Automating Judgement and Decisionmaking: Theory and Evidence from Résumé Screening

Bo Cowgill[*]
Columbia University

May 5, 2017

Preliminary and incomplete. Do not circulate.

Abstract

What types of decisionmaking tasks are better automated? And which are better left to judgement? I develop a formal model of the comparative advantages of human judgement and machines in decisionmaking. I subsequently test these predictions in a field experiment in applying machine learning for hiring workers for white-collar team-production jobs. The marginal candidate picked by the machine (but not by human screeners) is +17% more likely to pass a face-to-face interview with incumbent workers and receive a job offer offer, b) +15% more likely to accept job offers when extended by the employer and c) 0.2σ-0.4σ more productive once hired as employees. Consistent with the model, the algorithm's advantage comes in selecting candidates from high variance applicant pools, such as candidates who lack job referrals, those without prior experience, those with atypical credentials and those completing a PhD. I also find that machine candidates are +12% less likely to show evidence of competing job offers during salary negotiations. I also find that algorithmic judgment results in better performance in abstract, hard-to-measure dimensions of candidates – such as leadership, cultural fit and other non-cognitive "soft" dimensions – which is a prediction of the model. I conclude by discussing the implications of machine learning for a variety of decisionmaking tasks throughout the economy.

[*] The author thanks seminar participants at the 2015 Empirical Management Conference, the Institutions and Innovation Conference at Harvard Business School, the Tinbergen Institute, Universidad Carlos III de Madrid, NBER Economics of Digitization, 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, NBER Summer Institute (Labor), 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, Laura Gee, Dan Gross, John Horton, Danielle Li, Christos Makridis, Stephan Meier, John Morgan, Harikesh Nair, Paul Oyer, Tim Simcoe and Jann Spiess.

1 Introduction

Social skills and other *soft skills* are increasingly rewarded in the labor market. Deming (2015) and Weinberger (2014) provides compelling evidence that in the past 40 years, employment and wage growth has been strongest in jobs that require both cognitive skill and social skill. Deming (2015) shows that nearly all job growth in the US since 1980 has been in occupations intensive in social skills.

One reason for this increase is that social interactions and cultural understanding requires subtle, tacit knowledge that cannot be easily automated (Autor, 2015). Thus workers bearing these skills cannot be easily displaced by machines. Economists in this literature cite evolutionary theorists (Moravec, 1988) who argue that social interaction is an unconscious process that evolved over thousands of years – thus cannot be easily replicated by robots. In this sprit, Frey and Osborne (2013) identify *social intelligence tasks* as a key bottleneck to further automation. Armed with this reasoning, economists, policymakers and education reformeres have proposed curriculum changes, suggesting that tomorrow's students should learn social skills to protect their employability against technological innovation.

Are human social skills truly so productive and resistant to automation? The answer may affect the future returns to soft skills. Many of the widespread shortcomings documented in behavioral economics – attribution errors, homophily bias and others – are shortcomings in *social* judgements. In other contexts, the evolutionary origins of human behavior were irrelevant to automation[1] or present clear disadvantages.[2] Algorithms have been often suggested as a way to improve upon flawed human decision-making (Kahneman, 2011), not to coarsen it.

Do humans truly have better social skills?[3] The idea might be a seductive illusion kept alive by measurement challenges. Counterfactual social judgements are rarely observed. Even when they are, scoring and evaluating outcomes may be subjective, noisy or require years to realize.[4] The

[1] For example, many scholars belive that facial recognition is a subconscious process that required thousands of years to evolve (Parr, 2011). However, modern facial recognition software is capable of outperforming humans at labeling similar faces (Taigman et al., 2014; Schroff et al., 2015).

[2] For example, in dieting.

[3] A popular of evidence about automating social interactions is the failure of progress on the "Turing Test." The Turing Test is a laboratory game in which a human judge converses via text with an anonymous agent. After an unstructured text conversation, the judge must guess if the agent is a person or a computer program. A computer program "passes" if it successfully fools the judge into labeling it a human, which is difficult for machines to accomplish. The judge's Type II errors are rarely disclosed or discussed.

There are many reasons to be skeptical of the generalizability of this laboratory result. Some computer programs succeed by deliberately adopting common human *errors*, rather than emulating the productivity and utility of human judgement. See LaCurts (2011) for a discussion of the limits of Turing Test results.

Notably: In real-world settings – where firms face non-laboratory incentives to distinguish robots from humans – businesses rely on machine judges, not human judges. For example: Unwanted computer-generated email costs Americans over $20 billion annually and generates $200M in revenues for spam companies (Rao and Reiley, 2012). The job of separating computer-generated emails from "genuine" human-generated messages is a real-world "Turing Test." Real-world firms in this business delegate the task to algorithms, not humans.

[4] In many settings, agents may actively obfuscate how social outcomes are scored. In some circumstances, good social skills by one party may *require* ignoring or mislabeling one's own (or someone else's) previous bad social decisions. For example: Holiday gift-giving is an activity requiring social skills but that may result in mismatch (Waldfogel, 1993). Nonetheless, recipients of undervalued gifts often report happiness with the chosen gift, thus obfuscate scoring of the gift choice. Similarly: A boss may attempt to divert attention from bad hiring decisions in order to avoid embarrassment,

unit of analysis may not be clear: This literature often stresses "interactiveness," in which many social decisions are bundled together. The measurement issues around social skills also creates moral hazard in contracting. If identifying a job candidate's "cultural fit" cannot easily be verified, workers can exploit ambiguity for private gain. For example, a hiring manager may feign "cultural fit" as cover for employing cronies, relatives or favorites instead of a better-suited worker.

The features that make social interactions difficult to automate also make these tasks difficult to measure and contract upon.[5] By contrast, differences in hard skills (such as mathematics) may be much easier to both measure and contract upon.[6] These measurement and contracting differences – and not underlying technological performance differences – may be why we believe that hard skills can be effectively automated while soft skills cannot.

This paper examines these questions experimentally in the setting of hiring for team-based, white-collar jobs. Hiring decisions in this environment are intimately linked to social interaction and compatibility. In many industries, new employees often first must learn and be mentored by incumbent employees, and then later interact and collaborate with other workers in order to create value. A recent survey of ≈3,000 hiring managers, 43% listed "cultural fit" as the "single most important determining factor when making a new hire."[7] The importance of social relationships and compatibility are one reason that many firms tap into pre-existing social relationships (see Burks et al., 2015) in hiring decisions.

Deming (2015) writes that effective social skills require "the capacity that psychologists call *theory of mind* - the ability to attribute mental states to others based on their behavior, or more colloquially to 'put oneself into anothers shoes' (Premack and Woodruff, 1978; Baron-Cohen, 2000; Camerer et al., 2005)" or "[r]eading the minds of others."

By this definition of social skills, hiring managers must not only read the candidate's mind, but also read the minds of interviewers, co-workers, managers and clients who may interact with the candidate in the future. Screeners involved in hiring must make inferences about soft skills based on social signals in letters of reference, interviews and CVs. Efficient hiring requires selecting candidates likely to pass job interviews – face-to-face exchanges with incumbent workers in which subconscious and subtle non-verbal communication influence outcomes – and who are likely to cooperate productively with future teammates in on-the-job work.

Hiring decisions have attractive measurement properties: In this setting, the exercise of social perceptiveness involve discrete decisions with a natural unit of analysis: For a given candidate, should he or she be interviewed? This discreteness is distinct from the many bundled decisions of real-time conversations. The use of a field experiment enables researchers to observe counterfac-

avoid disruptive firings and/or focus attention how to best utilize the strengths of the hired worker. This may compel the boss to misrepresent the employee's quality – or to compensate by providing extra coaching and tutoring for the failed candidate. In both cases, social pressures require agents to confound measurement and obfuscate how outcomes are scored.

[5] A related contracting issue is: When outcomes take a long time to be realized – as is the case in hiring and many other tasks requiring social skills – it may be difficult to incentivize present-biased agents.

[6] There is often a single right answer and a single, provably optimal method for arriving at this answer; as such there is less subjectivity in scoring outcomes. The timeliness of measuring hard skills is easier. Computers derive much of their benefit from fast computation, and human subjects may quickly realize when they are stumped (hastening the time necessary to measure and compare outcomes).

[7] http://about.beyond.com/press/releases/20140520-National-Survey-Finds-College-Doesnt-Prepare-Students-for-Job-Search

tual outcomes and observe differences. How does a human and machine screener assess the same set of candidates? How often does the human and algorithm disagree? Who is more "correct" in these disagreements – particularly on dimensions of performance relating to "soft" skills such as cultural fit, leadership and social interaction?

To understand these issues, I develop a model of automating subjective judgements. In the model, a mass of human agents face a series of binary decisions. Each has a taste-based preference which is hidden from the principal. Each agent also has a statistical model of the the "correct" decision, but can only be compeled via contracting to make that decision regularly. Contracting may fail for two reasons: First, because outcomes are hard to verify which weakens incentives. Second, because the agents' statistical models are rudimentary and update slowly.

