 for Type 1 and $(1-q)p_0$ for Type 0. The algorithm's implicit quality ratio of Type 1 over Type 0 is $qp_1/(1-q)p_0$. This is less than the quality ratio of Type 1 over Type 0 (p_1/p_0) – unless noise is maximized by increasing the variance of F until $q = 1/2$. This would make the training data perfectly representative (i.e., humans were picking workers at random). Despite the reduction in bias, the algorithm will remain handicapped and exhibit some bias because of its training on biased training data.

Picking at random is extremely unlikely to appear in any real-world setting, since the purpose of most hiring is to select workers who are better than average and thus undersample sections of the applicant pool perceived to be weaker. A complete removal of bias therefore appears infeasible from training datasets from real-world observations, particularly observations of agents who are *not* optimizing labels for *ex-post* learning.

It is possible for an algorithm to achieve a total elimination of bias without using perfectly representative training data. This may happen if a procedure manages to "guess" the a totally unbiased algorithm from some other heuristic. Some of the algorithmic innovations suggested by the emerging fairness literature may achieve this. However, in order to achieve certainty that this is algorithm is unbiased, one would need a perfectly representative training dataset (i.e., one where the screeners were picking at random).

Proposition 7. *As the amount of noise in human decisions increases, the machine learning can correct increasingly small productivity distortions. The algorithm needs only a small amount of noise to correct errors with large productivity consequences.* Proof*: See Appendix A.7.*

The proposition means that if the screeners' bias displays a large amount of bias, only a small amount of noise is necessary for the algorithm to correct the bias. Similarly if screeners display a small amount of bias, then high amounts of noise are necessary for the algorithm to correct the bias. Large amounts of noise permit debiasing for both large and smaller biases, where as small noise permits only correction of large biases only.[7] Because we want all biases corrected, lots of noise is necessary to remove both large and small biases. However, Proposition 7 suggests that even a small amount of noise is necessary to reduce the most extreme biases.

The intuition behind Proposition 7 is as follows: Suppose that screeners were highly biased against Type 1 workers; this would conceal the large productivity differences between Type 1 and Type 0 candidates. The machine learning algorithm would need to see only a few realizations – a small amount of noise – in order to reduce the bias. Because each "experiment" on Type 1 workers shows so much greater productivity, few such experiments would be necessary for the algorithm to learn the improvement. By contrast, if the bias against Type 1 is small, large amounts of noise would be necessary for the algorithm to learn its way out of it. This is because each "experiment" yields a smaller average productivity gain. As a result, the algorithm requires more observations in order to understand the gains from picking Type 1 candidates.

A recent paper by Azevedo et al. (2018) makes a similar point about A/B testing. A company whose innovation policy is focused on large productivity innovations will need only a small test of each experiment. If the experiments produce large effects, they will be detectable in small sample sizes.

Proposition 7 effectively says there are declining marginal returns to noise. Although there may be increasing *cumulative* returns to noise, the marginal returns are decreasing. As Proposition 2 states, no corrections are possible if screeners are biased and feature no noise. The very first unit of noise – moving from zero noise to positive noise – allows for correction of any large productivity distortions. As additional noise is added, the productivity improvements from machine learning become smaller.

Proposition 8. *In settings featuring bias and sufficiently high noise, the algorithm's improvement in bias will be positive and increasing in the level of noise and bias. However, metrics of goodness-of-fit on the training data (and on additional observations from the data-generating process) have an upper bound that is low compared to settings with lower noise and/or lower bias.* Proof*: See Appendix A.8.*

[7] "Large" and "small" biases are used here in a relative sense – "small" biases in this model could be very harmful on a human scale, but are labeled "small" in this model only in comparison to still even greater biases.

The proof in Appendix A.8 comprares the algorithm's goodness-of-fit metrics on the training data in the setting of Proposition 5 (where debiasing happens) to Propositions 2 and 4, which codify bias. In the setting that facilitates debiasing, goodness-of-fit measures are not only low relative to the others, but also in absolute numbers (compared to values commonly seen in practice).

The implication of Proposition 8 is: If engineers avoid settings where models exhibit poor goodness-of-fit on the training data (and future samples), they will avoid the settings where machine learning has the greatest potential to reduce bias.

Proposition 9. *The "coefficient" or "weight" the machine learning algorithm places on $\theta = 1$ when ranking candidates does not equal the treatment effect of using the algorithm rather than human discretion for $\theta = 1$ candidates.* Proof*: See Appendix A.9.*

Proposition 9 discusses how observers should interpret the coefficients and/or weights of the machine learning algorithm. It shows that these weights may be highly misleading about the impact of the algorithm. For example: It's possible for an algorithm that places negative weight on $\theta = 1$ when ranking candidates could nonetheless have a strong positive benefit for $\theta = 1$ candidates and their selection outcomes. This would happen if the human penalized these characteristics even more than the algorithm did.

The internal weights of these algorithms are completely unrelated to which candidates benefit from the algorithm compared to a status quo alternative. The latter comparison requires a comparison to a counterfactual method of selecting candidates.

2.7 Extensions

2.7.1 Other Microfoundations for Noise and Bias

In the setup above, I model bias as taste-based discrimination, and noise coming from utility shocks within the same screener over time. However, both the noise and bias in the model can arise from different microfoundations. These do not affect conclusions of the model. I show these alternative microfoundations formally in Appendix A.10.

The formulation above models the bias against Type 1 candidates as "taste-based" (Becker, 1957), meaning that screeners receive direct negative payoffs for selecting one type of worker. A taste-based discriminator may be conscious of his/her taste-based bias (as would a self-declared racist) or unconscious (as would someone who feels worse hiring a minority, but can't say why). Either way, taste-based discrimination comes directly from the utility function.

Biased outcomes can also arise from statistical discrimination (Phelps, 1972; Arrow, 1973). Screeners exhibiting statistical discrimination (and no other type of bias) experience no direct utility preferences for attributes such as gender or race. "Statistical discrimination" refers to the process of making educated guesses about an unobservable candidate characteristic, such as which applicants' perform well as employees. If applicants performance is (on average) even slightly correlated with observable characteristics such as gender or race, employers may be tempted to use these variables as imperfect proxies for unobservable abilities. If worker quality became easily observable, screeners exhibiting statistical discrimination would be indifferent between races or

genders.

The framework in this paper can be reformulated so that the bias comes from statistical discrimination. This simply requires one additional provision: That the "educated guesses" are wrong and are slow to update. Again, the psychology and behavioral economics literature provides ample examples of decision-makers having wrong, overprecise prior beliefs that are slow to update.

Similarly, the noise variable η can also have alternative microfoundations. The formulation beginning in Section 2 proposes that η represents time-varying noise shocks within a single screener (or set of screeners). However, η can also represent noise coming from between-screener variation. If a firm employs multiple screeners and randomly assigns applications to screeners, then noise can arise from idiosyncrasies in the each screener's tastes.

The judgment and decision-making literature contains many examples of this between-screener variation as a source of noise.[8] This literature uses "noise" to refer to within-screener and between-screener random variations interchangeably. Kahneman et al. (2016) simply writes, "We call the chance variability of judgments *noise*. It is an invisible tax on the bottom line of many companies."

Similarly, the empirical economics literature has often exploited this source of random variation for causal identification.[9] This includes many papers in an important empirical setting for the CS literature on algorithmic fairness: Judicial decision-making. As algorithmic risk-assessment tools have grown in popularity in U.S. courts, a series of academic papers and expose-style journalism allege these risk assessment tools are biased. However, these allegations typically do not compare the alleged bias to what a counterfactual human judge would have done without algorithmic guidance.

