June 2022

Working Paper

Determining systematic differences in human graders for machine learning-based automated hiring

Mike H. M. Teodorescu, Nailya Ordabayeva, Marios Kokkodis, Abhishek Unnam, and Varun Aggarwal

This working paper is available online at: https://www.brookings.edu/center/center-on-regulation-and-markets/

Center on
Regulation and Markets
at BROOKINGS

The Center on Regulation and Markets at Brookings creates and promotes rigorous economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets. The Center provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields.

Disclosure

Aspiring Minds (acquired by SHL in 2020), the organization which provided data access, had the opportunity to review the content of this working paper. Varun Aggarwal is Chief AI Officer at SHL. Other than the aforementioned, the authors did not receive financial support from any firm or person for this article or from any firm or person with a financial or political interest in this article. Other than the aforementioned, the authors are not currently an officer, director, or board member of any organization with a financial or political interest in this article.

Determining systematic differences in human graders for machine learning-based automated hiring[1]

Mike H. M. Teodorescu
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, mike.teodorescu@bc.edu

Nailya Ordabayeva
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, ordabaye@bc.edu

Marios Kokkodis
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, kokkodis@bc.edu

Abhishek Unnam
Aspiring Minds, an SHL Company, Gurgaon, India, abhishek.unnam@aspiringminds.com

Varun Aggarwal
Aspiring Minds, an SHL Company, Gurgaon, India, varun@aspiringminds.com

Abstract

Firms routinely utilize natural language processing combined with other machine learning (ML) tools to assess prospective employees through automated resume classification based on pre-codified skill databases. The rush to automation can however backfire by encoding unintentional bias against groups of candidates. We run two experiments with human evaluators from two different countries to determine how cultural differences may affect hiring decisions. We use hiring materials provided by an international skill testing firm which runs hiring assessments for Fortune 500 companies. The company conducts a video-based interview assessment using machine learning, which grades job applicants automatically based on verbal and visual cues. Our study has three objectives: to compare the automatic assessments of the video interviews to assessments of the same interviews by human graders in order to assess how they differ; to examine which characteristics of human graders may lead to systematic differences in their assessments; and to propose a method to correct human evaluations using automation. We find that systematic differences can exist across human graders and that some of these differences can be accounted for by an ML tool if measured at the time of training.

[1] Center on Regulation and Markets Series on Economic and Regulation of Artificial Intelligence and Emerging Technologies

1. Introduction

Choosing people to hire is inherently hard (Klazema 2018), due to the difficulty of predicting how a candidate might perform outside of the observed settings of the interview—the candidates' latent characteristics (Kokkodis 2018, Geva and Saar-Tsechansky 2016). Resumes and recommendation letters provide observable characteristics such as skills or degree qualifications (Kokkodis et al. 2015, Abhinav et al. 2017). Due to increasingly dynamic trends and automation advancements, these observed characteristics are highly heterogeneous, as new skills are now born and old skills die faster than ever (Autor et al. 1998, Autor 2001, Kokkodis and Ipeirotis 2016, Institute of Business Value 2019, Kokkodis and Ipeirotis 2020). The latent applicant characteristics are the applicants' true knowledge and abilities on the listed skills and qualifications (Geva and Saar-Tsechansky 2016, Kokkodis and Ipeirotis 2020). Altogether, the heterogeneity of the observed qualifications combined with the unobserved applicant qualities create an environment of high uncertainty; employers make hiring decisions according to idiosyncratic assessments of fit between opening requirements and grader characteristics, where the attributes of the graders who determine the training set for the hiring tool are likely unknown to the employer. Recent economics literature show that taste-based discrimination can be a substantial contributor to bias in hiring (Cowgill and Tucker 2019).

To facilitate hiring decisions by reducing search costs (Pathak et al. 2010, Brynjolfsson et al. 2011, Fleder and Hosanagar 2009) and to improve hiring outcomes, many organizations invest in advanced machine learning algorithms. Natural language processing algorithms assess prospective employees through automated resume classification (Bollinger et al. 2012, Cowgill 2020) based on pre-codified skill databases (Nadkarni 2001, Lai et al. 2016). Human decision makers then use these machine assessments to efficiently filter and identify the best candidate for each opening (Kokkodis et al. 2015, Abhinav et al. 2017, Horton 2017). Even though such algorithms succeed in reducing search costs and improving hiring efficiency, they can often backfire by encoding unintentional biases against groups of candidates. For instance, an Amazon hiring tool discriminated against female applicants (Dastin 2018), while a Xerox hiring algorithm discriminated against applicants of lower socioeconomic status (O'Neil 2016). Despite the extensive literature on algorithmic fairness (Kusner et al. 2017, Kearns et al. 2017, Chen et al. 2019) and overall fairness in machine learning (FATML 2019), human biases transpiring into the training data still remains a problem in machine learning and is particularly pertinent in hiring where subjective qualities in interviews matter (Mann and O'Neil 2016). Hiring by referral remains a valuable tool, as candidates referred to by existing employees are often better performers and less likely to leave the firm after being hired

(Burks et al. 2015).

Fairness criteria such as demographic parity, classification accuracy parity, equal opportunity, equalized odds, and others (there are over 20 fairness criteria in the machine learning literature, without a consensus yet on which criteria should be used in every situation—see Mehrabi et al. 2021), are typically used to determine whether an algorithm is biased against a particular protected group, such as based on gender, age, marital status, disability status, and others (there is a vast literature on protected attributes, see for example Hardt et al. 2016, Ajunwa 2019, Awwad et al. 2020, Teodorescu et al. 2021). Unfortunately, in practice, it is very difficult to satisfy fairness criteria across combinations of protected attributes (subgroup fairness, see Kearns et al. 2018, Teodorescu and Yao 2021). Furthermore, it is known that theoretically it is impossible to satisfy three or more group fairness criteria at the same time (the impossibility theorem, Pleiss et al. 2017, Chouldechova and Roth 2018). Recent work has proposed human-machine augmentation of machine learning tools as a potential solution to choosing the appropriate fairness criteria for the data and task (Awwad et al. 2020, Teodorescu et al. 2021), though related literature in ethics shows that perceptionsof fairness by stakeholders, such as applicants to an automated interview system as in the setting in this paper, are essential to adoption of the tool and are not well understood yet in existing literature (Tarafdar et al. 2020, Morse et al. 2021). Given all these limitations of fairness criteria, we propose quantifying how the characteristics of human graders lead to systematic differences in training data and attempting to correct for these differences before training the machine learning tool, given that corrections ex post (i.e., running fairness criteria after the algorithm is trained in order to determine if the algorithm is fair according to a particular protected attribute) based on fairness criteria alone do not always lead to a solution (Teodorescu and Yao 2021). In Teodorescu and Yao (2021), a publicly available credit dataset is analyzed with standard prediction algorithms against several popular fairness criteria and under certain combinations of fairness criterion-protected attribute no optimum is found.

Our approach of determining systematic grader-specific differences in training data differs from standard fairness criteria applied ex post to machine learning algorithms. We use fairness criteria to determine if there are any differences across categories of protected attributes using a baseline neural network algorithm on our experimental sample before testing for any systematic differences across graders. Upon further analysis using a set of questions addressed to graders, we find that systematic grader-specific differences can exist in training data and can be accounted for if tested for appropriately prior to training the machine learning model, as shown in regression model results. The techniques here may be used outside of the hiring context as well, as the surveys of the individual grader personality attributes are based on general-purpose questionnaires of human preferences from well-known tests in the psychology literature and can be applied prior to the graders performing a classification task such as choosing who to hire.

To study and compare machine learning with human evaluations, a university research team and an industry R&D team from an international skill testing firm worked together for this research study, combining experience from academic research literature with algorithm design and software engineering experiences (Aspiring Minds has partnered in this study to create anexperiment to determine how potential personality characteristics and personal biases of graders who are essential in training an algorithm may affect hiring scores of candidates.) Hiring scores, combined with candidate characteristics, form the training data for a machine learning-based automated interview algorithm. The study coauthors from the firm provided an experimental version of the software which allowed us to turn off certain features in order to run a between-subjects experiment at the candidate level[1]. The firm (Aspiring Minds, a subsidiary of SHL) specializes in hiring assessments across various skills, including English proficiency, computer programming skills, customer service skills, and others. It conducts these assessments on millions of prospective employees on behalf of large firms it contracts for. The company also runs a video-based interview assessment platform which grades applicants automatically based on verbal and visual cues. For our study, the company shared anonymized videos and assessment values of machine learning graders.