I then compare the performance of this mass of humans against machine learning trained on historical data by the mass of human agents. The machine learning effectively integrates over each human individuals' idiosyncratic tastes, canceling out much (but not necessarily all) of the biases of the human processes. The algorithm's advantage thus comes from three sources: First, canceling out idiosyncratic biases. Second, attending to more variables and thus having a better statistical model, and third: Because the algorithm can pool information from the entire mass of humans, it can update its model and can learn more quickly. The algorithm is effectively a form of organizational centralization with superior coordination – in the form of learning – between agents (Alonso et al., 2008).

I then examine the results of a field experiment motivated by this model. The field experiment yields four main results.

First, the machine and human screeners disagree on about 30% of candidates. I find that the marginal candidate picked by the machine (but not by the human) is +17% more likely to pass a double-blind face-to-face interview with incumbent workers and receive a job offer offer. The marginal candidate picked by a human (but not the machine) is *less* likely to pass the double-blind interview. I show evidence that the machine candidates pass the interview panel in part because their worst (most negative) interview evaluation from the panel are more favorable. By contrast, the most positive evaluations from the panel are roughly similar to the human-picked candidates. The algorithm benefits candidates coming from an atypical career or educational backgrounds (for example, a school not attended by any other applicant in the firm's applicant pool).

Second, I find that are also more likely to accept job offers conditional on being extended. They are also about 12% less likely to show evidence of competing job offers during salary negotiations, and are 15% more likely to accept job offers when extended by the employer.

Third, these are about 0.2σ-0.4σ more productive once hired as employees.

Lastly, evaluations of the candidates shows that advantage of the machine comes from selecting candidates with superior 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 prior literature suggests humans – and not machines – have innately superior judgement.

I also find that 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 most weight on the machine signal.

While this setting has inherent limitations, these results show evidence of productivity gains from IT adoption in tasks requiring social judgements. Limiting or eliminating human discretion through this form of digitization 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 discusses the empirical setting and experimental design, and in section Section 4 I summarize results. Section **??** 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.

2 Empirical Setting

The job openings in this paper are technical staff such as programmers, hardware engineers and software-oriented technical scientists and specialists. Workers in this industry are involved in multi-person teams that design and implement technical products.

Successful contributions in this environment requires workers to collaborate with colleagues. In a typical project, a new product can be conceptualized as several interacting technical "modules" that function together as a coherent product. Each team member is tasked with designing and implementing a module, and ensuring that the technology of his or her module cooperates with others'. The team discusses as a group to achieve consensus on the macro-level segmentation of the product into "modules" and the assignment of various modules to teammates. Frequently circumstances arise that require these workers to switch module assignments. For example, some modules may take unexpectedly long and need to be subdivided. This resembles Deming's 2015 "trading tasks."

The internal promotion process in this market often involves peer feedback and subjective performance reviews. In fact, the incentives for positive subjective reviews from workplace peers are so strong that a number of scholars and journalists have expressed concern that these systems encourage "influence activities" (Milgrom and Roberts, 1988; Milgrom, 1988, Gubler et al., 2013) – that is, the system encourages social skills rather than programming. Eichenwald's 2012 journalistic account of Microsoft's promotion system[8] says that "[E]very employee has to impress not only his or her boss but bosses from other teams as well. And that means schmoozing and brown-nosing as many supervisors as possible."

Work in this industry thus involves substantial amounts of coordination, negotiation, persuasion and social perceptiveness – which the four skills in the the O*NET database in used by Deming (2015) to label jobs requiring social skills. This is especially true if one considers the behaviors necessary to be promoted. Consistent with this account, the occupations corresponding to this

[8] http://www.vanityfair.com/business/2012/08/microsoft-lost-mojo-steve-ballmer

work rank above the median in the O*NET database.[9]

Separately from the underlying job details, the *hiring process itself* in white-collar work often requires substantial coordination, negotiation, persuasiveness and social perception. Job candidates are often evaluated by a panel of interviewers who have differing needs and opinions, and whose feelings must be distilled into actionable decisions. For example: While a firm may be hiring for a role in one division, they may find another candidate who is better-matched for in a related division. Who has prioritized access to the candidate? Can a new position be created that combines both divisions? If so, what is the career path in this hybrid position, and what happens to the previous openings – does the hybrid job replace either or both earlier requisitions? Settling these questions may require discussion, persuasion and trading favors between divisions and/or members of the hiring panel.

In making an interviewing decision, a screener must put himself or herself into all of the shoes of many potentially affected parties – both those involved in the final job placement, as well as those involved in the hiring process. In addition, the screener must put himself or herself in the mind of the candidate: Will the candidate already have another job offer that he/she will like more? Will the candidate want the job after learning more details? How will the candidate react to peculiarities of pay, coworkers and procedures?

For these reasons, the O*NET occupation "human resource specialists" *also* ranks highly on all four O*NET measures of social skills (coordination, negotiation, persuasion and social perceptiveness). The actions automated in this paper are the decisions to to interview (or reject) candidates. This is only one part of the full hiring process. However, the initial decision is tied to the later outcomes through incentives: In this industry, HR specialists are awarded substantial performance incentives for selecting a candidate who passes screening. For this reason, the screeners (and their automated replacements) must be able to anticipate the social aspects of later screening and performance outcomes.

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.

The application process for jobs in this market proceeds as follows. First, candidates apply to the company through a website.[10] Next, a human screener reviews the applications of the candidate. This paper includes a field experiment in replacing these decisions with an algorithm.

[9] The exact categorization of these jobs in the O*NET database requires some interpretation. Based on title alone, the most similar occupation is "Software Developers, Applications." On these four measure, is at the median for three and slightly below for the fourth (social perceptiveness). However, the job description in O*NET for this occupation leaves out the design and coordination aspects of the job. These aspects are better captured in the occupations labeled "Computer and Information Systems Managers" and "Information Technology Project Managers," both of which rate highly on all four measures of social skills. Like the jobs in this paper, the "management" expressed in these latter O*NET occupations does not necessarily involve direct command authority over subordinates and often refer to managing processes through coordination.

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

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 aptitute, 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.

The setting from this study is a single firm with several products and services. The sample in this paper is only for one job opening, and for one geographic location where there is an office.[11] 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 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.

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 the "Top 1%-2%" – 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.

2.1 Selection Algorithm

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

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

application volumes for posted jobs throughout the economy. First, the digitization of job applications has lowered the marginal cost of applying. Second, the Great Recession caused a greater number of applicants to be looking for work. On the demand side, recent information technology improvements have enabled firms to handle the volume of online applications. Firms are motivated to adopt these algorithms in part of the volume/costs, and also because of the address potential mistakes in the judgements of human screeners.

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?"[12]

About two-thirds (67.5%) answered "yes."[13] 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[14] 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]."[15] 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.

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 article[16] about resume-filtering technology:[17] "[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.[18] 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

[12] 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).

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

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

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

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

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

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

were created to represent how many times each stem was used on each resume.[19]

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.[20] 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.

A few observations about the algorithm. First, the algorithm introduced no new data into the decision-making process. In theory, all of the covariates described above can also be observed by human resume screeners. The human screeners could also view an extensive list of historical outcomes on candidates. 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.

Second: This modeling approach imposes no constraints on the job applicant's message space. 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. As with a spoken conversation with a hiring manager, this algorithm did not impose constraints on what mix of information, persuasion, framing and presentation a candidate could use in her presentation of self.

By contrast, other "automated" job screening interventions drastically limit the candidate's message space. For example, the variables introduced in the screening algoritm studied by Hoffman, Kahn and Li (2016) are responses to human-designed, multiple-choice survey instrument which

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

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

are given weights by an algorithm. Hoffman, Kahn and Li (2016) provide convincing evidence that these tests can be very valuable to the employer. However they speak to a different research question than the subject of this paper for several reasons.

In these surveys, the survey questions are designed by humans. Additional information is available to the algorithm exists because a human – not a machine – decided to solicit this information from the candidate. A large part of the benefit an algorithm in this context may come from the fact that an experienced organizational psychologist knew to add a particular question to the survey. The success or failure of these applications may owe more to insightful human survey design than machine social skills.

The multiple-choice format vastly constrains the candidate's message space. This contrasts with normal human interaction. The communication style evolved by humans over millions of years of evolution is not constrained by multiple choice answers. This feature destroys the analogy to human social skills and makes the work of the computer much easier. In addition: The constrained format of the responses are, again, *also* designed by a human survey designer, like the questions themselves. The benefit of reduced message space should also be attributed to deliberate human design.

Third: The counterfactual human recruiters in these studies do not have the benefit of additional information nor the simplified message space. The "treatment" is a combination of new information, reduced message space and reweighing of information – of which only the latter was provided entirely by a machine. By contrast, the application in this paper provides a much cleaner comparison of human and machine judgment based on common inputs. For both sides of the experiment, the input is a text document with an enormous potential message space. The only human curation has been performed by the candidate – who acts adversarially to screening, rather than in cooperation with it. The performance improvement from digitization in this context comes entirely from reweighing information that humans are able to see.