A series of economics papers examine the random assignment of court cases to judges. Because human judges' approaches are idiosyncratic, random assignment creates substantial noisiness in how cases are decided. These researchers have documented and exploited this noise for all kinds of analysis and inference. The randomness documented in these papers suggests that courts exhibit the noisiness I argue is the key prerequisite for debiasing human judgement through algorithms.

However, clean comparisons with nonalgorithmic judicial decision-making are rare. One paper that does this is Kleinberg et al. (2017). It utilizes random assignment in judges for evaluating a machine learning algorithm for sentencing and finds promising results on reducing demographic bias.

[8] For example, this literature has shown extensive between-screener variation in valuing stocks (Slovic, 1969), evaluating real-estate (Adair et al., 1996), sentencing criminals (Anderson et al., 1999), evaluating job performance (Taylor and Wilsted, 1974), auditing financies (Colbert, 1988), examining patents (Cockburn et al., 2002) and estimating task-completion times Grimstad and Jørgensen (2007).

[9] For example, assignment of criminal cases to judges (Kling, 2006), patents applications to patent examiners (Sampat and Williams, 2014; Farre-Mensa et al., 2017), foster care cases to foster care workers (Doyle Jr et al., 2007; Doyle Jr, 2008), disability insurance applications to examiners (Maestas et al., 2013), bankruptcy judges to individual debtors (Dobbie and Song, 2015) and corporations (Chang and Schoar, 2013) and job seekers to placement agencies (Autor and Houseman, 2010).

2.7.2 Additional Bias: How Outcomes are Codified

Until now, the model in this paper has featured selection bias in which a lower-quality candidate joins the training data because of bias. This is a realistic portrayal of many fields, where performance is accurately measured for workers in the field, but entry into the field may contain bias. For example: In jobs in finance, sales, and some manual labor industries, performance can be measured objectively and accurately for workers in these jobs. However, entry into these labor markets may feature unjust discrimination.

In other settings, bias may also appear within the training data in the way outcomes are evaluated for workers who have successfully entered. For example: Suppose that every positive outcome by a Type 1 candidate is scored at only 90% as valuable as those by Type 0. In this extension, I will evaluate the model's impact when $\theta = 1$ candidates are affected by both types of bias.

Let $\delta \in [0, 1]$ represent the discount that Type 1's victories are given in the training data. High δs represent strong bias in the way Type 1's outcomes are evaluated. If $\delta = 0.9$, then Type 1's victories are *codified* as only 10% as valuable as Type 0's even if they are equally valuable in an objective sense. This could happen if (say) the test evaluators were biased against Type 1 and subtracted points unfairly.[10]

In Appendix B, I provide microfoundations for δ and update the propositions above to incorporate both types of bias. Again, noise is useful for debasing in many settings (Appendix Proposition 15). The introduction of the second type of bias actually increases the usefulness of noise. However, the existence of the second type of bias also creates limitations. For a threshold level of δ, the algorithm under this procedure will not decrease bias and can only entrench it (Appendix Proposition 11) no matter how much noise in selection.

These conclusions assume that evaluations could be biased (δ), but these evaluations are not themselves noisy (in the same way that selection decisions were). Future research will add a parameter for noisy posthire evaluations.

2.8 Model Discussion and Conclusion

This paper contains a model of how human judges make decisions, how these decisions are codified into training data, and how this training data is incorporated into a decision-making algorithm by engineers under mild assumptions.

I show how characteristics of the underlying human decision process propagate the later codification into training data and an algorithm, under circumstances common in practice.

The key feature of the model is that improvements to the human process are made possible only through experimental variation. This experimentation need not be deliberate and can come through random noise in historical decision-making.

Although the model was motivated by hiring, it could be applied to a wide variety of other settings in which bias and noise may be a factor. For example: Many researchers wonder if machine

[10] As with the earlier bias in hiring (b), the evaluation bias here (δ) could itself be the result of tastes or statistical inferences about the underlying quality of work.

learning or AI will find natural applications decreasing behavioral economics biases (loss aversion, hindsight bias, risk-aversion, etc). This model predicts that this is a natural application, but only these biases are realized in a noisy and inconsistent way.

Similarly, one can also use this model to assess why certain machine learning applications have been successful and which ones may be next. For example: Early, successful models of computer chess utilized supervised machine learning based on historical human data. The underlying humans most likely played chess with behavioral biases, and also featured within-player and between-player sources of noise. The model suggests that the plausibly high amounts of bias and noise in human chess moves make it a natural application for supervised AI that would reduce both the inconsistency and bias in human players.

Recent work by (Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2018) attempts to classify jobs tasks in the Bureau of Labor Statistics' O*NET database for their suitability for machine learning applications. The authors create "a 21 question rubric for assessing the suitability of tasks for machine learning, particularly supervised learning." A similar paper by Frey and Osborne (2017) attempts to classify job tasks easily automated by machine learning.

In these papers, the level of noise, random variation, or experimentation in the training data is not a criteria for "suitability for machine learning." Noisiness or quasi-experimental variation is not a major component of the theoretical aspects in either paper. In Brynjolfsson and Mitchell (2017); Brynjolfsson et al. (2018), noise is a *negative* predictor of "suitability for machine learning."

Instead, both analyses appear to focus mostly on settings in which human decision-making process can easily be mimicked rather than improved upon through learning. This is consistent with the goal of maximizing goodness of fit to historical data (as characterized above) rather than reducing bias. Proposition 2 suggests that applying machine learning in low noise environments will yield mimicking, cost savings, high goodness-of-fit measures and possible entrenchment of bias – rather than better, less-biased decision-making.

If the goal is learning and improving, noisiness should be positively correlated with adoption. Future empirical work in the spirit of (Frey and Osborne, 2017; Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2018) may be able to separately characterize jobs or tasks where machine learning yields cost-reduction, mimicking benefits from those where benefits arise from learning and optimizing using experiments.

Researchers in some areas of machine learning – particularly active learning, online learning, and multi-armed bandits – embrace randomization as a tool for learning. In many settings researchers enjoy the benefits of randomization for free because of noise in the environment.

By promoting consistent decision-making, adopting algorithms may actually eliminate useful experimental variation this paper has argued is so useful. Ongoing experimentation can be particularly valuable if the data-generating environment is changing. However, the experiments arising from environmental noise are inefficient and poorly targeted. Methods from the bandit and online learning contain much more statistically efficient use of noise than human psychology's behavioral quirks.

The setup described above may not apply well to all settings. In particular, there may be settings in which variables observable to humans (but unobservable in training data) could play a larger

role as they do *not* in the model above. We may live biased world featuring lots of noise – but still not enough to use in debiasing. For some variables (such as college major or GPA) we may have enough noise to facilitate debiasing, while for others (such as race or gender), historical bias may be too entrenched and consistent for algorithms to learn their way out. In addition, there are other ways that machine learning could reduce bias besides the mechanisms in this paper.

In many settings, however, these and other assumptions of the model are realistic. The small number of empirical papers featuring clean comparisons between human and algorithmic judgment (Kleinberg et al., 2017; Cowgill, 2017; Stern et al., 2018) demonstrate reduction of bias.

One interpretation of this paper is that it makes *optimistic* predictions about the impact of machine learning on bias, even without extensive adjustments for fairness. Prior research cited throughout this paper suggests that noise and bias are abundant in human decision-making, and thus ripe for learning and debiasing through the theoretical mechanisms in the model. Proposition 7 may have particularly optimistic implications – if we are in a world with lots of bias, we need only a little bit of noise for simple machines to correct it. If we are in a world with lots of noise (as psychology researchers suggest), simple algorithms should be able to correct even small biases.