One might reasonably ask, how do human and machine assessments differ when making hiring suggestions? To investigate this question, we compare machine evaluations with human evaluations of the same applicants[2]. We conduct two behavioral experiments(one with U.S. participants and one with participants recruited in India) with random assignment which involved human participants (human raters) of different genders and socioeconomic backgrounds. The experiments had two goals: (1) to identify systematic differences in human evaluations of candidates which may result from a combination of interview modality and individual characteristics of the evaluator—if such differences are present, they suggest that at least some evaluators exhibit bias;

[2] The between-subjects condition here is that graders were randomized to see candidates only in a certain condition, such as reviewing the interview audio-only, text-only, or video-only. This change, as well as the use of a different training set, required an experimental version of the software which was created by the industry partner Aspiring Minds under a research Memorandum of Understanding with the PI of the academic team.

and (2) to explore whether such differences lead to discrepancies between human versus machine-generated evaluations of candidates, potentially impacting hiring outcomes.

To address the first goal (i.e., examine the effects of interview modality and individual characteristics), participants were randomly assigned to one of three between-subjects conditions in which they assess the same candidate interviews presented in the form of: (1) text responses in the absence of visual or audio cues (text-only condition), (2) verbal responses in the absence of visual cues (audio condition), or (3) video responses (video condition, which could potentially reveal the greatest extent of rater bias). In addition to examining the effect of interview modality, we also assessed raters' individual-level characteristics that were previously linked to common biases in evaluations. Specifically, we administered validated scales of individual differences that could reveal human biases in evaluations of job candidates. These included: (1) numeracy which includes the cognitive reflection test and which is typically used to assess individuals' cognitive abilities and reliance on intuitive versus deliberative modes of thinking (Frederick 2005, Peterset al. 2006, Weller et al. 2013); (2) trust in feelings as information which captures individuals' reliance on their affective reactions in judgments and decisions (Avnet et al. 2012); (3) maximizing versus satisficing tendency which captures individuals' propensity to seek the best decision at the expense of efficiency versus willingness to accept good-enough outcomes in order to gain efficiency (Schwartz et al. 2002); (4) right-wing authoritarianism which captures individuals' endorsement of traditional hierarchies and authority and tendency to derogate low-status groups (Rattazzi et al. 2007); as well as (5) the scales of patriotism, cultural elitism, nationalism, and women's rights to understand potential biases related to candidates' gender, origin, and socioeconomic status in evaluations (Pratto et al. 1994).

To address the second goal (i.e., to determine whether human systematic differences in evaluations feed discrepancies between human- and machine-generated evaluations), we compared the machine ratings of anonymized interviews to the ratings of the same interviews generated by human raters. Doing so enabled us: (1) to test if biases identified in human ratings also characterize machine ratings and assessments, and (2) to recommend strategies for training algorithms using human ratings data to mitigate these biases.

2. Automated Interviewing Approach

In the next paragraphs, we present the automated interviewing approach we used in our study.

2.1. Aspiring Minds - partner for study

Aspiring Minds, an SHL company, is an assessment firm, specializing in AI-powered assessment and interviewing solutions. Some of the prominent AI-based products include Automata - an AI based coding assessment (Aggarwal et al. 2013, Srikant and Aggarwal 2014, Aggarwal et al. 2016), SVAR - automated spoken evaluation test (Shashidhar et al. 2015 a, b) and Writex - automated evaluation of email (Unnam et al. 2019) and essay writing skills. Recently, Aspiring Minds has taken the lead in developing a science-based video interview scoring which has been deployed to evaluate millions of candidates every year.

Over the last decade, Aspiring Minds have put forth a system to transfigure the evaluation of subjective assessments in the framework of machine learning. In (Aggarwal et al. 2013), the authors discuss how to apply the principles of machine learning in the testing of open responses. Automata (Srikant and Aggarwal 2014, Aggarwalet al. 2016) uses machine learning techniques to grade computer programs on parameters like correctness, coding style, and computational complexity, as well as semantic feedback on uncompilable codes.Related work by authors in Shashidhar et al. (2015 a, b) evaluates spoken speech, using machine learning to evaluate speaking skills. In that tool, the applicant dials a number and has a conversation with the system. Aspiring Minds has also explored the domain of blue-collar jobs and how automatedevaluation could be fruitful there. In Singh and Aggarwal (2016 a, b), they propose the use of touch-based surfaces/tablets as a medium to test cognitive skills. In Unnam et al. (2019), Aspiring Minds proposes a system for assessing email writing skills.

2.2. Autoview Automated Interviewing Algorithm - Review of the Approach

Advances in artificial intelligence have transformed the recruitment processes in companies (Singhania et al. 2020, Aggarwal et al. 2013, Bhatia et al. 2019). There have been recent efforts to introduce automation to the initial stages of the recruitment process in the form of resume matching and classification in Bhatia et al. (2019), such as scoring of spoken English, essays, computer code, and emails.

Unstructured interviews are often the most important tool employers use to make a hiring decision. However, the ubiquitous use of unstructured interviews is vulnerable to avariety of biases that can significantly impact the decisionmaking process. Reilly and Chao (2006) researches the validity, adverse impact, and fairness of such interview-based selection methods and provides discouraging evidence of interview validity. McDaniel et al. (1994) shows the relative superiority of structured interviews over traditional unstructured interviews for selection purposes.Video-based interviewing has gained traction recently because of the flexibility it provides to the interviewers and candidates, as well as because of cost savings for employers and the increase in remote work due to the COVID-19 pandemic. The structured nature of such interviews can help reduce biases in hiring decisions and improve the reliability and validity of the selection methods. Nevertheless, humans are susceptible to judgment and decisionmaking biases in even highly structured environments and when evaluating structured stimuli. For example, decisionmakers' reliance on intuitive (rather than deliberative) processing of information, reliance on subjective feelings (rather than cognition), desire for efficiency and willingness to settle for good-enough outcomes (rather than maximize outcomes), as well as beliefs about certain social groups (e.g., minorities, women, low-status groups) often bias judgments and subsequent behaviors toward individuals, groups, and situations, including in the hiring context (Avnet et al. 2012, Frederick 2005, Peterset al. 2006, Pratto et al. 1994, Rattazzi et al. 2007, Schwartz et al. 2002, Weller et al. 2013).

There has been extensive research and development of tools which quantify emotions from interview images and videos. In Chen et al. (2017), an algorithm that predicts personality attributes from video interviews is presented, while in Hemamou et al. (2019) models to predict hireability based on job requirements and video interview results are introduced. In Naim et al. (2016), actual interviews were recorded in a lab setting and then models were developed to predict social skills based on these videos. Likewise, Chen et al. (2016) proposed the doc2vec paradigm—a novel feature extraction method formulated as visual words learned from video analysis outputs. There has also been research on reactions to and outcomes of video interviewing methods and applications of AI fromjob applicants' perspective. For example, Brenner et al. (2016) and van Esch et al. (2018) examined applicants' reaction to video interviewing and how using AI in the recruitment process influences the likelihood of potential candidates completing the application process.

The present research examines the role and outcomes of one widely used automation tool—Autoview—which is an on-demand video assessment and interviewing platform that evaluates candidateson competencies, domain knowledge, and personality. Candidates can take the interview anytime, anywhere using multiple devices including a mobile and desktop/laptop, with just a working internet connection and a webcam to record their responses.

Employers consider soft skills along with hard skills (domain knowledge, technical skills) when evaluating a candidate. Autoview codes both types of skills to facilitate hiring evaluations. While soft skills are developed over time and are harder to quantify, they relate to how people interact with others and are strong indicators of on-the-job performance. Such skills are further categorized based on verbal and nonverbal forms of communication. Autoview assesses candidate soft skills using nonverbal cues such as facial expressions. Other than soft skills, Autoview also transcribes candidates' spoken content and uses Natural Language Processing (NLP) to score the content on workplace skills and functional knowledge. Functional Knowledge evaluates the domain skills specific to the job role s/he is applying to, using question independent model trained. Employers can add questions to the interview; however this feature was not used in the custom version of the software used in the experiments in this paper as we focused on a general interview. If employers provide their own question, they also need to provide a set of ideal answers that overlap between candidate responses and can be measured (standard NLP tools such as word vector similarity and doc2vec).

2.3. Dataset

Aspiring Minds collected data of job seekers of different ages and educational backgrounds from multiple countries. The full video set resulted in a total of 5845 videos from 810 job seekers and remains proprietary to Aspiring Minds as their training set. In order to conduct systematic comparisons of automated versus human ratings of job candidates while controlling for candidate-level variation (candidate characteristics and identity), in this paper we analyzed a randomly selected subset of 100 general interview videos provided by Aspiring Minds as a testset for fairness. The Autoview original model (Singhania et al. 2020) used a much larger, multi-race, multi-country dataset (we use a custom dataset and code branch for this research paper with fewer parameters, a much shorter interview, and much smaller pool of candidates, thus the results are not performed with the actual Autoview product). The candidate response data was collected through the proprietary platform AMCAT. Each candidate had to go through an instruction video, where the candidate was guided to set up a noise-free environment with proper lighting. Each candidate answered nine pre-recorded questions in English. Each candidate was given a preparation period of 30 seconds per question before the candidate could respond for a maximum of one minute. The sessions were recorded via the candidates' own webcam. The structured interview included both general and field-specific questions, the latter of which can be customized to an employer. For this limited research we did not include any employer specific or field-specific questions.