Lastly: Although the algorithm in this paper is computationally sophisticated, it is econometrically naive. The designers were not interested in interpreting the model causally. Similarly, the algorithm designers ignored the two-stage, selected nature of the historical screening process. Candidates in the training data are first chosen for interviews and then need to pass the interviews. If historical screeners selected the wrong candidates for interviews, this would lead to biased estimates of the relationship between characteristics and success. In economics, these issues were raised in Heckman (1979), but the programmers in this setting did not integrate these ideas into its algorithms.

3 Potential Outcomes Framework for Screening Experiment

How does one the effectiveness of one screening method (such as machine learning) compare to another (such as human evaluation)? In this section, I present a potential outcomes framework (Neyman, 1923/1990; Rubin, 1974, 2005) for answering this question generically. I then apply this framework into my setting to obtain causal estimates.

Many firms screen job candidates using a test such as a job interview or a skills assessment.

Candidates in these settings face multiple stages of screening: They must be selected for a test, and then pass the test in order to become employed.[21]

Because testing is expensive, the firm must target testing to candidates most likely to pass. The econometric setup below helps measure the effects of changing criteria for testing. This can be used by firms and researchers to study tradeoffs between the quantity and average yield of testing criteria.[22] This procedure takes the test as given, and evaluates criteria for selecting whom to test.[23]

The method can be applied in non-hiring settings. For example, doctors may want to administer costly tests – but only to patients who are likely to have a particular illness. Alternatively, police may want to spend investigative resources to evaluate ("test") allegations of criminality, but only in cases likely to uncover actual crime. College admissions officers may want to offer interviews, but only for applicants likely to pass (or likely to accept offers if extended). Venture capitalists may want to interview companies, but only those most likely to succeed.[24]

These settings feature similar testing and selection tradeoffs where this approach can be applied. For the exposition below, I will use generic testing language wherever possible and use hiring examples for clarification.

First I introduce notation. Each observation is "candidate" for testing, indexed by i. For a job candidate i, the potential hiring outcomes are a) working for this employer, b) working for another company or c) being unemployed.

However from the employer's perspective, the relevant counterfactuals for i are 1) hiring candidate i, 2) hiring someone else or 3) leaving the position unfilled. Because this paper is about employers' selection strategy, the statistics below will focus on the latter set of potential outcomes (the employer's) and not the candidate's (except where they interact).

Because this empirical strategy is oriented around the firm, I will at times code outcomes as zero for candidates who were rejected, who work elsewhere or who produce no output. The endogeneity of testing decisions will be addressed using a field experiment or instrument.

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).[25]

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

[21] In many settings, remaining employed or earning promotions or raises requires a similar process.

[22] For example, some firms may want to use a criteria that maximizes average test yields, conditional on a fixed budget of N tests. Alternatively, other firms may prefer to relax the N budget, and instead maximize the sum of total test yield – possibly at the expense of the average yield. In either case, the procedure below helps quantify the tradeoffs between testing criteria.

[23] I do not address whether the test itself is optimal.

[24] As I discuss later, the test itself is a form of "criteria" for employment. The procedure described here can be iteratively applied up and down the production function to select an optional hiring and/or promotion criteria (rather than testing criteria) from among several discrete alternatives.

[25] Binary outcomes is used to simplifies exposition. In addition, this maps to some of the outcomes examined in this paper, such as whether candidates passed an interview or not. However, this procedure can be used if the testing outcomes are non-binary or continuous. I discuss this later in this section.

a single firm's private evaluation, which may not necessarily be correlated with the wider labor market's assessment.[26]

For each candidate i, the econometrician 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.

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.[27] No assumptions about the distribution of θ are required, nor are assumptions about the correlation between A, B and θ.

For any given candidate, $A_i = 1$ means that Criteria A suggests testing candidate i and $A_i = 0$ means Criteria A suggests *not* testing 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 may provide a useful visualization. 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.[28]

However, often a researcher does not have full data about how each candidate fares in criteria A vs B. The act of testing an B candidate – for example, by scheduling a test – may automatically make the candidate unavailable for evaluation by A. The strategy for addressing this problem uses exogenous variation induced through a field experiment to make a causal inference.

In this setup, the potential outcomes framework below measures whether a counterfactual Criteria B changes the yield on tested candidates compared to A. Later, I will extend this framework to address related questions, such as offer acceptance rates, on-the-job performance and other downstream outcomes.

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 the success rate of $A \cap B$ candidates (candidates that both criteria approva), for all A candidates and for $A \setminus B$ candidates (ones that A approves and B doesn't). Then I will compare these test rates to make an inferences the effects of using A vs B.

[26] 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.

[27] 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. 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.

[28] Unless there is a SUTVA-violating interaction between candidates in testing outcomes.

To estimate $E[Y|T = 1, A = 0, B = 1]$, the pass rate of candidates who would be rejected by Criteria A, but tested by Criteria B, it isn't sufficient to test all B candidates or a random sample of B candidates. Some of the B candidates are also A candidates. The econometrician needs an instrument, Z_i, for decisions to test that is uncorrelated with A_i. Because the status quo selects only A candidates, the effect of the instrument is to select candidates who would otherwise not be tested.

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, for all candidates for whom $B_i = 1$. The instrument must affect who is interviewed – for exposition, assume that firm tests all candidates for whom $Z_i = 1$, irrespective of A_i.[29]

In order to measure the marginal yield of Criteria B, we need variation in Z_i within $B_i = 1$.[30] The instrument Z_i within $B_i = 1$ is "local" in that that it only varies for candidates approved by Criteria B. Z_i identifies a local average testing yield for Criteria B.

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 five conditions. I outline each condition in theory below, with some discussion of the implications in a hiring or testing setting. In the following section, I show that each condition is met for my specific empirical setting.

1. **SUTVA**: Candidate i's outcome depends only upon his treatment status, and not anyone else's. This permits us to write $T_i(Z) = D_i(Z_i)$ and $Y_i(Z_i, D(Z)) = Y_i(Z_i, D_i(Z_i))$.

   In a testing setup, this assumption might be problematic if candidates are graded on a "curve" or relative ranking, rather than against an absolute standard.[31] It would also be problematic if the firm (or candidates) in question were powerful enough in the labor market to create general equilibrium effects through the testing of specific candidates.

2. **Ignorable assignment of Z**. Z_i must be randomly assigned, or $0 < \Pr(Z_i = 1|X_i = x) = \Pr(Z_j = 1|X_j = x) < 1, \forall i, j, x$.

3. **Exclusion restriction**, or $Y(Z, T) = Y(Z', T), \forall Z, Z', T$. The instrument only affects the outcome through the decision to administer the test. For a given value of T_i, the value of Z_i must not affect the outcome.

   In a testing setting, one implication of the exclusion restriction is that the test must be graded fairly, so that the resulting pass/fail out are not biased to reflect the grader's preferences for

[29] These characteristics are true for this paper, but these assumptions can be relaxed to be more general.

[30] 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 $B = 0$.

[31] If candidates were graded by relative ranking, SUTVA would be violated when one candidate's strong performance adversely affects another's chances of passing.

Criteria A vs B. Biased test grading would violate the exclusion restriction.[32] Double-blind or objective evaluation may help meet the exclusion restriction.

A satisfied exclusion restriction lets us write $Y(Z,T)$ as $Y(T)$. Assumption 1 lets us write $Y_i(T)$ as $Y_i(T_i)$.

4. **Inclusion restriction**. The instrument must have a non-zero effect on who is tested ($E[T_i(1)T_i(0)|X_i] \neq 0$, or $Cov(Z,T|X) \neq 0$).
5. **Monotonicity**, or $T_i(1) \geq T_i(0)$ or $T_i(1) \leq T_i(0), \forall i$. This condition requires there to be no "defiers," for whom testing is less likely if the instrument is zero.

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

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

This setup is isomorphic to instrumental variables (Angrist et al., 1996), and the value above can be estimated through two-stage least-squares. The units of the estimand are *new successful tests per new administered tests* (success rate). 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.

This empirical framework is parallel to causal inference using instrumental variables. The key to the correspondence is that the outcome "caused" by the test is the *revelation* of an average value for θ over the tested candidates, so that the firm can at upon the new information. Importantly, the test itself does not cause θ to *change* for any candidate.

Next, I show how the IV conditions are met in my empirical setting. Then, I show how to extend this framework further into the production function to measure the effects on other downstream outcomes beyond early stage testing acceptance.

3.1 Application to my Empirical Setting via Field Experiment

How does the machine learning algorithm in my empirical setting compare to the status-quo human screening process ? In this paper, the firm performed a field experiment that allows a researcher to cleanly compare these methods using the framework above.