Given this, why have so many commentators raised alarms about algorithmic bias? One possible reason is the choice of benchmark. The results of this paper suggest that completely eliminating bias – a benchmark of zero-bias algorithmic perfection – may be extremely difficult to realize from naturally occurring datasets (Proposition 6). However, reducing bias of an existing noisy process may be more feasible. Clean, well-identified comparisons of human and algorithmic judgment are rare in this literature, but the few available (Kleinberg et al., 2017; Cowgill, 2017; Stern et al., 2018) suggest a reduction of bias. These results may come perhaps for the theoretical reasons motivated by this model.

The impact of algorithms compared to a counterfactual decision process may be an important component of how algorithms are evaluated for adoption and legal/social compliance. However, standard machine learning quality metrics – goodness of fit on historical outcomes – do not capture these counterfactual comparisons. This paper suggests that to maximize counterfactual impact, researchers should pick settings in which traditional goodness-of-fit measures may be lower (i.e., those featuring lots of bias and noise).

Relative comparisons are sometimes feasible only after a model has been deployed and tested. One attractive property of the model in this paper is that many of the pivotal features could plausibly be measured in advance – at the beginning of a project, before the deployment and model building – to estimate the eventual comparative effect vs a status quo. (Kahneman et al., 2016) described simple methods for "noise audits" to estimate the extent of noise in a decision process. Levels of bias could be estimated or calibrated through historical observation data, which may suggest an upper or lower bound for bias.

Eliminating bias may be difficult or impossible using "datasets of convenience." Machine learning theory should give practitioners guidance about when to expect practical, relative performance gains, based on observable inferences about the training data. This paper – which makes predictions about relative performance depending on the bias and noisiness of the training data generated by the status quo – is one attempt to do this.

3 Empirical Setting

The setting from this study is a single multinational conglomerate multiple products and services. The job openings in this paper are technical staff such as programmers, hardware engineers and software-oriented technical scientists and specialists. The jobs in question are full-time employees with benefits and typically work more than forty-hours per week. Workers in this labor market are involved in multi-person teams that design and implement technical products. The sample in this paper is only for one job opening (software engineer), and for one geographic location where the company has a presence.[11]

3.1 Status Quo Hiring Process: Professional Screeners

The application process for jobs in this market proceeds as follows. First, candidates apply to the company through a website.[12] Next, a human screener reviews the applications of the candidate. This paper includes a field experiment in replacing these decisions with an algorithm.

The next stage of screening is bilateral interviews with a subset of the firm's incumbent workers. The first interview often takes place over the phone. If this interview is successful, a series of in-person interviews are scheduled with incumbent workers, lasting about an hour. The interviews in this industry are mostly unstructured, with the interviewer deciding his or her own questions. Firms offer some guidance about interview content but don't strictly regulate the interview content (for example, by giving interviewers a script).

After the meetings, the employees who met the candidate communicate the content of the interview discussion, impressions and a recommendation. During the course of this experiment, the firm also asked interviewers to complete a survey about the candidate evaluating his or her general aptitude, cultural fit and leadership ability. With the input from this group, the employer decides to make an offer.

Next, the candidate can then negotiate terms of the offer not. Typically, employers in this market engages in negotiation only in order to respond to competing job offers. The candidate eventually accepts or rejects the offer. Those who accept the offer begin working. At any time the candidate could withdraw his application if he or she accepts a job elsewhere or declines further interest.

A few details inform the econometric specifications in this experiment. In this talent market, firms commonly desire as many qualified workers as it can recruit. Firms often do not have a quota of openings for these roles; insofar as they do they are never filled. "Talent shortage" is a common complaint by employers regarding workers with technical skills. The economic problem of the firms is to identify and select well-matched candidates, and *not* to select the best candidates for a limited set of openings. Applicants are thus not competing against each other, but against the hiring criteria.

In addition, the hiring company does not decline to pursue applications of qualified candidates on the belief that certain candidates "would never come here [even if we liked him/her]." For

[11] In this industry, candidates are typically aware of the geographic requirements upon applying.

[12] Some candidates are also recruited or solicited; the applications in this study are only the unsolicited ones.

these jobs, the employer in this paper believes it can offer reasonably competitive terms; it does not terminate applications unless a) the candidate fails some aspect of screening, or b) the candidate withdraws interest.

3.2 Selection Algorithm

3.2.1 Background

Firms offering products and consulting in HR analytics have exploded in recent years, as a result of several trends. On the supply side of applications, several factors have caused an increase in application volumes for posted jobs throughout the economy. The digitization of job applications has lowered the marginal cost of applying. In addition, the Great Recession (and subsequent recovery) caused a greater number of applicants to be looking for work. On the demand side, information technology improvements enable firms to handle the volume of online applications. Firms are motivated to adopt these algorithms in part of the volume/costs, and to address potential mistakes in human screeners' judgements.

How common is the use of algorithms for screening? The public appears to believe it is already very common. The author conducted a survey of ≈3,000 US Internet users, asking "Do you believe that most large corporations in the US use computer algorithms to sort through job applications?"[13]

About two-thirds (67.5%) answered "yes."[14] Younger and more wealthy respondents were more likely to answer affirmatively, as were those in urban and suburban areas.

A 2012 *Wall Street Journal* article[15] estimates that the proportion of large companies using resume-filtering technology as "in the high 90% range," and claims "it would be very rare to find a Fortune 500 company without [this technology]."[16] The coverage of this technology is sometimes negative. The aforementioned WSJ article suggests that someone applying for a statistician job could be rejected for using the term "numeric modeler" (rather than statistician). However, the counterfactual human decisions mostly left unstudied. Recruiters' attention is necessarily limited, and human screeners are also capable of mistakes which may be more egregious than the above example. One contribution of this paper is to use exogenous variation to observe counterfactual outcomes.

3.2.2 Selection Algorithm: Implementation Details

The technology in this paper uses standard text-mining and machine learning techniques that are common in this industry. The first step of the process is broadly described in a 2011 LifeHacker

[13] The phrasing of this question may include both "pure" algorithmic screening techniques such as the one studied in this paper, as well as "hybrid" methods, where a human designs a multiple-choice survey instrument, and responses are weighted and aggregated by formula. An example of the latter is studied in Hoffman, Kahn and Li (2016).

[14] Responses were reweighed to match the demographics of the American Community Survey. Without the reweighing, 65% answered yes.

[15] http://www.wsj.com/articles/SB10001424052970204624204577178941034941330, accessed June 16, 2016.

[16] As with the earlier survey, this may include technological applications that differen than the one in this paper.

article[17] about resume-filtering technology:[18] "[First, y]our resume is run through a parser, which removes the styling from the resume and breaks the text down into recognized words or phrases. [Second, t]he parser then sorts that content into different categories: Education, contact info, skills, and work experience."

In this setting, the predictor variables fall into four types.[19] The first set of covariates was about the candidate's education such as institutions, degrees, majors, awards and GPAs. The second set of covariates is about work experience including former employers and job titles. The third contains self-reported skill keywords that appear in the resume.