2.4. Raters and Rating process for ML Training Set

Aspiring Minds recruited the original training set raters online. The raters (graders) had experience working as HR professionals or psychology training. Each rater would rate videos of the candidates' response to questions on each of the three social skills. We provide a rubric in Table 1 with the scales and scores the original set of graders provided. Videos of very low quality (videos recorded with a variety of camera setups based on candidate computer configuration) were dropped. A comprehensive process withbest practices from Singhania et al. (2020) was followed. Each rater rated 50 videos and those with inter-rater correlation ofless than 0.5 and mean-difference of more than 1.0 (on a 5-point scale) were removed. Each video was rated by at least three raters, which resulted in a score per social skill as the mean of the raters' grades for that candidate under the respective social skill rubric.

Some of the questions were additionally tagged to either workplace skills or functional knowledge. These skills are evaluated from the candidates' spoken content, where raters referred to the underlying rubrics to assign a score based on the audio sample of the candidate response (please see Table 2). The question in consideration — "Describe a time when you made a mistake or when you did not apply the necessary effort in a task or project. What feedback did you receive? How did you change your approach as a result of this feedback?" evaluates the competency *Learning Attitude*.

Removing the noisy samples, we were left with around 800 data points for each of the workplace skills. For functional knowledge, the questions were more straightforward and close-ended, relating to definitions and concepts from a particular domain. An example of a domain question from the banking sector would be "What is overdraft protection?", which would then be evaluated using a rubric for functional knowledge based on audio samples. After cleaning the dataset and removing noisy audio samples, we used 1650 audio files collected on 30 questions from 5 sectors. We also maintain a set of 3-5 correct answers to each question which is required for model building.

Speech Prosody and Facial Features: Speech patterns such as emphases, intonation, rhythm, stress markers and others are extracted from the videos using OpenSMILE (Eyben et al. 2013). Similarly, facial features were extracted using OpenFace (Batruaitis et al. 2016). The exact details of this portion of the algorithm remain proprietary to Aspiring Minds.

Table 1 on the following page shows sample ratings for the social skill rubric**.**

Level	Description
NA	**Cannot be rated based on the video**
4	The candidate smiled as appropriate and answered the questions in an engaging, positive and warm way. He/she seems friendly.
3	*In between these 2 levels.*
2	The candidate seems neutral. He/she smiles once in a while and shows some positive emotion towards the conversation.
1	*In between these 2 levels.*
0	The candidate doesn't seem to be very friendly. He/she hardly smiles and shows little or hostile emotion.
-1	Video not clear

Table 1 Sample Rubric for the social skill - Positive Emotion

Word and Sentence Embeddings: Word embeddings are used to capture the meaning ofthe words used in the interview, rather than just the word itself. The measurement here is done by computing the similarity between the vector of the answer of the candidate and the vectors of appropriate answers; this generates a score between 0 and 1. We project the high dimensional word space (with each word as single dimension) to a low dimensional continuous vector space. These embeddings present a numerical representation of the contextual similarities between words, thereby in the transformed space, semantically similar words are mapped to nearby points. We used the Word2vec model using Google's pre-trained model (Mikolov et al. 2013). BERT ("Bidirectional Encoder Representations from Transformers") by Devlin et al. (2019) provides sentence embeddings which captures the context in which the words were used in the sentence. Using BERT as well provides a better representation of the sentence, one at a time.

Level	Description
5	The candidate describes a situation which portrays a particular area of improvement for the person such as the ability to effectively communicate or prioritization of tasks. The candidate describes the detailed feedback provided and how s/he incorporated the advice to improve in the future.
4	*In between the two levels*
3	The candidate describes a situation which portrays a particular area of improvement for the person such as the ability to effectively communicate or prioritization of tasks. The candidate describes some generic feedback provided and doesnt comment much on how s/he improved.
2	*In between the two levels*
1	The candidate describes a situational error which could happen to anyone, say being late in a meeting and getting transactional feedback to improve.

Table 2 Sample Rubric for the workplace skill - Learning Attitude

Bag of Words: We also used features counting unigrams (one word, for example “algorithm,” bigrams (pairs of words like “machine learning”), and trigrams (triplets of words that appear together commonly in the language, such as expressions like “natural language processing"). Stemming was applied to words prior to counting them (stemming turns the word to the dictionary root of the word) and stopwords (connectors like *and*, *or*, *a*, *the*, etc.) were removed then counted where each word is represented by a pair where each word is calculated a frequency (for a typical pipeline diagram see for example Younge and Kuhn 2016, Teodorescu 2017). We divide the term frequency counts by Inverse Document Frequency (TF-IDF) (Ramos et al. 2003) to prioritize the most important words/phrases. This approach is standard in the NLP literature.

Semantic and Cosine Distance: For functional knowledge, we used semantic-distance-based features. We calculate a comprehensive set overlap between the response and the set of correct answers pertaining to the question, also considering the synonyms and similar words obtained from wordnet (Fellbaum 2012). Similarly, cosine distance (Huang 2008) is also calculated between the word embeddings of the response and the correct answers after pre-processing.

2.4.1. Feature Selection and Modeling

Supervised learning models were applied to each social skill. Candidate-level scores were derived by averaging the scores across videos (multiple videos per candidate). Every social skill can be linked to a set of certain facial features. This relationship can be both positive and negative. For example, a candidate showing marked nervousness is rated low on confidence. In the first approach, we select these features for each social skill based on expert guidance. The features are then subject to feature transformations, i.e., transforming the higher-dimensional time-series features into a meaningful lower dimension. We don't need to apply any additional transformations on prosody features, already subject to systematic processing and smoothing before being extracted from the whole audio. Here we used a transformer-based model consisting of self-attention and additive attention layers. This model relies on the correlations present in the sample dataset. A deep learning model was used. The mathematical formulation and information relating to the model architecture is explained entirely in Singhania et al. (2020).

Separately, training question-specific models for each of the workplace skills was required. The final feature vector comprises word-level, sentence embedding-level, and bag-of-words count features. This is followed by feature selection techniques to vary the number of features selected in the model building process, models with the lowest cross-validation (4-fold) errors were selected. Classical supervised-modelling techniques were used in model building.

In order to build a question-independent model for evaluating functional knowledge, the dataset is divided into seen and unseen sets. The seen set comprises the questions used in training and unseen sets have different questions from all sectors. Features based on semantic and cosine similarity are calculated followed by supervised modelling techniques to train the final model. The model performance is evaluated both on seen and unseen sets to accurately measure the generalizability of the model.

3. Human Raters — Experiments

To obtain human ratings of candidates who had been rated by an algorithm, we ran two experiments with new sets of human graders different from those used by the company[3]. The goal of the

[3] A simpler version of the trained model of the video assessment company was used for the machine-based scores in our experiments. The original ML model was trained on a much larger multi-country multi-race dataset as presented in

experiments was to identify systematic differences in candidate evaluations and to compare human ratings of candidates to machine (automated) scores[4] in order to assess whether human ratings significantly diverge from machine ratings, and if so, why.

To examine the robustness and generalizability of the findings, we conducted one study in a U.S.-based participant pool recruited on Amazon Mechanical Turk and one study in an India-based online participant pool recruited through Qualtrics Panels. Geographic diversity should help us assess the role of human raters' cultural backgrounds in candidate evaluations. Furthermore, to test whether the susceptibility of human raters to various biases depends on individuals' expertise in hiring decision-making, we recruited India-based participants specifically from a pool of professionals with prior experience in Human Resources.

3.1. Participants

In both studies, 18- to 65-year-olds were recruited for a 15-minute individual online survey administered in Qualtrics. The U.S.-based study did not include participant employment as a recruitment criterion, whereas the India-based study specifically recruited participants as graders who were employed full-time and had experience in HR. Participants in the U.S.-based experiment were recruited on Amazon Mechanical Turk in exchange for a $1.50 participant compensation. Participants in the India-based experiment were recruited by the Qualtrics Panels service in exchange for a fee which combined participant compensation and panel service fees. A total of 926 participants completed the U.S.-based experiment (there were additional 67 partially complete responses). A total of 1,133 participants completed the India-based experiment, 317 of whom passed the Qualtrics Panels attention checks (the Qualtrics Panels service mandates the inclusion of attention checksto track the rate of service fulfillment; there were additional 256 partially completed responses). We analyzed all recorded responses without exclusions (in the India-based study the results did not differ across responses that passed versus failed the attention checks). The research design and materials were reviewed and approved by the Institutional Review Board of the academic authors' institution prior to data collection.

Singhania et al. (2020). The experiments in this paper were not part of the original Aspiring Minds training set and are a separate code branch - the videos used in these experiments were excluded from tuning the model.