[32] In many instances, test graders may have a preference for what which criteria are used. In the example above: Suppose the test grader was biased against the CEO's opinion (Criteria A) and wanted the evaluation to look poorly for the CEO. Such a grader he/she may CEO-approved candidates if he/she knew them, violating the exclusion restriction.

In my empirical setting, all incoming applications (about 40K candidates) were scored and ranked by the algorithm. Candidates with an estimated probability of 10% (or greater) of getting a job offer were flagged as "machine approved."[33] This group comprised about 800 applicants over roughly one year. [34]

The field experiment worked by generating a random binary variable Z for all machine-picked candidates (one or zero with 50% probability). Candidates who draw a one are automatically given an interview. Those who drew a zero – along with the non-machine approved candidates – are left to be judged by the status quo human process. The human evaluators were thus given access to a random half of machine-approved candidates, so that they could be independently evaluated along with those rejected by the screening algorithm.

The humans were not told how the machine evaluated each candidate – they were not told about the existence of the machine screening and had no choice than to evaluate the candidates independently.

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 both methods were sent blindly into an interview process. Neither the interviewers nor the candidates were told who (if anyone) came from an algorithmic process or a human-generated one. The experiment was not disclosed to the interviewers.

The IV conditions necessary to apply this framework are met as follows:

1. **SUTVA**: 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."[35]

2. **Ignorable assignment of Z**. Covariate balance tests in Table 2 appear to validate the randomization. The randomization was performed by the employer.

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.

[33] 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.

[34] 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.

[35] 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.

4. **Inclusion restriction**. The instrument must have a non-zero effect on who is tested ($E[T_i(1)T_i(0)|X_i] \neq 0$, or $Cov(Z, T|X) \neq 0$). In my empirical setting, this clearly applies. 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).

One characteristic of this approach is that the two methods are not required to 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.

Firms that want to fix the quantities can run quantity-limiting experiments. Alternatively: If one Criteria is based on a rankable variable, a researcher can examine subsets of the data that limit analysis to the top N candidates selected by either mechanism.

In my empirical setting, the machine learning algorithm identified 800 candidates, and the human screeners identified a far larger number (XXX). It's possible that the higher success rate is the result of extending offers to fewer, higher quality people. In order 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 tests available to both mechanisms to measure differences in yield, conditional on an identical "budget" of tests.

3.2 Offer Accepts, On-the-job productivity and other "downstream" outcomes

As I mention earlier, the empirical framework above is isomorphic to causal inference using instrumental variables. The key to the isomorphism is that the outcome "caused" by the test is the *revelation* of an average value for θ over the newly tested candidates, so that the firm can act upon the new information. Each candidate's θ does not change.

In some cases, the firm's ability to act upon the new information may be limited. For example: Candidates who pass the test may be extended a job offer, but some might reject the offer for another opportunity. A firm's evaluation of A vs B may depend on not only passthrough rates, but also offer acceptances.

It's possible that a new interviewing criteria identifies candidates who pass, but do *not* accept offers. For example: Testing candidates with elite degrees may be a good way to find test-passers. However, these candidates may have many other competing offers and exhibit a low yield on extended offers.

The benefits of successful first-stage screening could be wiped out by an ineffective second stage. Some firms screen candidates first with a written test and then with interviews. The test may help identify exceptional candidates. However, if interviewers reject those candidates (perhaps wrongly), then the net effect of the improved screening could be zero or negative.

For these situations, a researcher may care about the net effects of a change in early-stage screening criteria. For this purpose, one 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 Y, which only measures if the test was passed.

For a change in outcome variable to Y', 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 a "offer accepts" / "offer extends" marginal coefficient.

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' whose value is "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.[36]

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, 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.

3.2.1 Revisiting IV Assumptions

Introducing a new downstream outcome (Y') and endogenous variables (T') require revisiting the IV assumptions. Even if the IV requirements were met for Y and T (the original variables), this does not automatically mean the IV requirements are met for our second endogenous variable (T') and the downstream outcome (Y').

All IV assumptions must be revisited. Below, I mention a few particular areas where the IV criteria may fail for downstream outcomes in a testing or hiring setting – even if they are first met in upstream ones.

SUTVA. In my empirical setting, there are no cross-candidate comparisons ("grading on a curve") necessary to pass the test; if they were, it would introduce SUTVA violations.

However, even if cross-candidate comparisons were absent from test-grading, they might reappear downstream in offer-acceptances. If an employer has a finite, inelastic number of "slots"

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

(Lazear et al., 2016), then test-passers' acceptance decisions could interact with each other. A candidate who accepts a spot early may block a later one from being able to accept, creating a SUTVA violation.

This does *not* happen in my empirical setting, where the employer wants to hire as many people as could pass the test and does not have a finite quota of offers or slots.[37] I raise this issue as an example of how downstream SUTVA requirements can fail, even if they pass upstream.

Random/ignorable assignment of Z. An instrument orthogonal to early testing outcomes is not necessarily orthogonal to downstream testing decisions. In the setting of this paper and for experiments more generally.

Inclusion restriction (instrument strength). An instrument Z that has a strong effect on which candidates are tested is not necessarily a strong effect on which candidates are hired. Z could be a much weaker instrument for a downstream T' than for the earlier T. This is partly because there are fewer candidates who passed T and were eligible to take T' – effectively there is a smaller sample size.

3.3 Comparison of this Method to others in Literature

As Oyer et al. (2011) discuss, field experiments varying hiring criteria are relatively rare ("What manager, after all, would allow an academic economist to experiment with the firms screening, interviewing or hiring decisions?"). The goal of the above sections has been to lay out some simple econometric theory for designing hiring experiments (or otherwise making causal inferences about hiring outcomes from observational data).

In this section, I contrast the approach above to those used by other fields studying personell assessment. My experiment allows me to compare my experimental estimates to those obtained by other methods – including methods advocated by government policymakers – and thus quantify the bias in these alternatives (for my setting). I also evaluate the assumptions behind these methodologies empirically.

The field of industrial and organizational psychology ("I/O psychology") has developed an influential literature on personnel assessment and hiring criteria. Among other things, this literature examines statistical relationships between on-the-job performance and observable characteristics at the time of hiring.

Although it began with psychological characteristics, this field has grown to encompass many forms of employment testing (including measure that aren't particularly psychological). This strand of research – and its methodological recommendations – have been very influential in law and public policy surrounding employment selection mechanisms.[38]

[37] It is possible that SUTVA violations may arise if multiple test-passers were to make a single group decision about where to work together (or apart) as a group. For example, if Candidates i and j wanted to join the same firm and made decisions together, this could violate SUTVA. The candidates in this study applied individually to the employer via an online job application."Joint" offer-acceptance decisions are more common in merger or acquisition settings. It is impossible to know if this is happening in this dataset, but the author inquired with the recruiting staff if they knew of any "joint" offer acceptance decisions in this sample. The recruiters reported no known instances.

[38] For example, the Uniform Guidelines on Employee Selection Procedures ("UGESP"), was adopted in 1978 by the

The I/O psychology literature about hiring criteria often ignores the sample selection issues discussed in (Heckman, 1979). As I show below, ignoring sample selection bias is also strongly advocated in the government evaluation guidelines surrounding employment testing in the United States.

Most I/O psych papers – and human resource practitioners following government guidance – have a dataset containing performance outcomes only on hired workers, without experimental variation in who is hired.[39] Because of this non-random sampling, the coefficients in these papers are biased estimates of the underlying population relationships. The direction of the bias cannot be signed, and the true correlation may not even share the same direction as in the selected sample.

In some cases, researchers have justified the above approach using an assumption of linearity or monotonicity, but this assumption is rarely empirically tested or made explicit. These issues lead to mistakes about the costs and benefits of adopting various employee selection criteria.

The experiment in this paper allow me to compare the estimates of analysis of the selected sample, compare it to the experimental estimates and measure the extend of the bias. I can also empirically evaluate the assumption of linearity and monotonicity.

Using my notation, the I/O psych approach compares performance outcomes between $A \cap B$ candidates to $A \setminus B$. In other words, it compares candidates whom both the status quo *and* the new method select versus candidates approved only by the status quo, but not the new method. All candidates in the sample must have passed A in order to have performance outcomes.

Civil Service Commission, the Department of Labor, the Department of Justice, and the Equal Opportunity Commission in part to enforce the anti-employment discrimination sections of the 1964 Civil Rights Act. The UGESP creates a set of uniform standards for employers throughout the economy around personnel selection procedures from the perspective of federal enforcement. The UGESP are not legislation or law; however, they provide highly influential guidance to the above enforcement agencies and been cited with deference in numerous judicial decisions.

These guidelines extensively reference and justifies itself using the standards of academic psychology. For example, the UGESP requires that assessment tests that are "consistent with professional standards," and offers "the A.P.A. Standards" (an American Psychological Association book called Standards for Educational and Psychological Testing (2014)) as the embodiment of professional standards. No other profession or academic discipline is referenced at all in the UGESP, including economics.