The final set of covariates were about the other keywords used in in the resume text. The keywords on the resumes were first merged together based on common linguistic stems (for example, "swimmer" and "swimming" were counted towards the "swim" stem). Then, resume covariates were created to represent how many times each stem was used on each resume.[20]

Although many of these keywords do not directly describe an educational or career accomplishment, they nonetheless have some predictive power over outcomes. For example: Resumes often use adjectives and verbs to describe the candidate's experience in ways that may indicate his or her cultural fit or leadership style. For example: Verbs such as "served" and "directed" may indicate distinct leadership styles that may fit into some companies' better than others. Such verbs would be represented in the linguistic covariates – each resume would be coded by the number of times it used "serve" and "direct" (along with any other word appearing in the training corpus). If the machine learning algorithm discovered a correlation between one of these words and outcomes, it would be kept in the model.

For each resume, there were millions of such linguistic variables. Most were excluded by the variable selection process described below. The training data for this algorithm contained historical resumes from previous four years of applications for this position. The final training data dataset contained over one million explanatory variables per job application and several hundred thousand successful (and unsuccessful) job applications.

The algorithms used in this experiment machine learning methods – in particular, LASSO (Tibshirani, 1996) and support vector machines (Vapnik, 1979; Cortes and Vapnik, 1995) – to weigh covariates in order to predict success of the historical applications for this position. Applications were coded as successful if the candidate was extended an offer. A standard set of machine learning techniques – regularization, cross-validation, separating training and test data – were used to select and weigh variables.[21] These techniques (and others) were ment to ensure that the weights were not overfit to the training data, and that the algorithm accurately predicted which candidates would succeed in new, non-training samples.

[17] http://lifehacker.com/5866630/how-can-i-make-sure-my-resume-gets-past-resume-robots-and-into-a-humans-hand

[18] Within economics, this approach to codifying text is similar to Gentzkow and Shapiro (2010)'s codification of political speech.

[19] Demographic data are generally not included in these models and neither are names.

[20] The same procedure was used for two-word phrases on the resumes.

[21] See Friedman et al. (2013) for a comprehensive overview of these techniques. Athey and Imbens (2015) has an excellent surveys for economists.

3.2.3 Economic Features of the Algorithm

A few observations about the algorithm. First: Although the algorithm in this paper is computationally complex, it is econometrically naive. The designers sought to predict who would receive an offer using historical data. As discussed in the theoretical section of this paper (2), this approach could plausibly codify historical biases directly into the model. Relatedly, the algorithm designers ignored the two-stage, selected nature of the historical screening process. In economics, these issues were raised in Heckman (1979), but the programmers in this setting did not integrate these ideas into its algorithms. Lastly, the designers were also uninterested in interpreting the model causally and chose engineering approaches that neglected this possibility.

A few other features of the algorithm are worth mentioning. The algorithm introduced no new data into the decision-making process. In theory, all of the covariates on the resume above can also be observed by human resume screeners. The human screeners could also view an extensive list of historical outcomes on candidates through the company's HR database of historical candidates, which was available to be browsed (the algorithm's training data came from this database). In a sense, any comparisons between humans and this algorithm is inherently unfair to the machine. A human can quickly consult the Internet or a friend's advice to examine an unknown' school's reputation. The algorithm was given no method to consult outside sources or bring in new information that the human couldn't.

In addition, this modeling approach imposes no constraints on the job applicant's message space. The algorithm in this paper imposed no constraints on what mix of information, persuasion, framing and presentation a candidate could use in her presentation of self. The candidate can fill the content of her resume with whatever words she chooses. The candidate's experience was unchanged by the algorithm and his/her actions were not required to be different than the status quo human process.

This differs from other studies of hiring in which job testing interventions alter the candidate's message space. For example, the job tests studied by Hoffman, Kahn and Li (2016) are multiple-choice responses to human-designed, multiple-choice survey instrument. The answers are later weighted by an algorithm. This intervention not only changes how variables are weighed, but also the message space between candidates and screeners. Part of the benefit of such tests may come from the introduction of new variables from a human organizational psychologist, rather than the re-weighting of previously known variables. In addition, the multiple-choice format vastly constrains the message space, which simplifies the algorithm's weighing.

In this experiment, the algorithm introduces no new variables. For both arms of the experiment, the input is a text document with an enormous potential message space. The curation of the message space was performed by candidates (who act independently of each other and adversarially to screening). This setup facilitates a clean comparison of human and machine judgment based on common inputs.

Changes to the message space not only affect the interpretation of results. They may could also in theory affect downstream outcomes. The experience of answering an organizational psychologist's survey questions could affect how a candidate feels about the employer, performs in interviews or views a job offer. The questions may signal a employers' type, and make certain features of employee experience salient. These considerations are off the table in this experiment because the

algorithm left the candidate experience unaltered. As discussed in Section 4, this experiment was double-blind; candidates as well as screeners were unaware of the experiment's existence, as well as which specific candidates' treatment status.

3.3 Data

In the next section, I describe experimental design and specifications. However first I describe the variables in the analysis.

Characteristics For each candidate,

Outcomes. For the analysis in this paper, I code an applicant as being interviewed if he/she passed the resume screen and was interviewed in any way (including the phone interview). I code candidates as passing the interview if they were subsequently extended a job offer.

Table 1 contains descriptive statistics and average success rates at the critical stages above. As described in the next section, the firm used a machine learning algorithm to rank candidates. Table 1 reports separate results for candidates more than 10% likely to be offered a job – the subjects of the experiment in this study – and the remainder of applicants.

Table 1 shows that the candidates above the machine's threshold are positively selected on a number of traits. They also tend to pass rounds of screening at much higher rates even without any intervention from the machine. One notable exception is the offer acceptance rate, which is lower for the candidate that the machine ranks highly. One possible explanation for this is that the algorithms' model is similar to the broader market's, and highly ranked candidates may attract competitive offers.

4 Experimental Specifications

As Oyer and Schaefer (2011) discuss, field experiments varying hiring criteria are relatively rare ("What manager, after all, would allow an academic economist to experiment with the firm's screening, interviewing or hiring decisions?"). In this section, I outline some simple econometric specifications to introduce the experiment and clarify what measurement it enables. The goal of the experiment is to measure the causal effect of changing hiring criteria on characteristics (including the productivity) of selected workers.

There are three major measurement challenges. First, many outcomes (for example, interview performance or on-the-job productivity) are observable only for selected workers. Second, new selection criteria may partially overlap with the old criteria. Candidates identified by a new mechanism might have been selected anyway, and their outcomes should not be fully attributed to the new policy.

Finally, the information environment may contaminate measurement. If algorithms' suggestions are not hidden from human screeners, they may influence human judgements. Unless information flow is controlled, performance from one selection methodology could be misattributed.

Contamination is particularly difficult in candidate-level hiring experiment inside a firm. In

most HR departments, a recruiter has access to a database containing the status of all job applications. The recruiter can see if a candidate he/she rejected was nonetheless interviewed or hired, and may investigate why. The resulting information could affect the recruiter's future assessments. If shared more broadly, this information could also contaminate downstream evaluations by interviewers or by managers. Controlling the information environment is therefore critical for the assessment.

A experiment helps overcome these three challenges. To make the identification strategy transparent, below I present a stylized potential outcomes framework. The framework has been adapted to the hiring setting, and my empirical section mimics this setup. As I will show, the experiment is a form of an encouragement design that can be analyzed through an instrumental variable strategy.