[3] In this experiment we only compare the social skills scores produced by the trained model.

3.2. Experimental design and method

Both studies collected human evaluations of a total of 100 actual interviews captured by Aspiring Minds. A sample of 100 interviews used in the experiments was randomly selected from a total of 810 total interviews available from Aspiring Minds, with an even split between featured male and female job candidates. In both studies, after completing a consent form, participants evaluated three randomly selected (out of 100) candidate interviews. Interview format presented to participants was manipulated between-subjects with random assignment (text only vs. audio only vs. video andaudio), such that all three interviews were presented to participants in a text-only, audio-only, or text-and-audio format.

3.3. Experimental measures

3.3.1. Candidate evaluations After viewing the video of each candidate, participants completed several measures about the candidate. As the key measure, participants indicated their likelihood of hiring the candidate (1 = very unlikely to 7 = very likely). This constituted the main dependent variable in our analyses.

We additionally captured several exploratory measures, including assessment of candidate strength, liking, friendliness, warmth, engagement, competence, experience, knowledgeability, expertise, confidence, expressiveness, articulateness, positivity (vs. negativity), and emotionality (all measured with single-item 7-point scales). In the audio and video conditions, we additionally captured evaluations of aspects only relevant for participants in the audio and video conditions: the strength of accent and understanding of candidate speech (both single-item 7-point scales). In the video condition, we also captured evaluations of the candidates' gesticulation and smile (both single-item 7-point scales).

3.3.2. Participants' individual characteristics

To capture participants tendency to engage in heuristic (vs. deliberative) processingof information and reliance on subjective affect (vs. cognition) in judgments and decisions, we administered the numeracy scale which included the cognitive reflection task (three items) (Frederick 2005) and additional validated numeracy items (six items) (Peters et al. 2006, Weller et al. 2013), and the trust in feelings scale (two items) (Avnet et al. 2012). To capture individuals' propensity to maximize the outcomes of their decisions (vs. satisfice - i.e., willingness to accept a good-enoughoutcome for the sake of efficiency), we administered the maximizing vs. satisficing tendency scale (six items) (Schwartz et al. 2002). Finally, to examine individual beliefs that may generate biases related to candidates' gender, cultural, and socioeconomic background, we administered the rightwing authoritarianism scale (five items) (Rattazzi et al. 2007), as well as the scales (Pratto et al. 1994) of patriotism (six items), nationalism (four items), cultural elitism (four items), and women's rights endorsement (four items). We examined the main effects of these individual-level grader characteristics to identify systematic human biases in evaluations of job candidates, as well as the interactive effects of grader characteristics with job applicant characteristics (such as applicant gender) to assess whether human biases apply similarly (or differently) to different job applicants.

4. Data

Aspiring Minds shared the video interviews of the 100 job candidates along with their anonymized basic demographics (hereafter referred to as "applicants"). This dataset was then transcribed into audio-only and video-only files for the experimental conditions.

The India-based experiment recruited participants on Qualtrics Panels with three between-subjects experimental conditions (text-only, audio-only, video-only). The experiment generated 3,871 individual grades (nine assessments per interview, times three experimental conditions: 1,234 grades in the text only condition, 1,248 grades in the audio only condition, 1,389 grades in the video only condition). The baseline text-only condition has not yet been analyzed in this version of the paper.

The U.S.-based experiment recruited U.S. participants on Amazon Mechanical Turk with three between-subjects conditions (text only, audio only, video only). The experiment generated 2,772 individual grades (9 assessments per interview, times three experimental conditions: 993 grades in the text only condition, 952 grades in the audio only condition, 907 grades in the video only condition). As above, only audio- and video-based conditions were used in this version of the paper.

Dependent variables We use two dependent variables: *Likelihood to hire* and *AI-Human difference*. Likelihood to hire captures the likelihood that a study participant would hire the interviewed applicant. All are on a Likert scale of 1-7.

In addition, we analyzed the AI-Human difference which captures the difference in hiring likelihood of a given candidate between the AI hiring ratings and the human ones. We estimate this score as follows:

$$AI - Human\ Difference = AI\ Score - Likelihood\ to\ Hire \qquad (1)$$

where AI score is the average of the AI predicted dimensions “positive emotion”, “composure”,and “engagement”. Conceptually, AI-Human difference shows the difference between the AI hiring system and the human raters. The larger this difference between the AI vs. human assessment of the candidate, the higher the potential for bias in the assessment of the candidate due to human factors.

Table 3 **Descriptive statistics of the dependent and focal variables India-based study**

		Mean	Median	StD	Min	Max
India-based study:						
Focal variables	Numeracy	2	1	2.2	0	8
	Trust-feel	6	6	1.2	1	7
	Maximize	5	6	1.2	1	7
	Elitist	6	6	1.1	1	7
	Women rights	6	6	0.93	1	7
	Patriotism	6	6	0.86	1	7
	Right-wing	6	6	0.94	1	7
	Gender	0.47	0	0.5	0	1
	Video (audio as baseline)	0.52	1	0.5	0	1
Dependent variables	Likelihood to hire	4.8	5	2.1	1	7
	AI-Human	−1.6	−2.1	2.1	−4.7	2.6

Table 4 **Descriptive statistics of the dependent and focal variables US-based study**

Mechanical Turk study:						
Focal variables	Numeracy	4.2	4	2.1	0	9
	Trust-feel	5	5	1.3	1	7
	Maximize	4.7	4.7	0.94	2.2	7
	Elitist	4.3	4.2	1.1	1	7
	Women rights	6	6	1.1	1	7
	Patriotism	4.5	4.7	1.3	1	7
	Right-wing	3.8	4	1.6	1	7
	Gender	0.49	0	0.5	0	1
	Video (audio as baseline)	0.49	0	0.5	0	1
Dependent variables	Likelihood to hire	3.9	4	1.7	1	7
	AI-Human	−0.74	−0.81	1.7	−4.4	2.6

Table 3 shows the descriptive statistics of the Indian experiment for the focal and dependent variables; **Table 4** represents the same for the US-based reviewers. Figures **1** and **2** in Appendix A shows the correlations of the focal variables in the two studies.

5. Machine Learning Fairness Criteria Study – Automated Interviewing

We study whether the models employed in this experiment using an automated interviewing approach are fair with respect to candidates' gender. Each of the candidate was provided with a total interview score, which was the average of the predicted socials skills on a scale of 0 to 100. The graders' propensity to hire (originally on a scale of 1-7) was scaled 0 to 100. The cut-score (denoted by θ) applied was to reject the lower third of scores[5]. We evaluate the resulting models on two popular group fairness criteria - Average Odds (Bellamy et al. 2018, Hardt et al. 2016) and Equal Opportunity (Hardt et al. 2016). Additionally, we calculate classification accuracy (Fawcett 2006), the proportion of candidates classified accurately by the model. We evaluated these metrics against human scores provided under different subject conditions - Text, Video and Audio from U.S. and India-basedraters.

We calculated the differences in these metrics, between the protected group (female) and privileged group (males). Significance is calculated using 2-sample bootstrap t-technique at 95% confidence level. The two groups with sizes (47 and 53) were sampled 10,000 times with replacement and the corresponding p-value was calculated.

The results of these fairness criteria checks are in Table 5. With regards to classification accuracy, there isn't a considerable difference. We see that there is significant difference in Average Odds and Equal Opportunity, additionally it is the protected groups which benefit in all scenarios. These differences could be mitigated using different methods. First, one could use techniques of post-processing (Kamiran et al. 2012, Pleiss et al. 2017, Hardt et al. 2016) to induce fairness in hiring. Secondly, certain assessment practices may be used. It is interesting to note that (as presented in **Table 6**), as we lower the cut-score, the differences in Average Odds (AO) and Equal Opportunity (EO) also decrease. Therefore, a uniform reduction in cut-score, accommodating the candidates impacted, can be used to mitigate potential differences in assessments across protected groups by an AI system.

[5] Companies generally use automated hiring assessment tools to eliminate the lower tier performers and select the next round of interviews.

Table 5 Differences in Classification Accuracy (Acc), Average Odds (AO), Equal Opportunity (EO) evaluated for total interview score pertaining to gender.

Condition	Raters type	Acc	AO	EO
Text	US	-0.06	0.34*	0.32*
Video	US	-0.06	0.33*	0.30*
Audio	US	-0.07	0.30*	0.25*
Text	India	-0.07	0.43*	0.31*
Video	India	-0.06	0.32*	0.36*
Audio	India	-0.07	0.38*	0.30*

$^{*}p < 0.05$

Table 6 Differences in Classification Accuracy (Acc), Average Odds (AO), Equal Opportunity (EO) evaluated for total interview score pertaining to gender across cut-scores.