The UGESP were adopted in 1978 and contains extensive statistical commentary about hiring criteria. Since 1978, statistical practice in a number of social science fields has changed substantially (Angrist and Pischke, 2010). However, the UGESP have not been substantially revised and are still in use today. They can be accessed at `https://www.gpo.gov/fdsys/pkg/CFR-2014-title29-vol4/xml/CFR-2014-title29-vol4-part1607.xml` (last accessed December 5, 2016).

[39] For example: A famous psychology paper (Dawes, 1971) shows that for psychology graduate students, a simple linear model more accurately predicts academic success than professors' ratings. In a followup paper, Dawes (1979) showed this result held, even when the linear predictor was misspecified.

Dawes interpreted this finding was to mean that linear predictors should be used in the graduate students' admissions. However McCauley (1991) showed that two decades after Dawes' finding, linear predictors were still not often used often not graduate student selection for PhD programs. This author's casual survey indicates this practice is still rare still rare in academic psychology as of this writing, but is gaining in popularity in the corporate world as discussed in Section 2.1.

A closer reading of Dawes (1971) shows the sample consists only of *matriculated* graduate students at one University. For reasons studied in Heckman (1979), Dawes' correlations within a selected sample may not generalize to the entire applicant pool. The direction of the bias cannot be signed, and the true correlation may not even have the same direction as in the selected sample.

An experiment would be necessary to measure the causal effect of changing selection criteria on ultimate graduate student achievement. Despite the popularity of Dawes' 1971 finding, no one to date has performed this experiment, or attempted to address the potential bias via another source of identification.

There are several reasons this would produce a biased estimate of the effect of shifting hiring policies. A shift from policy A to B would both i) deny tests to some applicants who were formerly getting tests under policy A and no longer do so under B, as well as ii) grant tests to new candidates who would be tested under policy B but who weren't tested under policy A.

Data about the latter set of outcomes (ii) is not identified within purely observational data, and thus a quasi-experimental strategy for identification would be needed.[40]

Most I/O psychology papers lack the necessary experimental variation. In addition, the government guidelines flowing from this literature also encourage ignoring sample selection bias. The Uniform Guidelines on Employee Selection Procedures ("UGESP"), a set of policy guidelines for enforcing federal non-discrimination issues introduced in Footnote 38. These guidelines offer technical definitions of how to evaluate selection procedures, such as the requirement that the "relationship between performance on the [job] procedure and performance on the criterion measure [test] is statistically significant at the 0.05 level of significance[.]"

This has been interpreted by courts, prosecutors and practitioners to mean correlations within the set of historically hired workers – a comparison of A to $A \cap B$ that ignores $B \setminus A$. The UGESP specifically clarifies it does not require an experiment, or another intervention that would sample from outside the firm's status quo, in order to evaluate a particular hiring method: "These guidelines do not require a user to hire or promote persons for the purpose of making it possible to conduct a criterion-related study." (Section 14B).

Without an experimental variation, a firm (or regulator) would be unable to evaluate one of the critical effects of shifting to B: The new candidates who would be admitted by B but not by A. Such candidates would not appear in the historical data at all. Without an intervention – the kind the UGESP says is unnecessary – there is no way to evaluate such candidates.

By contrast the candidates selected both by A and B – those who play a central role in the evaluation of the UGESP and the I/O psych literature – are irrelevant to evaluating a potential shift of A vs B, because such candidates would be admitted under both policies.

Ignoring selection could result in either over- or under- estimate the true effect. As such, the method advocated in the UGESP and I/O psych papers does not deliver a useful boundary condition such as an upper- or lower- bound.

Overestimation: The $A \cap B$ candidates have passed both Criteria. If both Criteria A and B have some merit, then the $A \cap B$ candidates contains "superstar" who were able to pass both standards. Comparing these superstars to candidates who only passed one test (A) may thus overestimate the benefit of switching to B; much of the effect of switching policies would come from selecting B candidates who only passed one test. A fairer approach (advocated in this paper) is to make comparisons between candidates who passed exactly one test. However, this method would require testing (or hiring) candidates who wouldn't otherwise be tested.

Underestimation: It's also possible that the I/O psych method could understate, rather than overstate, the benefit of a new method. Suppose that Criterial B identifies lots of high-performing

[40] Besides an experiment, one way to avoid this issues is to test all applicants. Within the economics literature, Pallais and Sands (2016) used the strategy of hiring all applicants to a job opening in her study of referrals in hiring for routine cognitive tasks (basic computations and data entry) on oDesk.

candidates who were previously not identified at all. In this case, most of the benefit of adopting B would come from these new candidates. Candidates who passed both requirements (A and B) may not perform as well as purely B-only ones. This would mean that the I/O psych analysis understates the benefit of B.

In my experiment, I find that the I/O psych both over- and under- estimates the true effects, depending on the outcome in question. Results are summarized in Table **??**. For interview performance, I find that the I/O psych leads to overstatements of the benefits of the machine learning algorithm by ×1.5 (or +10%).

However, the I/O psych method *understates* the benefits of machine learning on offer acceptance rates – and in fact, reports the wrong *sign* of the effect. In Table **??**, I test and reject the monotonicity and linearity assumptions.

I elaborate on these empirical results in Section **??**; I preview them here to motivate and differentiate the empirical approach.

Other studies have recently made causal inferences about the effects of hiring policies. In particular, Autor and Scarborough (2008), Hoffman, Kahn and Li (2016) and Horton (2013). Although these studies have not laid out a potential outcomes framework, they can be re-interpreted in light of the above. I do this in Section **??**.

4 Results

In Table 2 Panel B, I report the performance of treatment and control groups in the hiring process. The first result is substantial disagreement between machine and human judgement. The machine and humans agree on roughly 50% of candidates, and disagree on 30%. The remaining ≈20% withdrew their applications prior to the choice to interview. Roughly 30% of candidates in the experiment were approved by the machine, but counterfactually disapproved by humans. The most common reason cited for the human rejection in this group is lack of qualifications. Separate regressions show that the machine appears more generous towards candidates without work experience and candidates coming from rare educational backgrounds.

Table 2, Panel B also shows that many of the candidates in the treatment group succeed in subsequent rounds of interviews. The yield of candidates is about 8% higher in the treatment group. Table 2, Panel C assesses whether machine picked candidates are more likely to pass subsequent rounds of screening conditional on being picked. The conditional success rates are generally higher for the treatment group, but not statistically significant.

Table 3 examines these differences as regressions and adds controls, which tighten the standard errors on the differences. These results show that the average candidate in the treatment group is more likely to pass the interview than a candidate selected by a human screener from the same applicant pool. It also shows that the machine candidates are more likely to accept an offer.

In Panel C of Table 3, I examine marginal success rates of machine's candidates using instrumental variables. Here, I use the experiment as an instrument for which candidates are interviewed (or given an offer). The marginal candidate passes interviews in 37% of cases – about 17% more than the average success rate in the control group. 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 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 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.[41] These peer reviews cover both the logical structure, formatting and readability of the code as outlined in company guidelines.[42] 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.

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

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

6 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 2% 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.

7 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.

8 Conclusion

The idea that a computer could have better "social skills" than a human may sound counterintuitive. Autor's 2015 discussion of the future of automation mentions social skills as the type of work that can't easily automated because we humans our don't consciously "know the rules" and thus cannot program a machine to perform this work. However, Autor (2015) specifically mentions machine learning – the intervention examined in this paper – as one of two methods for which these tasks could one day be automated.

The idea that a computer could have better "social skills" than a human may still sound counterintuitive. However, a precedent already exists. There is no reason that the finest interlocutor of the emotions of a species must be another member of the same species. Many domesticated animals can be trained to perform tasks (such as sitting on command, fetching newspapers, racing at high speeds under control of a jockey) that don't appear in nature.

Training these animals requires a nuanced understanding of an animal's psychology, emotions and relationship to humans. And yet, the most effective dog and horse trainers are humans – not other dogs or horses. The master manipulators of canine psychology and social instincts are human beings, not other dogs.[43] When it comes to making a dog fetch a newspaper, whose emotional intelligence is better – a human's or another dog's? In the same way, computers may be equal or better than humans at the nuanced social judgements necessary to select and persuade humans.

One possible caveat to the strong performance of the algorithm in this paper is that the algorithm had to be "trained" to model historical human decisions. Without the historical human decision, the algorithm's performance would not have been possible. However, the algorithm could have been easily trained using bandit methods (rather than with historical data). All the firm would have wasted was some possibly bad interviews. This might have been a more efficient, less biased way to train the algorithm if one were starting from scratch.

The results of this study should not be over-interpreted. Despite the positive performance of machines in the selection task, computers are currently inept at many tasks requiring emotional intelligence such as therapy, sales or customer support. As cited by Deming (2015), computer scientists have yet to create a chat robot that can successfully pose as a human (the "Turing Test.")