I will begin with notation. Each observation is a job applicant, indexed by i. Each candidate applying to the employer has a true, underlying "type" of $\theta_i \in \{0, 1\}$, representing whether i can pass the test if administered. The potential outcomes for any candidate are $Y_i = 1$ (passed the test) or $Y_i = 0$ (did not pass the test, possibly because the test was not given). Because this empirical strategy is oriented around the firm's strategy, candidates outcomes are coded as zero for candidates are rejected or work elsewhere.[22] For each candidate i, the researcher observes either $Y_i|T = 1$ (whether the test was passed if it occurred) or $Y_i|T = 0$ (whether the test was passed if it didn't occur, which is zero). The missing or unobserved variable is how an untested candidate would have performed on the test, if it had been given.

This framework – and the subsequent experiment – is about the causal effects of adopting a new selection criteria. Suppose we want to compare the effects of adopting a new testing criteria, called Criteria B, against a status quo testing criteria called Criteria A. Criterion A and B can be a "black box" – I will not be relying on the details of how either criteria are constructed as part of the empirical strategy.[23] For any given candidate, $A_i = 1$ means that Criteria A suggests selecting candidate i and $A_i = 0$ means Criteria A suggests *not* selecting i (and similarly for $B = 1$ and $B = 0$). I will refer to $A = 1$ candidates as "A candidates" and $B = 1$ candidates as "B candidates." I'll refer to $A = 1$ & $B = 0$ candidates as "$A \setminus B$ candidates," and $A = 1$ & $B = 1$ as "$A \cap B$ candidates." The Venn diagram in Figure 1 visualizes the scenario.

For many candidates, Criterion A and B will agree. As such, the most informative observations in the data for comparing A and B are where they disagree. If the researcher's data contains A and B labels for all candidates, it would suffice to test randomly selected candidates in $B \setminus A$ and $A \setminus B$ and compare the outcomes. Candidates who are rejected (or accepted) by both methods are irrelevant for determining which strategy is better.[24]

[22] θ represents a generic measure of match quality from the employer's perspective. It may reflect both vertical and horizontal measures of quality. The tests in question may evaluate a candidate in a highly firm-specific manner (Jovanovic, 1979). Y reflects the performance of the candidate on a single firm's private evaluation, which may not necessarily be correlated with the wider labor market's assessment. It is possible that the candidate applied and/or took another test through a different employer, possibly with a different outcome. These outcomes are not used in this procedure for two reasons. First, firms typically cannot access data about evaluations by other companies. Second: Even if they could, the other firm's evaluation may not be correlated with the focal firm's.

[23] In this paper, A is human discretion and B is machine learning. However, A could also be "the CEO's opinion" and B could be "the Director of HR's opinion." One Criteria could be "the status quo," which may represent the combination of methods currently used in a given firm.

[24] Unless there is a SUTVA-violating interaction between candidates in testing outcomes, discussed later.

In many settings, researchers do not know the full extent of disagreement between A and B. This problem is widespread, including the empirical setting of this paper. In HR departments with lots of open information about candidates, the act of measuring B may contaminate evaluation by A. As a result, measuring the amount of intersection and disagreements requires a strategy.

I propose a strategy for addressing this problem below using an instrument (such as a field experiment) for causal inference.[25]

The framework proceeds in two steps. First, I estimate the test success rate of $B \setminus A$ candidates – that is, candidates who would be hired *if and only if* Criteria B were being used and who would be *rejected* if A were used.

Next, I will then compare the above estimate to a series of benchmarks. For most questions about Criteria A versus B, the relevant baseline is the $A \setminus B$ candidates (ones that A approves and B doesn't). However, other benchmarks may be interesting or relevant as well. In my empirical section, I also compare the estimand to the uccess rate of $A \cap B$ candidates ("intersection" candidates that both criteria approve) and for all A candidates.

To measure the success rate of "B only" candidates ($E[Y|T = 1, A = 0, B = 1]$, or outcomes of candidates who would be rejected by Criteria A, but tested by Criteria B), the researcher needs an instrument, Z_i, which selects B candidates in a way that is uncorrelated with the each candidate's assessment on A. Because the status quo selects only A candidates, the effect of the instrument is to select candidates who would otherwise not be tested.[26]

For exposition, suppose the instrument Z_i is a binary variable at the candidate level. It varies randomly between one and zero with probability 0.5; it could be a 50/50 coin flip for all candidates for whom $B_i = 1$. In order to measure the marginal yield of Criteria B, we need variation in Z_i within $B_i = 1$.[27] The instrument Z_i within $B_i = 1$ is "local" in that that it only varies for candidates approved by Criteria B. The instrument Z_i must affect probability that each candidate is interviewed or tested. For the experiment in this paper, the firm tests all candidates for whom $Z_i = 1$, irrespective of A_i.

We can now think of all candidates as being in one of four types: a) "Always tested" – these are candidates for whom $T_i = 1$ irrespective of whether Criterion A or B are used ($A_i = B_i = 1$), b) "Never tested," for which $T_i = 0$ irrespective of Criteria A or B ($A_i = B_i = 0$). The instrument does not effect whether these two groups are treated. Next, we have c) "Z-compliers," who are tested only if $Z_i = 1$, and d) "Z-defiers," who are tested only if $Z_i = 0$.

Identification of this "local average testing yield" requires the typical five IV conditions. I outline each condition in theory in Appendix C, with some interpretation of these assumptions in a hiring setting. In the following section (4.1), I show that each condition is met for my empirical setting.

Under these assumptions, we can estimate the average yield of $A = 0$ & $B = 1$ candidates as:

[25] Aside from the restrictions above, no additional assumptions about the distribution of θ are required, nor are assumptions about the correlation between A, B and θ.

[26] One cannot test all B candidates or a random sample of them, because some of the B candidates are also A candidates.

[27] Additional random variation in Z_i beyond $B = 1$ is not problematic, but isn't necessary for identifying $E[Y|T = 1, A = 0, B = 1]$. Z_i can be constant everywhere $B = 0$.

$$E[Y|T=1, A=0, B=1] = \frac{E[Y_i|Z_i=1, B_i=1] - E[Y_i|Z_i=0, B_i=1]}{E[T_i|Z_i=1, B_i=1] - E[T_i|Z_i=0, B_i=1]} \tag{1}$$

The value above can be estimated through two-stage least-squares in a procedure akin to instrumental variables (Angrist et al., 1996). The first stage is the instrument on the binary decision to test, and the second stage is the testing decision on the test outcome (success/failure). The resulting estimand is a success rate of the candidates tested by B but not A. This estimand has units of "*new successful tests* over *new administered tests*."[28]

Next, I show how the IV conditions are met in my empirical setting. Then, I show how to extend this framework into other downstream outcomes after screening performance (such as on-the-job performance).

4.1 Application to Empirical Setting via Field Experiment

In my empirical setting, all incoming applications (about 40K candidates) were scored and ranked by the algorithm shortly after each application was submitted. For the contamination issues mentioned above, the algorithm worked behind the scenes and without the knowledge of rank-and-file recruiting staff or future interviewers and managers.[29] The algorithm was calibrated to intervene on candidates with an estimated probability of 10% (or greater) of getting a job offer were flagged as "machine approved."[30] This group comprised about 800 applicants over roughly one year. While this seems like a small number of candidates, this group comprised about 30% of the firm's hires from this applicant pool over the same time period.