Condition	Raters type	Cut-score	Pass % (Rater)	Pass % (AI)	Acc	AO	EO
Text	US	45	54	85	-0.07	0.21*	0.08
Text	US	52	32	72	-0.06	0.22*	0.1
Text	US	55	29	75	-0.06	0.27*	0.22
Text	US	58	22	68	-0.06	0.34*	0.32*
Text	India	55	88	75	-0.07	0.38*	0.26*
Text	India	58	82	68	-0.07	0.43*	0.31*
Text	India	60	80	64	-0.06	0.48*	0.35*
Text	India	63	66	56	-0.05	0.45*	0.48*

$^{*}p < 0.05$

6. Systematic Human Differences - Results

We estimate the following specifications:

$$\mathbf{DV}_i \sim \mathbf{X}_i + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{X}_i$ are the focal variables described in Tables 3 and 4, and DV ∈ {Likelihood to hire, AI-Human difference}.

Human Systematic Differences: Table **7** shows the results of the two studies looking at the grader characteristics. Column A1 includes only responses from the India-based study; column A2 includes only responses from the MTurk study; column A3 pools both studies together (since both utilized identical experimental design and measures)and uses study fixed effects and interaction variables. The two studies show general human biases around multiple attributes. In both studies, evaluators with higher Trust-feel (i.e., tendency to rely on feelings in decision-making) tend to rate candidates higher. U.S.-Based graders with higher Numeracy (i.e., tendency to rely on deliberative, rather than heuristic, processing) tend to give higher scores, however this effect is not visible for the Indian-based graders (note however that the applicants are also Indian, so this might be an interaction that depends on whether reviewers are observing candidates from the same cultural background.

The two studies are consistent in terms of Maximize (i.e., tendency to seek the best possible outcome rather settle for good-enough), where the Maximize score has a highly positive and significanteffect on hiring score. The two studies are also consistent in terms of Elitist (i.e., perception of one's cultural superiority), where a higher Elitist score of the interviewer increases the candidates' hiring likelihood. This latter result was unexpected.

The Women rights (i.e., endorsement of women's rights) scores decrease the likelihood of an applicant to get hired across the Indian-based study and the pooled population but is not significant in the U.S.-based reviewer study. On the other hand, higher Trust-feel (trust in feelings) has a positive effect on hiring scores in all the studies (see Table **7**).

Grader characteristics related to Patriotism (i.e., patriotic feelings towards one's country) and Right-wing (i.e., tendency to endorse the hierarchy and authority) are only significant in the U.S.-based reviewer population. The video flag Video (audio as baseline) (i.e., scores in the video vs. audio condition) was not significant in either study. U.S. participants with high U.S.-directed Patriotism were more likely to rate negatively applicants, while U.S. participants with higher Right-wing scores showed a positive bias for the applicants. This requires further research.

AI vs. Human: *Does AI correct for some of these biases?* Using the AI-Human difference variable, Table **8** shows how each focal characteristic increases or decreases the difference between AI and human scores. Conceptually, if a coefficient from Table **7** (which shows the effects of human raters' characteristics on candidate ratings) is reversed in Table **8**, then AI has counteracted against the impact of these human characteristics on candidate ratings, which could be interpreted as AI correcting for the observed human bias.

The results show that indeed, AI corrects for multiple types of human biases. Specifically, the coefficients of Numeracy, Trust-feel, Maximize, Elitist, Patriotism, and Right-wing are reversed in Table **8**, in both columns A1 and A2, which indicates that AI acts against the influence of these human characteristics in (human-based) experimental data on candidate evaluations.

An additional interesting observation from Table **8** is that the coefficient of Video (audio as baseline) × Gender is positive. This suggests that when male applicants are being interviewed with video, the difference between AI and humans increases **-** i.e., humans may be particularly biased in favor of male (vs. female) candidates in video (vs. audio) interview settings, which AI works to counteract.

We ran a regular linear regression model with clustered standard errors on the applicants, aswell as variants of the model with interaction effects. The findings were as follows: numeracy score and patriotism score both had negative large effects significant at the 0.0001 level on hiring outcome. Degree of perceived warmth interacted with applicant gender and had significant main and interactive effects. First, generally, the warmer the applicant appeared, the more likely they were to be hired. Second, female applicants were more likely to be hired than male applicants at every level of warmth except the lowest level (i.e., 1 = very cold on the 7-point scale).

Interestingly, applicants' (performance) score at high-school graduation had no impact on their likelihood to be hired after college; it may be that a high-school graduation score is not salient or observed by the human grader or is not a good indicator of good job performance after college. Endorsement of women's rights, i.e., how supportive graders report they are of women's rights and gender equality, on average yielded very high scores (6/7) and thus did not have an impact on hiring score. Nationalism did not have a significant impact on hiring decisions in the models we have run so far. Trust-in-feelings and maximizing-vs.-satisficing tendencies both had positive significant effects on hiring decisions, which indicated more positive evaluations of candidates among individuals who rely on their affective responses and who seek the best possible decisions. Similarly, elitism had a positive and significant effect on hiring decisions, which is interesting and worth investigating more with a different population of graders. The full-page results **Tables 7** and **8** follow next.

Table 7 Human biases in hiring choices

	India study	US study	Pooled
DV=Hire score	(A1)	(A2)	(A3)
Numeracy	0.02	-0.12^{***}	0.02
	(0.02)	(0.03)	(0.02)
Trust-feel	0.36^{***}	0.10^{*}	0.36^{***}
	(0.05)	(0.05)	(0.04)
Maximize	0.21^{***}	0.21^{**}	0.23^{***}
	(0.05)	(0.07)	(0.05)
Elitist	0.34^{***}	0.14^{*}	0.40^{***}
	(0.07)	(0.07)	(0.06)
Women rights	-0.32^{***}	0.03	-0.29^{***}
	(0.07)	(0.06)	(0.06)
Gender	0.16	0.23	−0.35
	(0.65)	(0.75)	(0.44)
Patriotism	−0.01	-0.17^{**}	−0.12
	(0.09)	(0.06)	(0.08)
Right-wing	0.03	0.13^{*}	−0.01
	(0.08)	(0.06)	(0.07)
Video (audio as baseline)	0.16.	−0.06	0.13
	(0.09)	(0.11)	(0.08)
Study (Mturk = 1)			1.09^{*}
			(0.52)
Numeracy × Gender	−0.04	−0.06	−0.06.
	(0.05)	(0.04)	(0.03)
Trust-feel × Gender	−0.13	−0.13.	-0.12^{*}
	(0.08)	(0.07)	(0.05)
Maximize × Gender	0.15	0.01	0.09
	(0.10)	(0.10)	(0.07)
Elitist × Gender	0.27^{*}	0.04	0.15^{*}
	(0.11)	(0.10)	(0.07)
Women rights × Gender	−0.01	0.00	−0.02
	(0.12)	(0.08)	(0.06)
Video (audio as baseline) × Gender	−0.26	−0.03	−0.14
	(0.18)	(0.16)	(0.12)
Patriotism × Gender	−0.22	0.09	−0.01
	(0.15)	(0.09)	(0.08)
Right-wing × Gender	−0.06	−0.07	−0.08
	(0.13)	(0.09)	(0.07)
Video (audio as baseline) × Study (Mturk = 1)			−0.13
			(0.10)
Numeracy × Study (Mturk = 1)			-0.16^{***}
			(0.03)
Trust-feel × Study (Mturk = 1)			-0.27^{***}
			(0.05)
Elitist × Study (Mturk = 1)			-0.34^{***}
			(0.07)
Women rights × Study (Mturk = 1)			0.34^{***}
			(0.07)
Right-wing × Study (Mturk = 1)			0.15.
			(0.08)
N	2415	1717	4132
R^2	0.22	0.11	0.24

Robust standard errors in parentheses. The constant term is estimated but omitted. Patriotism × Study (Mturk = 1), Numeracy × Study (Mturk = 1), and Right-wing × Study (Mturk = 1) effects for column A3 are insignificant and removed. (*** $p < .001$, ** $p < .01$, * $p < .05$).