However, even the results above should be viewed as a lower bound of quality. The statistical modeling approach in this paper is "naive," underutilizing decades-old techniques from other disciplines (such as Heckman's 1979 sample-selection methods or Robbins et al.'s 1952 multi-armed bandit techniques[44]) that could plausibly improve performance. More generally, computational power and input data are increasing over time, and firms continue to invest in new ways to automate social skills. At the time of this writing, several technology companies (Apple, Facebook, Yahoo! Google and others) recently announced large investments in "chatbot" digital assistants with whom users can converse as one would with other humans.

Future research (including a future version of this paper) should better examine where subjective decisionmaking tasks fit into the labor market. One source of data about jobs that require subjective

[43] As with analogy with humans/computers, there may be some agency problems. IE, perhaps dogs have a superior innate ability to persuade other dogs to fetch newspapers, but we haven't found the proper contract.

[44] Related methods in machine learning are called "active learning." For a survey, see Settles, 2010.

decision making is from the US Labor Department's O*NET database used by Autor, Deming and others. O*NET categorizes occupations based on the tasks, skills and type of work involved. These occupations can be linked to current and historical employment and wage data.

One such category in O*NET contains occupations involving "Judging the Qualities of Things, Services, or People." This category contains professions with heavy subjective decisionmaking requirements. According to O*NET, the occupation requiring the most subjective assessment is *Clinical Psychologist*. Human resource and screening professionals rank near the top. The occupation requiring the least such is *Model* (as in fashion, art or photography).[45] This category may be a good source of data for future research about how automating subjective judgements may impact the labor market and production processes.

8.1 Addendum: Alternative Explanations to Soft-Skill Wage Growth

If social skills can indeed be mechanized, what other explanations may explain the returns to social skills cited by Deming (2015) and Weinberger (2014)?

The technology is still new. Most data in the "social skills" literature describes changes over the past 40+ years. For example, Deming (2015) studes post-1980 America and Weinberger (2014) studies data since 1972. The modern renaissance in big data and machine learning has only come in the last decade.

Legal artifacts. Should machine "social skills" exceed humans' in quality, there may be other barriers to adoption. Pre-existing legal and social institutions are oriented around human, rather than machine, discretion. For example: One of the key obstacles to self-driving cars is how regulators, courts, insurance companies and the police will treat this new technology.

In the space of hiring, several precedents thwart wider adoption of algorithmic screening methods. One particularly interesting case is *Wal-Mart vs Dukes* (1990), a case which (perhaps inadvertently) created different liability standards for human and machine decisions. *Wal-Mart vs Dukes* (1990) was a class-action discrimination case in which 1.6 million female WalMart employees (current and former) claimed discrimination, citing statistics about pay and promotion rates. The court ruled in Wal-Mart's favor, stating that the women did not have enough in common to establish a class.

The Court's reason was that Wal-Mart's internal personnel policies were delegated (Dessein, 2002) to local managers who enjoyed substantial discretion over pay and promotion to local managers. Writing for the majority, Justice Antonin Scalia claimed that "On its face, [Wal-Mart's decentralization policy] is just the opposite of a uniform employment practice that would provide the commonality needed for a class action. It is a policy against having uniform employment practices." For the women suing Wal-Mart, Scalia wrote there is no "common answer to the crucial question *why was I disfavored?*"

Through the lens of *Wal-Mart vs Dukes* (1990) (and related laws and precedents), automation is a form of centralization that increases a firm's legal vulnerability. Similarly, a firm's hiring algorithm

[45] A separate O*NET category, "Evaluating Information to Determine Compliance with Standards," involves occupations requiring more "objective" judgments.

can be dissected and scruitinizedin ways that a human decision cannot be. Some firms may not desire the transparency that an algorithm would introduce.

"Taste-based discrimination" against machines. To some observers, machines performing human tasks is repugnant (Roth, 2007). In psychology, a small literature documents "algorithm aversion" (Dietvorst et al., 2015a,b), or reluctance to delegate decisions to algorithms even when shown evidence of the machine's superior skill.

Table X contributes additional evidence for this hypothesis with a survey of 1500 Internet users, weighted to match the demographics of the American Community Survey. Survey respondents were randomly allocated into a 2×2 design and asked how they would feel if he/she applied for a job – and were offered an interview (or not) and the decision was made by a human (or an algorithm). Respondents were given a 7-point Likert scale for responses. One might speculate that job applicants would care only about the application's outcome (rejection vs interview) than the decision process (human vs machine). Nonetheless, respondents preferred being evaluated by a human, both in rejection and acceptance.

References

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek, "When does coordination require centralization?," *The American economic review*, 2008, *98* (1), 145–179.

Angrist, Joshua D and Jörn-Steffen Pischke, "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *The Journal of Economic Perspectives*, 2010, *24* (2), 3–30.

_ **, 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.

Association, American Educational Research, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (U.S.), *Standards for Educational and Psychological Testing*, American Psychological Association, 2014.

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, "Why are there still so many jobs? The History and Future of Workplace Automation," *The Journal of Economic Perspectives*, 2015, *29* (3), 3–30.

_ **and David Scarborough**, "Does job testing harm minority workers? Evidence from retail establishments," *The Quarterly Journal of Economics*, 2008, pp. 219–277.

Baron-Cohen, Simon, "Theory of mind and autism: A fifteen year review.," 2000.

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.

Burks, Stephen V, Bo Cowgill, Mitchell Hoffman, and Michael Housman, "The value of hiring through employee referrals," *The Quarterly Journal of Economics*, 2015, p. qjv010.

Camerer, Colin, George Loewenstein, and Drazen Prelec, "Neuroeconomics: How neuroscience can inform economics," *Journal of economic Literature*, 2005, pp. 9–64.

Commission, Equal Employment Opportunity et al., "Uniform Guidelines on Employee Selection Procedures," *Federal register*, 1978, *43* (166), 38295–38309.

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

Dawes, Robyn M, "A case study of graduate admissions: Application of three principles of human decision making.," *American psychologist*, 1971, *26* (2), 180.

_ , "The robust beauty of improper linear models in decision making.," *American psychologist*, 1979, *34* (7), 571.

Deming, David J, “The growing importance of social skills in the labor market,” Technical Report, National Bureau of Economic Research 2015.

Dessein, Wouter, “Authority and communication in organizations,” *The Review of Economic Studies*, 2002, *69* (4), 811–838.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, “Algorithm aversion: People erroneously avoid algorithms after seeing them err.,” *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

_ , _ , and _ , “Overcoming Algorithm Aversion: People Will Use Algorithms If They Can (Even Slightly) Modify Them,” *Available at SSRN 2616787*, 2015.

Eichenwald, Kurt, “Microsoft’s Lost Decade,” *Vanity Fair*, 2012.

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

Frey, Carl Benedikt and Michael A Osborne, “The future of employment: how susceptible are jobs to computerisation,” 2013.

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.

Gubler, Timothy, Ian Larkin, and Lamar Pierce, “The dirty laundry of employee award programs: Evidence from the field,” *Harvard Business School NOM Unit Working Paper*, 2013, (13-069).

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

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

Horton, John J, “The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment,” *Forthcoming, Journal of Labor Economics*, 2013.

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

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

Kahneman, Daniel, *Thinking, fast and slow*, Macmillan, 2011.

LaCurts, Katrina, “Criticisms of the turing test and why you should ignore (most of) them,” *Official Blog of MITs Course: Philosophy and Theoretical Computer Science*, 2011.

Lazear, Edward P, 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.

McCauley, Clark, "Selection of National Science Foundation Graduate Fellows: A case study of psychologists failing to apply what they know about decision making.," *American Psychologist*, 1991, *46* (12), 1287.

Milgrom, Paul and John Roberts, "An economic approach to influence activities in organizations," *American Journal of sociology*, 1988, pp. S154–S179.

Milgrom, Paul R, "Employment contracts, influence activities, and efficient organization design," *The Journal of Political Economy*, 1988, pp. 42–60.

Moravec, Hans, *Mind children: The future of robot and human intelligence*, Harvard University Press, 1988.

Neyman, Jerzy S, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Statistical Science*, 1923/1990, *5* (4), 465–472.

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

Pallais, Amanda and Emily Glassberg Sands, "Why the Referential Treatment? Evidence from Field Experiments on Referrals," *Journal of Political Economy*, 2016, *124* (6), 1793–1828.

Parr, Lisa A, "The evolution of face processing in primates," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2011, *366* (1571), 1764–1777.

Premack, David and Guy Woodruff, "Does the chimpanzee have a theory of mind?," *Behavioral and brain sciences*, 1978, *1* (04), 515–526.

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

Rao, Justin M and David H Reiley, "The economics of spam," *The Journal of Economic Perspectives*, 2012, *26* (3), 87–110.

Robbins, Herbert et al., "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, 1952, *58* (5), 527–535.

Roth, Al, "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, 2007, *21* (3), 37–58.

Rubin, Donald B, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 1974, *66* (5), 688.