After these candidates were identified, a random variable $Z \in \{0, 1\}$ was drawn for all machine-picked candidates (50% probability of each). Candidates randomly assigned a "1" were automatically granted an interview. Those randomly assigned "0" – along with the algorithm-rejected candidates – proceeded through the status quo channels (assessment by humans process)

To avoid contamination, the algorithm's role in selecting or rejecting candidates was hidden from other HR workers, interviewers or future co-workers. For $Z = 1$ candidates who the algorithm approved and nudged to get an interview, a normal database entry suggested that a human screener selected the candidate. For all other candidates, no "algorithmic approval/rejection" notice was appended to the HR file. This practice – paired with the conditional randomization – allows the HR staff to examine the resume as they normally would (independently and without contamination from the machine). As mentioned above, the *existence* of the machine selection was hidden from

[28] β_{2SLS} is the ratio of the "reduced form" coefficient to the "first stage" coefficient. In this setup, the "reduced form" comes from a regression of Y on Z, and the "first stage" comes from a regression of T on Z. Applied in this setting, the numerator measures new successful tests caused by the instrument, and the denominator estimates new administered tests caused by the instrument. The ratio is thus the marginal success rate – new successful tests per new tests taken.

[29] Hiding similar information is normal. Most workers are not told all the details of who are how they are selected.

[30] The threshold of 10% was chosen in this experiment for capacity reasons. The experiment required the firm to spend more resources on interviewing in order to examine counterfactual outcomes in disagreements between the algorithm and human. Thus the experiment required an expansion of the firm's interviewing capacity. The ≈10% threshold was selected in part because the firm's interviewing capacity could accommodate this amount of extra interviews without overly distracting employees from productive work.

rank-and-file recruiting staff and line managers (for contamination reasons). The human screeners no choice than to evaluate the candidates independently.

This obfuscation is common; most workers are spared the exact details of how they wound up getting an offer. For example, they are not told whether they were a top choice or a backup candidate; or which managers liked or disliked the candidate during the interview process; or whether the candidate was hired as part of a normal program or a special outreach to diversity candidates. Knowing this information could affect relationships for candidates eventually hired. Many firms compartmentalize information about how candidates are assessed to promote independent assessments and avoid information cascades.

The random binary variable Z acts as an instrument for interviewing that can be used with the potential outcomes framework above. Candidates selected for an interview (from either method) were sent blindly into an interview process. Neither the interviewers nor the candidates were told about the experiment or which candidates (if anyone) came from which selection process. The IV conditions mentioned previously (in Appendix C) are met as follows:

1. **SUTVA**: SUTVA would be violated if the treatment group's outcomes interact with the control group's. This would be problematic if an employer had an inelastic quota of hiring slots. In my empirical setting and many others, the employer's policy is to make an offer to anyone who passes the test. "Passing" depends on performance on the test relative to an objective standard, and not by a relative comparison between candidates on a "curve."[31]

2. **Ignorable assignment of** Z. Covariate balance tests in Table 2 appear to validate the randomization.

3. **Exclusion restriction**, or $Y(Z,T) = Y(Z',T), \forall Z, Z', T$. In my empirical setting, the instrument is a randomized binary variable Z_i. This variable was hidden from subsequent screeners. Graders of the test did not know which candidates were approved (or disapproved) by Criteria A or B, or which candidates (if any) were affected by an instrument. The existence of the experiment and instrument were never disclosed to test graders or candidates – the evaluation by interviewers was double-blind.

4. **Inclusion restriction**. The instrument must have a non-zero effect on who is tested. In my empirical setting, this is clearly met. The experiment strongly affected who was interviewed. B candidates were +30% more likely to be interviewed if when $Z_i = 1$.

5. **Monotonicity**. The instrument here was used to guarantee certain candidates an interview, and not to deny anyone an interview (or make one less likely to be interviewed).

The econometric setup above does not require that the two methods test the same *quantity* of candidates. This is a useful feature that makes the approach more generic: Many changes in testing or hiring policy may involve tradeoffs between the quantity and quality of examined candidates.

[31] This policy is common in many industries where hiring constraints are not binding – for example, when there are few qualified workers, or workers who are interested in joining the firm, compared to openings. As Lazear et al. (2016) discuss, much classical economic theory does not model employers face an inelastic quota of "slots." Instead it models employers featuring a continuous production function where tradeoffs are feasible between worker quantity, quality and cost.

In my empirical setting, the machine learning algorithm identified 800 candidates, and the human screeners identified a larger number (TODO). It's possible that the higher success rate is the result of extending offers to fewer, higher quality people. To address this, I will compare the outcomes of machine-only candidates not only to the average human-only candidate, but also to the average candidate selected by both mechanisms (of which there were much fewer). Then, I will fix the quantities of interviews available to both mechanisms to measure differences in yield, conditional on an identical "budget" of interviews.

4.2 Offer accepts, on-the-job productivity and other "downstream" outcomes

In some cases, firms may care about downstream outcomes after the job test or interview. For example: They may care about who accepts extended job offeres, or who performs well as an employee after testing and hiring. It's possible that a new interviewing criteria identifies candidates who pass, but do *not* accept offers (perhaps because many other firms have simultaneously recruited these candidates). The framework above can be extended to measure how these outcomes are affected by changes to screening.

For these empirical questions, a research can use a different Y (the outcome variable measuring test success). Suppose that $Y_i' = 1$ if the candidate was tested, passed *and* accepted the offer. This differs from the original Y, which only measures if the test was passed. Using this new variable, the same 2SLS procedure can be used to measure the effects of changing Criteria A to B on offer-acceptance or other downstream outcomes. Such a change would estimate a local average testing yield whose units are *new accepted offers / new tests*, rather than *new tests passed (offers extended) / new tests*.

In some cases, a researcher may want to estimate the offer acceptance rate, whose units are "offer accepts" / "offer extends." The same procedure can be used for this estimation as well, with an additional modification. In addition changing Y to Y', the researcher would also have to change the endogenous variable T to T' (where $T' = 1$ refers to being extended an offer). In this setup, the instrument Z_i is an instrument for receiving an offer rather than being tested. This can potentially be the same instrument as previously used. The resulting 2SLS coefficient would deliver an estimand whose units are "offer accepts" / "offer extends" for the marginal candidate.

Accepting offers is one of many "downstream" outcomes that researchers may care about. We may also care about how downstream outcomes such as productivity and retention once on the job, as well as the characteristics of productivity (innovativeness, efficiency, effort, etc). This would requiring using an outcome variable Y' representing "total output at the firm" (assuming this can be measured), whose value is zero for those who aren't hired. T' would represent being hired, and Z_i would need to instrument for T (being hired). This procedure would estimate the change in downstream output under the new selection scheme.[32]

We can think of these extensions as a form of imperfect compliance with the instrument. As the econometrician studies outcomes at increasingly downstream stages, the results become increasingly "local," and conditional on the selection process up to that stage. For example, results about accepted job offers may be conditional on the process process for testing, interviewing, persuasion,

[32] In some cases, such as the setting in this paper, it could make more sense to study output per day of work.

compensation and bargaining with candidates in the setting being studied. The net effectiveness of A vs B ultimately depends on how these early criteria interact with downstream assessments.

The analysis of downstream outcomes requires the IV assumptions to be revisited. Even if the IV assumptions are met for the initial phase, they may not for the later outcomes. In Appendix D, we revisit the IV assumptions for important downstream outcomes after hiring, including the main ones analyzed in the empirical section of this paper.

5 Results

I begin by analyzing the strength of the intervention. Table **??** shows how the randomization influenced who is interviewed. Approximately 50% of the machine-approved candidates were interviewed, even without the machine's encouragement. The randomized intervention added an additional +30% of candidates who were approved by the algorithm, and rejected by the humans.[33] The most common reason cited for the human rejection in this group is lack of qualifications.

I next examine the performance of the 30% of candidates that the machine liked, and the humans rejected in interviews. Table **??** shows that the marginal candidate passes interviews in 37% of cases – about X% more than the candidates picked by humans only, and X% more than the candidates selected both by the human and the machine.