Table 8 **AI vs. Humans**

	India study	US study	Pooled
DV=AI-Human difference	(A1)	(A2)	(A3)
Numeracy	−0.02	0.12^{***}	−0.03
	(0.02)	(0.03)	(0.02)
Trust-feel	-0.37^{***}	-0.09^{*}	-0.36^{***}
	(0.05)	(0.05)	(0.04)
Maximize	-0.21^{***}	-0.22^{***}	-0.23^{***}
	(0.05)	(0.07)	(0.05)
Elitist	-0.34^{***}	-0.15^{*}	-0.41^{***}
	(0.07)	(0.07)	(0.06)
Women rights	0.33^{***}	−0.02	0.30^{***}
	(0.07)	(0.06)	(0.06)
Gender	−0.52	−0.58	−0.01
	(0.65)	(0.73)	(0.44)
Patriotism	0.02	0.17^{**}	0.12
	(0.09)	(0.06)	(0.08)
Right-wing	−0.03	-0.13^{*}	0.02
	(0.08)	(0.06)	(0.07)
Video (audio as baseline)	-0.25^{**}	−0.04	-0.23^{**}
	(0.09)	(0.11)	(0.08)
Study (Mturk = 1)			-1.12^{*}
			(0.51)
Numeracy × Gender	0.01	0.06	0.03
	(0.03)	(0.04)	(0.02)
Trust-feel × Gender	0.13.	0.12.	0.11^{*}
	(0.07)	(0.07)	(0.05)
Maximize × Gender	-0.19^{*}	−0.01	-0.13^{*}
	(0.09)	(0.10)	(0.06)
Elitist × Gender	-0.32^{**}	−0.03	-0.18^{*}
	(0.10)	(0.10)	(0.07)
Women rights × Gender	−0.03	−0.00	0.04
	(0.11)	(0.08)	(0.06)
Video (audio as baseline) × Gender	0.47^{***}	0.31^{*}	0.41^{***}
	(0.13)	(0.15)	(0.10)
Patriotism × Gender	0.25.	−0.07	0.03
	(0.15)	(0.09)	(0.08)
Right-wing × Gender	0.19	0.06	0.08
	(0.13)	(0.09)	(0.07)
Video (audio as baseline) × Study (Mturk = 1)			0.14
			(0.10)
Numeracy × Study (Mturk = 1)			0.16^{***}
			(0.03)
Trust-feel × Study (Mturk = 1)			0.28^{***}
			(0.05)
Elitist × Study (Mturk = 1)			0.34^{***}
			(0.07)
Women rights × Study (Mturk = 1)			-0.34^{***}
			(0.07)
Right-wing × Study (Mturk = 1)			-0.16^{*}
			(0.08)
N	2415	1717	4132
R^2	0.23	0.12	0.24

Robust standard errors in parentheses. The constant term is estimated but omitted. Patriotism × Study (Mturk = 1), Numeracy × Study (Mturk = 1), and Right-wing × Study (Mturk = 1) effects for column A3 are insignificant and removed for brevity. (*** $p < .001$, ** $p < .01$, * $p < .05$).

7. Discussion

7.1. Generalizable Correction Approach for Sources of Grader Bias

A biased grade could result from two separate sources: error in the algorithm itself which results in a systematic error in the grade given by the AI system, or systematic differences from human graders who provide inputs to the training data. The latter has been the primary focus of this paper – finding how human graders can systematically differ in their assessments of the same candidates in hiring interviews in metrics unrelated to candidate performance. When one observes only the outcome without knowing the characteristics of the human graders that serve as the basis for the training data for the algorithm, it is difficult to distinguish the difference between the human grade and the ML grade as the source of the bias. The grade assigned by the machine (denoted as "Grade by AI") may have a fixed bias, while the bias of the humans depends on the particular grader:

$$\text{Grade by AI} = \text{bias of AI} + \text{ideal grade}$$

$$\text{Grade by human} = \text{bias of human(grader feature vector)} + \text{ideal grade}$$

The difference between the two is the difference in biases. The systematic differences generated by human graders could be inferred from the grader feature vector, with questionnaires and tests as described earlier in this paper. However, not all systematic differences due to human graders may be observable. When these observable features are negligible (for example, no trust in feelings, no elitism, etc.), the grades assigned by humans may approach an ideal grade, but such ideal human graders may not exist. Furthermore, it is likely there are personality features which are unknowable by such scales, including some which perhaps have not yet been quantified. We may however be able to determine the differences between human evaluators and algorithm; these differences may include, for instance, the fixed bias of the algorithm and the variable biases of the graders, with both categories of bias unknown:

$$diff(expert_h, candidate_k) = algorithm\_grader(candidate_k) - expert\_grader_h(candidate_k)$$

We denote by $G_0(k)$ the ideal, unbiased, and unknown grade for the k'th candidate and the bias of the algorithm by B_{Alg}, that of the human expert by $B_{Eh}(k)$, and further the subscript h representing the *h'th* human grader/expert grade. With these notations, we can rewrite the formula above as:

$$diff(h,k) = G_{Alg}(k) - G_{E_h}(k) = (G_0(k) + B_{Alg}) - (G_0(k) + B_{E_h}(k)) = B_{Alg} - B_{E_h}(k).$$

To solve this, we would make a relatively strong assumption: that the bias of the algorithm is constant with regards to the candidate. This may be the case in some systems. However, we are aware that the human experts' grade biases are dependent on their feature vector, thus (here *H_h* is the bias vector representing expert *h*):

$$B_{E_h}(k) = B_{E_h}(k, H_h)$$

We would then define the ideal evaluator / grader as the average of the evaluators with zero-components of the feature vector *H*, i.e., the following limit is satisfied for any *k:*

$$\lim_{H_h \to 0} B_{E_h}(k) = 0 \quad \forall k$$

To determine the average bias of the algorithm knowing the feature vectors of the human graders (like in our experiments), one could follow this generalizable procedure: determine the subsets of evaluators with the bias vectors (or more generally, feature of grader vectors) H_h amongst the lowest 5%, 10%, 15%, … etc. and the sequences of average differences:

$$D_5 = diff(h, k_{5\%}),\ D_{10} = diff(h, k_{10\%}),\ D_{15} = diff(h, k_{15\%})$$

which would indicate, maintaining the assumptions in this analysis and notations, that the limit:

$$\lim_{j \to 0} D_j = B_{Alg}$$

would determine the ***average bias of the algorithm***, which once known can now be corrected. This proposed method is preliminary; analysis on the data using this approach is ongoing and will be the subject of future studies following this working paper.

7.2. Grade Distribution Considerations for Fairness and Numerus Clausus Approach

Equality of average grades across a protected attribute can be easily tested but does *not ensure* fairness even in the simpler fairness criteria to calculate, such as demographic parity applied to one attribute. For example, assume that employers decide to select candidates based on a minimum grade, which sets an implicit threshold for θ to be accepted to the next round. This threshold, if applied to the entire hiring process (for example 5% of the first round, as filtered by the ML tool, will go on for a second round of interviews) may not result in an equal number of candidates from each gender. The threshold-based process will result in a number hired, or a number promoted to the next phase of interviewing. This approach, while often used in practice (for instance, admissions to college would usually have a maximum target of admitted or in hiring would have a maximum number of new hires) does *not* guarantee demographic parity, even if – in a very simplified example with just one protected attribute – the average grade at which males are hired is the same as the grade at which female candidates are hired.

If we consider the number of male candidates N_m and number of female candidates N_f, the number of *selected* male and *selected* female candidates given the same grade threshold θ will not necessarily be the same, as we may not assume we are drawing candidates from identical distributions of grades. Instead, the number of selected male and selected female candidates will be based on the integrals:

$$N_{male_{selected}} = N_m \cdot \int_{\theta}^{Max_{grade}} p(grade|male)d(grade)$$

$$N_{female_{selected}} = N_f \cdot \int_{\theta}^{Max_{grade}} p(grade|female)d(grade)$$

In order to satisfy the demographic parity criterion, the following equality would need to hold:

$$N_m \cdot \int_{\theta}^{Max_{grade}} p(grade|male)d(grade) = N_f \cdot \int_{\theta}^{Max_{grade}} p(grade|female)d(grade)$$

whereas in order to satisfy the Equality of Opportunity criterion for male and female candidates, the probability of being hired would need to be equal across the protected attribute:

$$p_{hired}(grade|male) = p_{hired}(grade|female)$$

However, as many firms select for hiring or next round interviewing a fixed number of candidates per position, let's denote that fixed number of selected candidates as *N*, a *numerus clausus* condition is often applied in practice, which would in this case be:

$$N = N_m \cdot \int_{\theta}^{Max_{grade}} p(grade|male)d(grade) + N_f \cdot \int_{\theta}^{Max_{grade}} p(grade|female)d(grade)$$

Once the firm sets the number of candidates to be selected, given that the number of applicants N_m, N_f is known, as well as what the maximum grade is for the assessment given, one can solve for the threshold θ. A limitation of this discussion as well as of the entire paper is that we assume that the threshold θ is the same across firms, i.e., that the hiring process (be it ML-based, or human-decision-based), if applied the same across firms, will generate fair outcomes. It is however possible that different firms using the same hiring assessment tool may have different preferences in performance, thus the hiring threshold θ may be a firm-based decision. In that case, an ML model can be trained on a set of candidates of the firm in question (which would require access to the firm applicants, hiring records, and its hiring managers as graders for the training set) and the algorithm customized to the firm. That would require further a large enough pool of candidates for the training set, which would only be feasible in practice with large firms. While this is outside the scope of this paper, if such a study were possible in the future, the internal firm employee records could in fact provide a measure of 'ground truth' which would be the actual performance of the hired employees at for example three- or five-years post hire. Such performance data on the job would provide clarity regarding the True Positives and False Positives; however, in practice that data would be very difficult to obtain. That data would also not provide information on the False Negatives (people who were applicants, were denied hiring but would have been qualified, and went to another employer). Potentially that information could be inferred from platforms such as LinkedIn by looking at employee titles of peer employers at typical promotion thresholds for the field (for example, three- or five-year) and searching the applicants to see if they were promoted at their new employer. Such an endeavor is well outside the scope of this paper and would require a considerable effort, follow-up time following the experiment, and permission to run such an experiment at a large enough firm.