Rubin, Donald B., "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 2005, *100* (469), 322–331.

Scalia, Justice Antonin, "Wal-Mart Stores, Inc vs Dukes et al.," *131 Supreme Court*, 2011, *2541.*

Schroff, Florian, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition" 2015, pp. 815–823.

Settles, Burr, "Active learning literature survey," *University of Wisconsin, Madison*, 2010, *52* (55-66), 11.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition" 2014, pp. 1701–1708.

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

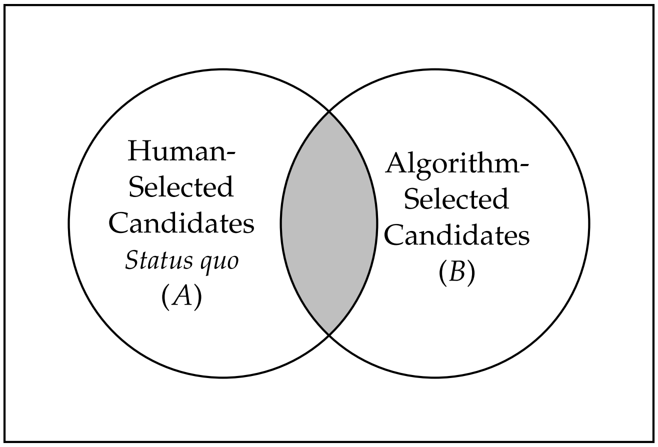
Vapnik, Vladimir Naumovich, *Estimation of dependences based on empirical data [In Russian]* 1979. English Translation by Kotz, Samuel in 1982 by publisher Springer-Verlag New York.

Waldfogel, Joel, "The deadweight loss of Christmas," *The American Economic Review*, 1993, *83* (5), 1328–1336.

Weinberger, Catherine J, "The increasing complementarity between cognitive and social skills," *Review of Economics and Statistics*, 2014, *96* (4), 849–861.

Tables and Figures

Figure 1: Caption for figure below.



Notes: The above is a useful visualization of the empirical setting. The left circle A represents candidates selected by the status quo. The right circle B represents candidates selected by the machine learning. The shaded area $A \cap B$ represents candidates accepted by both. The goal of the empirical analysis is to compare the average outcomes between the unshaded areas of A and B ($A \setminus B$ vs $B \setminus A$).

Table 1: Descriptive Statistics

Panel A: Characteristics

	Above Thresh	Below Thresh	Difference
Has Doctorate	0.280	0.066	0.214***
Ever Attended Elite School	0.576	0.211	0.365***
Ever Attended Top Tier School	0.315	0.339	-0.025**
Ever Attended Non-Selective School	0.035	0.122	-0.088***
Average Elite of all Schools Attended	0.513	0.164	0.349***
Referred	0.147	0.054	0.093***
No Work Experience (New Graduate)	0.189	0.182	0.007
Rare School	0.387	0.735	-0.348***

Panel B: Cumulative Acceptance Rates

	Above Thresh	Below Thresh	Difference
Interview	0.563	0.318	0.244***
Job Offer	0.112	0.007	0.105***
Accept Offer	0.080	0.006	0.074***

Panel C: Incremental Acceptance Rates

	Above Thresh	Below Thresh	Difference
Interview	0.563	0.318	0.244***
Job Offer	0.199	0.020	0.179***
Offer Accept	0.714	0.854	-0.140*

Notes: This table presents descriptive statistics of the sample of applicants. “Above the threshold” refers to candidates whom the machine estimated to be in the top 1-2% of applicants. “Below the threshold” candidates refer to the remaining candidates. Randomization in the experiment took place among candidates above the threshold.

Panel A contains applicant characteristics. In Panel A, “Above the threshold” includes the characteristics of both control and treatment candidates (Table 2 shows covariate balance between treatment and control).

Panel B shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel B can be read to mean, “Of all applicants who applied, X% were extended an offer.”

Panel C shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel C can be read to mean, “Of all applicants who *were interviewed*, Y% were extended an offer.”

In Panels B and C, I include only the “Above” candidates in the control group because these outcomes were affected by the experiment.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 2: Covariate Balance and Acceptance Rates (Basic Averages)

Panel A: Characteristics

	Treatment	Control	Difference
Has Doctorate	0.26	0.28	-0.02
Ever Attended Elite School	0.62	0.58	0.04
Ever Attended Top Tier School	0.28	0.31	-0.04
Ever Attended Non-Selective School	0.03	0.03	-0.00
Average Elite of all Schools Attended	0.56	0.51	0.04
Referred	0.13	0.15	-0.02
No Work Experience (New Graduate)	0.22	0.19	0.03
Rare School	0.41	0.39	0.02

Panel B: Cumulative Acceptance Rates

	Treatment	Control	Difference
Interview	0.84	0.56	0.28***
Job Offer	0.21	0.11	0.10***
Accept Offer	0.17	0.08	0.09***

Panel C: Incremental Acceptance Rates

	Treatment	Control	Difference
Interview	0.84	0.56	0.28***
Job Offer	0.25	0.20	0.05
Offer Accept	0.80	0.71	0.09

Notes: The above comparisons are of raw means. See later tables for estimated differences of averages with controls, and for IV-based estimates of marginal effects.

Panel A presents covariate balance between treatment and control groups in the "Above the threshold" applicants.

Panel B shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel B can be read to mean, "Of all applicants who applied, X% were extended an offer."

Panel C shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel C can be read to mean, "Of all applicants who *were interviewed*, Y% were extended an offer."

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 3: Average Success Rates (Regressions w/ Controls)

Panel A: Cumulative Acceptance Rates

	Interview	Job Offer	Accept Offer
Treatment	0.28***	0.095***	0.082***
	(0.032)	(0.025)	(0.023)
R^2	0.12	0.13	0.13
Observations	770	770	770

Panel B: Incremental Acceptance Rates

	Interview	Job Offer	Accept Offer
Treatment	0.30***	0.078**	0.10
	(0.033)	(0.036)	(0.077)
R^2	0.19	0.22	0.27
Observations	770	544	124

Panel C: Marginal Success Rates using Instrumental Variables

	Job Offer	Accept Offer
Interview	0.36***	
	(0.089)	
Job Offer		0.90***
		(0.12)
F-stat (1st Stage)	81.6	16.1
Mean Outcome of Control	0.20	0.71
Difference	0.16*	0.18*
Mean Outcome of Population	0.026	0.84
Difference	0.33***	0.06***
Observations	38242	38242

Notes: The above tables contain linear regressions of arriving at each stage (passing the previous) conditional on the treatment and controls. Controls include the month of the application, whether the applicant was referred and education and experience controls. Panel A shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). Panel B shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage).

Panel C shows marginal passthrough rates of the machine's additional candidates. This differs from the average succes rates, as measured in Panel A and B and in earlier tables. In the average success rate numbers above, the "Treatment" includes both the candidates both methods selected as well as those that the machine selected but the human screeners would not have. In Panel C, I use the treatment/control status to isolate the success rates of the "marginal" candidate whom the machine liked but the human screeners would have rejected. I then compare these "marginal" success rates to the incremental success rate in the control group. The results are reported in the "Difference" row. The marginal candidates pass the interview and accept offers at a higher rate than the control group.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 4: Fixing Quantitites

Panel A: Cumulative Acceptance Rates

	Treatment	Control	Difference
Interview	0.534	0.563	-0.028
Job Offer	0.175	0.112	0.063***
Accept Offer	0.142	0.080	0.062***

Panel B: Incremental Acceptance Rates

	Treatment	Control	Difference
Interview	0.534	0.563	-0.028
Job Offer	0.327	0.199	0.128***
Accept Offer	0.812	0.714	0.097

Panel C: Incremental Acceptance Rates (Regression w/ Controls)

	Job Offer	Accept Offer
Treatment	0.13***	0.18**
	(0.043)	(0.088)
R^2	0.021	0.082
Observations	422	111

Notes: The above comparisons are of raw means. The above analysis restricts the machine's discretion to match the quantity in the human-selected condition.

Panel A shows cumulative acceptance rates (the rate of applicants who made it from the beginning to each stage). The second row of Panel A can be read to mean, "Of all applicants who applied, X% were extended an offer."

Panel B shows incremental acceptance rates (pass rates of applicants were accepted until the previous stage). The second row of Panel B can be read to mean, "Of all applicants who *were interviewed*, Y% were extended an offer."