Table **??**. The marginal candidate accepts a job offer extended about 87% of cases, which is about 15% higher than the average in the control group. Tests of statistical significance of theses differences are reported in the bottom of Panel C, Table 3. In Table 5, I show that the machine candidates are are less likely to negotiate their offer terms.

In the above analysis, the machine was permitted to interview more candidates than the human. A separate question is whether the machine candidates would perform better if its capacity was constrained to equal the human's. In Table 4, I repeat the above exercise but limit the machine's quantity to match the human's. In this case, the results are sharper. The machine selected candidates improve upon the human passthrough rates.

5.1 Job Performance

The candidates who are hired go on to begin careers at their firm, where their career outcomes can be measured. I examine variables relating to technical productivity. The jobs in this paper involve developing software. As with many companies, this code is stored in a centralized repository (similar to `http://github.com`) that facilitates tracking programmer's contributions to the base of code.

This system permits reporting about each programmer's lines of code added and deleted. I use these as rudimentary productivity measures. Later, I use these variables as surrogate outcomes (Prentice, 1989; Begg and Leung, 2000) for subjective performance reviews and promotion using

[33] In both treatment and control groups the remaining ≈20% withdrew their applications prior to the choice to interview, usually because they already accepted another job.

the Athey et al. (2016) framework.

The firm doesn't create performance incentives on these metrics, in part because it would encourage deliberately inefficient coding. The firm also uses a system of peer reviews for each new contribution of code.[34] These peer reviews cover both the logical structure, formatting and readability of the code as outlined in company guidelines.[35] These peer reviews and guidelines bring uniformity and quality requirements to the definitions of "lines of code" used in this study.

Despite the quality control protocols above, one may still worry about these outcome metrics. Perhaps the firm would prefer fewer lines of elegant and efficient code. A great programmer should thus have fewer lines of code and perhaps delete code more often. As such, I examine both lines of code added and deleted in Table 7. These are adjusted to a per-day basis and standardized. The conclusions are qualitatively similar irrespective of using adding or deleting lines: The marginal candidate interviewed by the machine both adds and deletes more lines of code than those picked by humans from the same pool.

5.2 Cultural Fit and Leadership Skills

During the sample period of the experiment, the employer in this experiment began asking interviewers for additional quantitative feedback about candidates. The additional questions asked interviewers to assess the candidate separately on multiple dimensions. In particular, they asked interviewers for an assessment of the candidate's "general aptitude," "cultural fit" and "leadership ability." The interviewers were permitted to assess on a 1-5 scale. These questions were introduced to the interviewers gradually and orthogonally to the experiment.

Because of the gradual introduction, do not have assessments for all of the candidates in the experiment. In order to expand the sample size, I combine the variation from the experiment with regression discontinuity around the 10% threshold. For the regression discontinuity, I use the Imbens and Kalyanaraman (2011) bandwidth. The machine picked candidates aren't different from the human picked ones in general aptitude, but are more highly rated in soft dimensions such as cultural fit and leadership.

5.3 Benefit driven by better weighting of a small number of variables.

- Benefit driven by better weighting of a small number of variables.
- Rather than utilizing more non-obvious variables.
- Theoretically possible to reverse-engineer a low-dimensional linear model that delivers 70-90% of benefits of the full ML model.
- Cognitive limitations and *"attending to more variables"* is more important for *learning* and less for scoring (Hanna et al., 2014; Schwartzstein, 2014)

[34] For a description of this process, see https://en.wikipedia.org/wiki/Code_review.

[35] See descriptions of these conventions at https://en.wikipedia.org/wiki/Coding_conventions and https://en.wikipedia.org/wiki/Programming_style.

5.4 Combining Human and Machine Signals

In the "treatment" branch of the experiment, all machine-approved candidates were automatically given an interview. Before these candidates' were interviewed, they were shown to human screeners who were informed that the algorithm had suggested interviewing this candidate. The human screeners were next asked if they agreed with the machine's decision to interview. This is a similar setup to the control group, except that in the control group the machine's preference was blind.

After learning the machine's choice, the human screeners agreed on 85% of non-withdrawn applications (70% of total applications). By contrast, in the control group – where human screeners were asked for *independent* evaluations without knowing the machine's choice – the humans agreed on only 60% of non-withdrawn applications (50% of total applications).

This large difference suggests that the human screeners substantially change their minds after learning the machine's choice. The humans' propensity to agree with the algorithm speaks to how much the human screeners *themselves* place faith in their own private signals of quality. We observe this difference, even though the screeners were not told details of how the algorithm worked or about its performance.

After recording their agreement (or disagreement), the screeners were also asked to assess the treatment candidate on a 1-5 scale. In Table 12, I measure whether these human provided signals contain information using "horserace" regressions (Fair and Shiller, 1989).

I find that in isolation, the human evaluations contain some predictive information. That is, they can predict which candidates among the machine-selected candidates will successfully pass interviews. However, when both signals can be combined, nearly all weight should be placed on the machine's score of the candidates. Once the algorithm's ranking enters the regression, the human evaluation offers no additional predictive power.

Regarding candidates' acceptance of extended offers, I show in Panel B of Table 12 the human's assessment has no predictive power, even in isolation. The machine's ranking does.

5.5 Heterogeneity

5.6 Role of Noise

6 Conclusion

Implications for management?

6.1 Coefficients versus Treatment Effects

References

Adair, Alastair, Norman Hutchison, Bryan MacGregor, Stanley McGreal, and Nanda Nanthakumaran, "An analysis of valuation variation in the UK commercial property market: Hager and Lord revisited," *Journal of Property Valuation and Investment*, 1996, *14* (5), 34–47.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Exploring the Impact of Artificial Intelligence: Prediction versus Judgment," 2017.

Anderson, James M, Jeffrey R Kling, and Kate Stith, "Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines," *The Journal of Law and Economics*, 1999, *42* (S1), 271–308.

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

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks," *ProPublica, May*, 2016, *23*.

Arrow, Kenneth J., *The Theory of Discrimination*, Princeton University Press, 1973.

Athey, Susan and Guido Imbens, "NBER Summer Institute 2015 Econometric Lectures: Lectures on Machine Learning," 2015.

_ , Raj Chetty, Guido Imbens, and Hyunseung Kang, "Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index," *arXiv preprint arXiv:1603.09326*, 2016.

Autor, David H and Susan N Houseman, "Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from "Work First"," *American Economic Journal: Applied Economics*, 2010, pp. 96–128.

Azevedo, Eduardo M, Deng Alex, Jose Montiel Olea, Justin M Rao, and E Glen Weyl, "A/b testing," 2018.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki, "The impact of big data on firm performance: An empirical investigation," Technical Report, National Bureau of Economic Research 2018.

Becker, Gary S, "The economics of discrimination Chicago," *University of Chicago*, 1957.

Begg, Colin B and Denis HY Leung, "On the use of surrogate end points in randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2000, *163* (1), 15–28.

Brynjolfsson, Erik and Tom Mitchell, "What can machine learning do? Workforce implications," *Science*, 2017, *358* (6370), 1530–1534.

_ , _ , and Daniel Rock, "What Can Machines Learn, and What Does It Mean for Occupations and the Economy?," in "AEA Papers and Proceedings," Vol. 108 2018, pp. 43–47.

Busse, Meghan R, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso, "The psychological effect of weather on car purchases," *The Quarterly Journal of Economics*, 2015, *130* (1), 371–414.