7.3. Future Work

One limitation of this study is that we were not able to follow up with candidates after they took the interview. We would also be interested to determine if we could find actual job success of the randomly sampled candidates in order to determine actual ability on the job. While, unfortunately, we were not able to obtain this information, if a future study is done in direct collaboration with employers or an aggregator of employment data, this may be possible to test.

Highschool GPAs did not appear to affect hireability, which was unexpected. The interaction of grader elitism with applicant gender was strong in the Indian grader study but not in the U.S. grader study. It may be interesting to test if these results hold in the case of technical interview questions, as opposed to a general introductory question as in this study.

There may be an ability to use the software effect or a home setting effect since candidates are taking the interview from home or university dorms with varying hardware and bandwidth configurations. We understand that the developers have applied certain corrections to the video streams to mitigate this, however due to the proprietary nature of the software we are not able to directly evaluate that component.

8. Conclusions

This working paper indicates that there are systematic differences across human graders which may seep into training data and which differ based on the gender of the job applicant as well as the ethnicity of the grader and of the applicant. In our U.S.-based graders, higher grader Numeracy and Patriotism resulted in a negative effect on the likelihood to hire. In our India-based study, a higher Women rights score decreases the likelihood of an applicant to get hired. However, higher grader Trust-feel and Elitist scores increase the likelihood a grader would see a candidate favorably and assign a high hiringscore. More research needs to be done to come up with recommendations applicable to industry as well as potential policy recommendations regarding measuring bias in training data for HRML applications. We do observe that the different grader scores across the various dimensions we measured about grader personality end up affecting likelihood to hire score differently based on the country (U.S. graders or India-based graders). We would like to be able expand the study to multiple countries, as well as to vary the profession of the grader. For instance, instead of a grader with a background in HR, it may be interesting to pick graders who have engineering backgrounds and to determine if they too are susceptible to trust in feelings, for example. In the later portions of the paper, we propose a method to correct systematic differences across graders provided grader features are observed.

A more general question regarding fairness criteria themselves is that the decision of which individual features are considered protected attributes is determined by the legal landscape but may not be necessarily complete (Morse et al. 2021). Indeed, the definition of a protected attribute may evolve to include further attributes over time, in which case algorithms may need to be retested and retuned.

Our final research objective was to determine if the use of a machine learning algorithm (which was in itsessence a deep neural network) can resolve some of these biases. We indeed found that ML corrects for multiple types of human biases, reversing Numeracy, Trust-feel, Maximize, Elitist, Patriotism,and Right-wing. Thus, knowing the source of human grader bias can enable the machine tool to remove it. Much more work remains to be done in this research topic - we share this as a discussion paper.

9. Acknowledgements

We thank USAID and MIT D-Lab for funding the study as part of grant *Appropriate Use of Machine Learning in Developing Country Contexts* AID-OAA-A-12-00095, which funded Mike Teodorescu's research, experiment expenses, and a research assistant, while he worked in collaboration with a diverse group of fairness researchers at MIT D-Lab over a period of about two years as an MIT Visiting Scholar. Summer research support, research assistant support, and IRB review were provided by Boston College, where Mike Teodorescu, Nailya Ordabayeva, and Marios Kokkodis have primary faculty appointments. A later version of this working paper was sponsored in part by Brookings Institution as part of the Working Paper Series on AI. Many scholars, government researchers, and industry experts provided feedback during the various revisions of this paper, and we wish to thank them for very helpful feedback and for their support. We thank Aubra Anthony, Shachee Doshi, Craig Jolley, Amy Paul (all USAID), Kendra Leith (MIT), Daniel Frey (MIT), Yazeed Awwad (MIT), and Melissa Mangino (MIT). We thank Rohit Takhar of Aspiring Minds. The first author thanks Research Assistants Peixuan Huang, Michelle Youn, and Madison Choo. We thank the Brookings Institution for sponsoring this working paper. We also thank Sanjay Patnaik, Bo Cowgill, and Rob Seamans for valuable feedback during the writing of this working paper draft as well as anonymous reviewers from the Brookings Center on Regulation and Markets who provided very helpful and insightful feedback. Mike Teodorescu also thanks the attendees of the inaugural Wharton/Columbia Management Analytics and Data Conference for feedback. Mike Teodorescu thanks USAID and MIT D-Lab for funding field research travel to India.

10. Author statement

Mike Teodorescu served as the Principal Investigator, conceived the study, obtained funding for the study, analyzed the data and conceived the models, and prepared most of the text of the paper, and was co-responsible for the IRB submission with Nailya Ordabayeva. Nailya Ordabayeva co-designed the randomized experiment, contributed to literature review, and to analysis of results. Marios Kokkodis contributed to literature review and analysis of results. Abhishek Unnam and Varun Aggarwal contributed video interview data, ran fairness metrics on the automated interview research project code, and are employees of Aspiring Minds. Abhishek Unnam and Varun Aggarwal are responsible for running the neural network tuning and other internal systems to the Aspiring Minds assessment products which are not available to the academic research team. This research is produced according to a Memorandum of Understanding between MIT and Aspiring Minds and is coauthored with industry researchers from Aspiring Minds.

References

2008. AMCAT https://www.aspiringminds.com/platform/delivery-technology/.

2008. Aspiring minds, an SHL company, https://www.aspiringminds.com/.

Abhinav, Kumar, Alpana Dubey, Sakshi Jain, Gurdeep Virdi, Alex Kass, Manish Mehta. 2017. Crowdadvisor:A framework for freelancer assessment in online marketplace. *International Conference on Software Engineering*. 93–102.

Aggarwal, Varun, Shashank Srikant, Vinay Shashidhar. 2013. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study.

Aggarwal, Varun, Shashank Srikant, Gursimran Singh. 2016. Question independent grading using machinelearning: The case of computer program grading. KDD.

Ajunwa, Ifeoma. 2019. The paradox of automation as anti-bias intervention. *Cardozo L. Rev.*

41 1671.Autor, David H. 2001. Wiring the labor market. *Journal of Economic Perspectives* **15**

25–40.

Autor, David H, Lawrence F Katz, Alan B Krueger. 1998. Computing inequality: Have computers changedthe labor market? *The Quarterly Journal of Economics* **113** 1169–1213.

Avnet, Tamar, Michel Tuan Pham, Andrew T Stephen. 2012. Consumers trust in feelings as information.

Journal of Consumer Research **39** 720–735.

Awwad, Yazeed, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, Mike Teodorescu. 2020.*Exploring fairness in machine learning for international development*. Technical Report, CITE MIT D-Lab, https://hdl.handle.net/1721.1/126854.

Baltruaitis, T., P. Robinson, L. Morency. 2016. Openface: An opensource facial behavior analysis toolkit.

2016 IEEE Winter Conference on Applications of Computer Vision (WACV). 1–10.

Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360:An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bhatia, Vedant, Prateek Rawat, Ajit Kumar, Rajiv Ratn Shah. 2019. End-to-end resume parsing and finding candidates for a job description using BERT. *arXiv preprint arXiv:1910.03089*.

Bollinger, Jacob, David Hardtke, Ben Martin. 2012. Using social data for resume job matching. *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*. ACM, 27–30.

Brenner, Falko S., T. Ortner, D. Fay. 2016. Asynchronous video interviewing as a new technology in personnel selection: The applicants' point of view. *Frontiers in Psychology* **7**.

Brynjolfsson, Erik, Yu Hu, Duncan Simester. 2011. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* **57** 1373–1386.

Burks, S. V., Cowgill, B., Hoffman, M., & Housman, M. 2015. The value of hiring through employee referrals. *The Quarterly Journal of Economics*, **130(2)**, pp. 805-839.

Chen, Jiahao, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, Madeleine Udell. 2019. Fairness under unaware- ness: Assessing disparity when protected class is unobserved. *Proceedings of the Conference on Fairness,Accountability, and Transparency*. ACM, 339–348.

Chen, Lei, Gary Feng, Chee Wee Leong, Blair Lehman, Michelle Martin-Raugh, Harrison Kell, Chong MinLee, Su-Youn Yoon. 2016. Automated scoring of interview videos using doc2vec multimodal feature extraction paradigm. *Proceedings of the 18th ACM International*

Conference on Multimodal Interaction.ICMI '16, Association for Computing Machinery, New York, NY, USA, 161168. doi:10.1145/2993148. 2993203.