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 5: Negotiations

	Negotiated Offer	Negotiated Offer
Treatment	-0.13**	-0.12**
	(0.062)	(0.056)
Controls	No	Yes
P-value	0.037	0.041
R^2	0.051	0.12
Observations	124	124

Notes: As described in Section 2, candidates extended a job offer sometimes request improved offer terms. In the regressions above, each observation is a candidate who was extended an offer. The dependent variable is whether the candidate received an updated job offer reflecting negotiations. The outcome variable thus reflects whether the candidate was successful in improving the terms (even in minor ways), and *not* whether the candidate requested changed terms. Success in improving the terms typically happens only if the candidate can persuade the firm that he/she has competing offers. It is possible that some candidates requested changes and were denied. However, the firm's managers report that they generally try to update the offer in some way in response to a request for different terms if there are competing offers.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 6: Job Performance

	Lines of Code (Added)	Lines of Code (Deleted)
Tenure at Firm (Days)	0.38**	0.14***
	(0.15)	(0.041)
F-stat (1st Stage)	11.5	11.5
Mean Outcome of Control	-0.088	-0.12
Difference	0.47***	0.26***
Observations	770	770

Notes: This table measures the on-the-job productivity of candidates in both groups. The regressions above use instrumental variables. Each observation is a candidate. The endogenous variable is tenure length at the firm measured in days. This is zero for non-hired candidates and positive for candidates who were hired and started work. The instrumental variable is the experiment, which affects tenure by altering who is hired at all.

The outcome variable is lines of code added (and deleted). See a discussion of this variable in Section 7. The resulting coefficient on tenure can be interpreted as lines of code added per additional day of work. I normalize the outcome variable using the mean and standard deviation of everyone hired through the experiment and compare this coefficient on the average lines of code per day submitted in the human-selected control group. The results of this test are reported in the "Difference" row.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 7: Leadership, Cultural Fit and General Aptitude

	Cultural Fit	Leadership	General Aptitude
Interviewed (inst w/ Treatment & RD)	1.89***	1.80***	1.02**
	(0.090)	(0.093)	(0.44)
F-stat (1st Stage)	455.2	387.3	96.6
Mean Outcome of Control	1.52	1.60	1.28
Difference	0.37***	0.2**	-.26
Observations	54328	54328	54328

Notes: This table presents results on the interview evaluations of the marginal candidate preferred by the algorithm, but rejected by a human. The coefficient estimated in the first row is the average assessment in each dimension of these marginal candidates. In the "Difference" row, I report the difference between these marginal machine-picked candidates against the average candidate in the experiment.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 8: Combination of Algorithmic + Human Evaluation

Panel A: Passing Interviews

	Job Offer	Job Offer	Job Offer	Job Offer	Job Offer	Job Offer
Human would Interview	0.14**					0.099
	(0.063)					(0.061)
Human Score		0.031***			0.020*	
		(0.012)			(0.012)	
Algorithm would Interview			0.25***			0.24***
			(0.051)			(0.051)
Algorithm Score				0.13***	0.12***	
				(0.024)	(0.025)	
R^2	0.18	0.18	0.22	0.26	0.27	0.22
Observations	333	333	333	333	333	333

Panel B: Accepting Offers

	Accept Offer	Accept Offer	Accept Offer	Accept Offer	Accept Offer	Accept Offer
Human would Interview	0.041					0.031
	(0.21)					(0.22)
Human Score		0.040*			0.027	
		(0.023)			(0.021)	
Algorithm would Interview			0.040			0.097
			(0.16)			(0.16)
Algorithm Score				0.11***	0.10***	
				(0.029)	(0.030)	
R^2	0.21	0.24	0.20	0.31	0.34	0.21
Observations	82	82	82	82	82	82

Notes: This table contains "horserace" regressions (Fair and Shiller, 1989) predicting candidate outcomes from machine and human assessments. See Section 7 for discussion.

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 9: Heterogeneous Treatment Effects: Education Degree and Quality

Panel A: Extended Offer

	All	Doctorate	Elite School	Top Tier	Non-Select
Treatment	0.15***	0.34***	0.14***	0.18**	0.089
	(0.042)	(0.086)	(0.054)	(0.076)	(0.22)
Adj. R^2	0.026	0.14	0.021	0.040	-0.067

Panel B: Interviewed

	All	Doctorate	Elite School	Top Tier	Non-Select
Treatment	0.36***	0.49***	0.31***	0.35***	0.44**
	(0.033)	(0.074)	(0.041)	(0.060)	(0.18)
Adj. R^2	0.22	0.32	0.19	0.20	0.16

Panel C: Marginal Pass Rate

	All	Doctorate	Elite School	Top Tier	Non-Select
Machine-only Success Rate	0.40***	0.70***	0.45***	0.52**	0.20
	(0.11)	(0.19)	(0.17)	(0.22)	(0.46)
F-stat	120.7	44.1	57.8	35.0	6.17
Machine + Human Rate	0.20	0.10	0.22	0.15	0.14
Difference	0.2*	0.59***	0.23	0.37*	0.06
Human Only Rate	0.023	0.027	0.033	0.018	0.0077
Difference	0.38***	0.67***	0.42**	0.5**	0.19

Notes:

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 10: Heterogeneous Treatment Effects: Education Topic

Panel A: Extended Offer

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Treatment	0.15***	0.15***	0.11	0.064	-0.014	0.38***	-0.097
	(0.042)	(0.045)	(0.069)	(0.11)	(0.22)	(0.11)	(0.19)
Adj. R^2	0.026	0.026	0.011	-0.014	-0.059	0.16	-0.030

Panel B: Interviewed

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Treatment	0.36***	0.36***	0.32***	0.37***	0.29**	0.38***	0.45***
	(0.033)	(0.036)	(0.055)	(0.10)	(0.13)	(0.085)	(0.16)
Adj. R^2	0.22	0.22	0.19	0.22	0.042	0.21	0.30

Panel C: Marginal Pass Rate

	All	CS	Other Sci/Eng	EE	Bus/Econ	Math	Hum/SocSci
Machine-only Success Rate	0.40***	0.41***	0.35*	0.17	-0.050	1.00***	-0.21
	(0.11)	(0.12)	(0.21)	(0.29)	(0.74)	(0.33)	(0.45)
F-stat	120.7	97.6	35.0	13.8	5.01	20.4	8.46
Machine + Human Rate	0.20	0.20	0.19	0.16	0.27	0.13	0.28
Difference	0.2*	0.21*	0.15	0.01	-0.32	0.87***	-0.49
Human Only Rate	0.023	0.022	0.020	0.029	0.028	0.049	0.048
Difference	0.38***	0.39***	0.33	0.14	-0.08	0.95***	-0.26

Notes:

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 11: Heterogeneous Treatment Effects: Background and Experience

Panel A: Extended Offer

	All	Referred	Not Referred	No Experience	Experience
Treatment	0.15***	0.12	0.15***	0.36***	0.094*
	(0.042)	(0.12)	(0.044)	(0.089)	(0.051)
Adj. R^2	0.026	-0.00068	0.029	0.14	0.0086

Panel B: Interviewed

	All	Referred	Not Referred	No Experience	Experience
Treatment	0.36***	0.16**	0.40***	0.43***	0.35***
	(0.033)	(0.065)	(0.037)	(0.077)	(0.040)
Adj. R^2	0.22	0.067	0.25	0.28	0.21

Panel C: Marginal Pass Rate

	All	Referred	Not Referred	No Experience	Experience
Machine-only Success Rate	0.40***	0.76	0.38***	0.84***	0.27*
	(0.11)	(0.77)	(0.11)	(0.23)	(0.14)
F-stat	120.7	5.73	119.8	30.8	76.9
Machine + Human Rate	0.20	0.24	0.19	0.18	0.20
Difference	0.2*	0.52	0.19*	0.66***	0.07
Human Only Rate	0.023	0.041	0.021	0.033	0.020
Difference	0.38***	0.72	0.36***	0.81***	0.25*

Notes:

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Table 12: Heterogeneous Treatment Effects: Prob Selection by Human Screeeners

Panel A: Extended Offer

	All	High Prob.	Low Prob.	High Uncertainty	Low Uncertainty
Treatment	0.15***	0.12*	0.18***	0.12**	0.14*
	(0.042)	(0.065)	(0.062)	(0.053)	(0.072)
Adj. R^2	0.026	0.011	0.052	0.013	0.027

Panel B: Interviewed

	All	High Prob.	Low Prob.	High Uncertainty	Low Uncertainty
Treatment	0.36***	0.23***	0.49***	0.30***	0.46***
	(0.033)	(0.044)	(0.048)	(0.041)	(0.053)
Adj. R^2	0.22	0.14	0.26	0.19	0.22

Panel C: Marginal Pass Rate

	All	High Prob.	Low Prob.	High Uncertainty	Low Uncertainty
Machine-only Success Rate	0.40***	0.53*	0.38***	0.40**	0.31**
	(0.11)	(0.28)	(0.13)	(0.17)	(0.15)
F-stat	120.7	26.9	104.0	50.7	74.9
Machine + Human Rate	0.20	0.31	0.11	0.24	0.14
Difference	0.2*	0.22	0.26**	0.16	0.17
Human Only Rate	0.023	0.060	0.012	0.029	0.015
Difference	0.38***	0.47*	0.37***	0.37**	0.3*

Notes:

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are robust.

Appendix

A Comparison to Dawes