Card, David and Gordon B Dahl, "Family violence and football: The effect of unexpected emotional cues on violent behavior," *The Quarterly Journal of Economics*, 2011, *126* (1), 103–143.

Chang, Tom and Antoinette Schoar, "Judge specific differences in Chapter 11 and firm outcomes," *Unpublished working paper, National Bureau of Economic Research Cambridge*, 2013.

Chouldechova, Alexandra and Max G'Sell, "Fairer and more accurate, but for whom?," *arXiv preprint arXiv:1707.00046*, 2017.

Cockburn, Iain M, Samuel Kortum, and Scott Stern, "Are all patent examiners equal? The impact of examiner characteristics," Technical Report, National Bureau of Economic Research 2002.

Colbert, Janet L, "Inherent risk: An investigation of auditors' judgments," *Accounting, Organizations and society*, 1988, *13* (2), 111–121.

Corchón, Luis C, Marco Serena et al., "Contest theory," *Handbook of game theory and industrial organization*, 2018, *2*, 125–146.

Cortes, Corinna and Vladimir Vapnik, "Support-vector networks," *Machine learning*, 1995, *20* (3), 273–297.

Cowgill, Bo, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening," *Working Paper*, 2017.

_ **and Eric Zitzewitz**, "Mood Swings at Work: Stock Price Movements, Effort and Decision Making," 2008.

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

Dimakopoulou, Maria, Susan Athey, and Guido Imbens, "Estimation Considerations in Contextual Bandits," *arXiv preprint arXiv:1711.07077*, 2017.

Dobbie, Will and Jae Song, "Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection," *The American Economic Review*, 2015, *105* (3), 1272–1311.

Edmans, Alex, Diego Garcia, and Øyvind Norli, "Sports sentiment and stock returns," *The Journal of Finance*, 2007, *62* (4), 1967–1998.

Engelberg, Joseph and Christopher A Parsons, "Worrying about the stock market: Evidence from hospital admissions," *The Journal of Finance*, 2016.

Fair, Ray C and Robert J Shiller, "The informational content of ex ante forecasts," *The Review of Economics and Statistics*, 1989, pp. 325–331.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, "What is a Patent Worth? Evidence from the US Patent "Lottery"," Technical Report, National Bureau of Economic Research 2017.

Frey, Carl Benedikt and Michael A Osborne, "The future of employment: how susceptible are jobs to computerisation?," *Technological forecasting and social change*, 2017, *114*, 254–280.

Friedler, Sorelle A. and Christo Wilson, eds, *Conference on Fairness, Accountability and Transparency, 23-24 February 2018*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 2018.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer-Verlag New York, 2013.

Gentzkow, Matthew and Jesse M Shapiro, "What Drives Media Slant? Evidence from US Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71.

Grimstad, Stein and Magne Jørgensen, "Inconsistency of expert judgment-based estimates of software development effort," *Journal of Systems and Software*, 2007, *80* (11), 1770–1777.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Hardt, Moritz, Eric Price, Nati Srebro et al., "Equality of opportunity in supervised learning," in "Advances in Neural Information Processing Systems" 2016, pp. 3315–3323.

Heckman, James, "Sample Selection Bias as a Specification Error.," *Econometrica*, 1979.

Hirshleifer, David and Tyler Shumway, "Good day sunshine: Stock returns and the weather," *The Journal of Finance*, 2003, *58* (3), 1009–1032.

Hoffman, Mitch, Lisa B Kahn, and Danielle Li, "Discretion in Hiring," 2016.

Imbens, Guido and Karthik Kalyanaraman, "Optimal bandwidth choice for the regression discontinuity estimator," *The Review of economic studies*, 2011, p. rdr043.

Joseph, Matthew, Michael Kearns, Jamie H Morgenstern, and Aaron Roth, "Fairness in learning: Classic and contextual bandits," in "Advances in Neural Information Processing Systems" 2016, pp. 325–333.

Jovanovic, Boyan, "Job matching and the theory of turnover," *The Journal of Political Economy*, 1979, pp. 972–990.

Jr, Joseph J Doyle, "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care," *Journal of political Economy*, 2008, *116* (4), 746–770.

_ et al., "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *American Economic Review*, 2007, *97* (5), 1583–1610.

Kahneman, Daniel, "Remarks by Daniel Kahneman," *NBER Economics of AI Conference*, 2017.

_ , M Rosenfield, Linnea Gandhi, and Tom Blaser, "Noise: How to overcome the high, hidden cost of inconsistent decision making," *Harvard Business Review*, 2016, *10*, 38–46.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2017, *133* (1), 237–293.

_ , **Sendhil Mullainathan, and Manish Raghavan**, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

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

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva, "Counterfactual fairness," in "Advances in Neural Information Processing Systems" 2017, pp. 4066–4076.

Lambrecht, Anja and Catherine E Tucker, "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads," 2016.

Lazear, Edward P and Sherwin Rosen, "Rank-order tournaments as optimum labor contracts," *Journal of political Economy*, 1981, *89* (5), 841–864.

_ , **Kathryn L Shaw, and Christopher T Stanton**, "Who Gets Hired? The Importance of Finding an Open Slot," Technical Report, National Bureau of Economic Research 2016.

Li, Danielle, "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 2017, *9* (2), 60–92.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *The American Economic Review*, 2013, *103* (5), 1797–1829.

Marschak, Jacob, "Binary choice constraints and random utility indicators," Technical Report, YALE UNIV NEW HAVEN CT COWLES FOUNDATION FOR RESEARCH IN ECONOMICS 1959.

Mullainathan, Sendhil and Ziad Obermeyer, "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, 2017, *107* (5), 476–480.

Oyer, Paul and Scott Schaefer, "Personnel Economics: Hiring and Incentives," *Handbook of Labor Economics*, 2011, *4*, 1769–1823.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, pp. 659–661.

Prentice, Ross L, "Surrogate endpoints in clinical trials: definition and operational criteria," *Statistics in medicine*, 1989, *8* (4), 431–440.

Rind, Bruce, "Effect of beliefs about weather conditions on tipping," *Journal of Applied Social Psychology*, 1996, *26* (2), 137–147.

Sampat, Bhaven and Heidi L Williams, "How do patents affect follow-on innovation? Evidence from the human genome," *available at http://economics.mit.edu/files/9778*, 2014.

Schwartzstein, Joshua, "SELECTIVE ATTENTION AND LEARNING," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Schwarz, Norbert and Gerald L Clore, "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states.," *Journal of personality and social psychology*, 1983, *45* (3), 513.

Selten, Reinhard, "Reexamination of the perfectness concept for equilibrium points in extensive games," *International journal of game theory*, 1975, *4* (1), 25–55.

Slovic, Paul, "Analyzing the expert judge: A descriptive study of a stockbroker's decision process.," *Journal of Applied Psychology*, 1969, *53* (4), 255.

Stern, Léa H, Isil Erel, Chenhao Tan, and Michael S Weisbach, "Selecting Directors Using Machine Learning," 2018.

Sweeney, Latanya, "Discrimination in online ad delivery," *Queue*, 2013, *11* (3), 10.

Taylor, Robert L and William D Wilsted, "Capturing judgment policies: A field study of performance appraisal," *Academy of Management Journal*, 1974, *17* (3), 440–449.

Tibshirani, Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267–288.

Vapnik, Vladimir Naumovich, *Estimation of dependences based on empirical data [In Russian]* 1979. English Translation by Kotz, Samuel in 1982 by publisher Springer-Verlag New York.

Wager, Stefan and Susan Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2017, (just-accepted).