Chen, Lei, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, Mohammed Ehsan Hoque. 2017. Auto- mated video interview judgment on a large-sized corpus collected online. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 504–509.

Chouldechova, Alexandra, Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

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

Cowgill, B., & Tucker, C. E. 2019. Economics, fairness and algorithmic bias. *NBER Working Paper*, http://conference.nber.org/confer/2019/YSAIf19/SSRN-id3361280.pdf

Dastin, Jeffrey. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October* **9** 2018.

Devlin, J., Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectionaltransformers for language understanding. *NAACL-HLT*.

Eyben, Florian, Felix Weninger, Florian Gross, Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM International Con- ference on Multimedia*. MM 13, Association for Computing Machinery, New York, NY, USA, 835838. doi:10.1145/2502081.2502224.

FATML. 2019. Conference on fairness, accountability, and transparency. https://fatconference.org/. [Online;accessed 02-December-2019].

Fawcett, Tom. 2006. An introduction to roc analysis. *Pattern recognition letters* **27** 861–874. Fellbaum, Christiane. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Fleder, D., K. Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systemson sales diversity. *Management Science* **55** 697–712.

Frederick, Shane. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* **19**
25–42.

Geva, Tomer, Maytal Saar-Tsechansky. 2016. Who's a good decision maker? Data-driven expert workerranking under unobservable quality. *International Conference on Information Systems*. AIS.

Hardt, Moritz, Eric Price, Eric Price, Nati Srebro. 2016. Equality of opportunity in supervised learning.
D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, eds., *Advances in Neural Information Processing Systems,* vol. 29. Curran Associates, Inc.

Hemamou, Léo, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, Chloé Clavel. 2019. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. *Proceed- ings of the AAAI Conference on Artificial Intelligence*, vol. 33. 573–581.

Horton, John J. 2017. The effects of algorithmic labor market recommendations: Evidence from a field experiment. *Journal of Labor Economics* **35** 345–385.

Huang, Anna. 2008. Similarity measures for text document clustering. *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4. 9–56.

Institute of Business Value. 2019. The enterprise guide to closing the skills gap. https://www.ibm.com/downloads/cas/EPYMNBJA. [Online; accessed 02-December-2019].

Kamiran, Faisal, Asim Karim, Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

Kearns, Michael, Seth Neel, Aaron Roth, Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*.

Kearns, Michael, Seth Neel, Aaron Roth, Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *International Conference on Machine Learning*. PMLR,2564–2572.

Klazema, Michael. 2018. Why is hiring the right employee so difficult? *Workplaces* [Online; accessed: 28-April-2019].

Kokkodis, Marios. 2018. Dynamic recommendations for sequential hiring decisions in online labor markets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* ACM, 453–461.

Kokkodis, Marios, Panagiotis G. Ipeirotis. 2016. Reputation transferability in online labor markets. *Man- agement Science* **62** 1687–1706.

Kokkodis, Marios, Panagiotis G. Ipeirotis. 2020. Demand-aware career path recommendations: A reinforce- ment learning approach. *Management Science* (forthcoming).

Kokkodis, Marios, Panagiotis Papadimitriou, Panagiotis G. Ipeirotis. 2015. Hiring behavior models for onlinelabor markets. *International Conference on Web Search and Data Mining*. ACM, 223–232.

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

Lai, Vivian, Kyong Jin Shim, Richard J Oentaryo, Philips K Prasetyo, Casey Vu, Ee-Peng Lim, David Lo. 2016. Careermapper: An automated resume evaluation tool. *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 4005–4007.

Mann, Gideon, Cathy ONeil. 2016. Hiring algorithms are not neutral. *Harvard Business Review* **9**.

McDaniel, Michael, Deborah Whetzel, Frank Schmidt, Steven Maurer. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. the journal of applied psychology, 79, 599-616. *Journal of Applied Psychology* **79** 599–616. doi:10.1037/0021-9010.79.4.599.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021. A survey onbias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54** 1–35.

Mikolov, Tomas, G.s Corrado, Kai Chen, Jeffrey Dean. 2013. Efficient estimation of word representations invector space. 1–12.

Morse, L., Teodorescu, M. H. M., Awwad, Y., & Kane, G. C. 2021. Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms. *Journal of Business Ethics*, pp. 1-13.

Nadkarni, Uday P. 2001. Skills database management system and method. US Patent 6,266,659.

Naim, Iftekhar, Md Iftekhar Tanveer, Daniel Gildea, Mohammed Ehsan Hoque. 2016. Automated analysisand prediction of job interview performance. *IEEE Transactions on Affective Computing* **9** 191–204.

O'Neil, Cathy. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pathak, Bhavik, Robert Garfinkel, Ram D Gopal, Rajkumar Venkatesan, Fang Yin. 2010. Empirical analysisof the impact of recommender systems on sales. *Journal of Management Information Systems* **27** 159–188.

Peters, Ellen, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* **17** 407–413.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, Kilian Q. Weinberger. 2017. On fairness and cal-ibration. *Proceedings of the 31st International Conference on Neural Information Processing Systems.* NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 56845693.

Pratto, Felicia, Jim Sidanius, Lisa M Stallworth, Bertram F Malle. 1994. Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of personality and social psychology* **67** 741.

Ramos, Juan, et al. 2003. Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, vol. 242. New Jersey, USA, 133–142.

Rattazzi, Anna Maria Manganelli, Andrea Bobbio, Luigina Canova. 2007. A short version of the right-wing authoritarianism (rwa) scale. *Personality and Individual Differences* **43** 1223–1234.

Reilly, Richard, Georgia Chao. 2006. Validity and fairness of some alternative employee selection procedures. *Personnel Psychology* **35** 1 – 62. doi:10.1111/j.1744-6570.1982.tb02184.x.

Schwartz, Barry, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, Darrin R Lehman. 2002. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of personality and socialpsychology* **83** 1178.

Shashidhar, Vinay, Nishant Pandey, Varun Aggarwal. 2015a. Automatic spontaneous speech grading: A novel feature derivation technique using the crowd. ACL.

Shashidhar, Vinay, Nishant Pandey, Varun Aggarwal. 2015b. Spoken english grading: Machine learning withcrowd intelligence. KDD.

Singh, Bhanu Pratap, Varun Aggarwal. 2016a. Apps to measure motor skills of vocational workers. Paul Lukowicz, Antonio Krüger, Andreas Bulling, Youn-Kyung Lim, Shwetak N. Patel, eds., *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp2016, Heidelberg, Germany, September 12-16, 2016* . ACM, 340–350. doi:10.1145/2971648.2971739.

Singh, Bhanu Pratap, Varun Aggarwal. 2016b. An automated test of motor skills for job selection and feedback. Tiffany Barnes, Min Chi, Mingyu Feng, eds., *Proceedings of the 9th International Con- ference on Educational Data Mining, EDM 2016, Raleigh, North Carolina, USA, June 29 - July2, 2016*. International Educational Data Mining Society (IEDMS), 694–699.

Singhania, Abhishek, Abhishek Unnam, Varun Aggarwal. 2020. Grading video interviews with fairnessconsiderations. *ArXiv* **abs/2007.05461**.

Srikant, Shashank, Varun Aggarwal. 2014. A system to grade computer programming skills using machine learning. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery anddata mining*. ACM, 1887–1896.

Tarafdar, M., Teodorescu, M., Tanriverdi, H., Robert, L., & Morse, L. (2020). Seeking ethical use of AI algorithms: Challenges and mitigations, ICIS Proceedings, *Forty-First International Conference on Information Systems*, India 2020.

Teodorescu, M. (2017). Machine Learning methods for strategy research. Harvard Business School Research Paper Series, (18-011).

Teodorescu, Mike HM, Lily Morse, Yazeed Awwad, Gerald C Kane. 2021. Failures of fairness in automation require a deeper understanding of human-ml augmentation. *MIS Quarterly* **45(3)**, doi:10.25300/MISQ/2021/16535.

Teodorescu, Mike HM, Xinyu Yao. 2021. Machine learning fairness is computationally difficult and algorithmically unsatisfactorily solved. *2021 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 1–8.

Unnam, Abhishek, R. Takhar, Varun Aggarwal. 2019. Grading emails and generating feedback. *EDM* .

van Esch, Patrick, J Black, Joseph Ferolie. 2018. Marketing ai recruitment: The next phase in job application and selection. *Computers in Human Behavior* **90** 215–222. doi:10.1016/j.chb.2018.09.009.

Younge, K. A., & Kuhn, J. M. (2016). Patent-to-patent similarity: a vector space model. Available at SSRN 2709238, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238.

Weller, Joshua A, Nathan F Dieckmann, Martin Tusler, CK Mertz, William J Burns, Ellen Peters. 2013. Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making* **26** 198–212.

Appendix A

Figure 1 Correlations of the focal variable in the India study

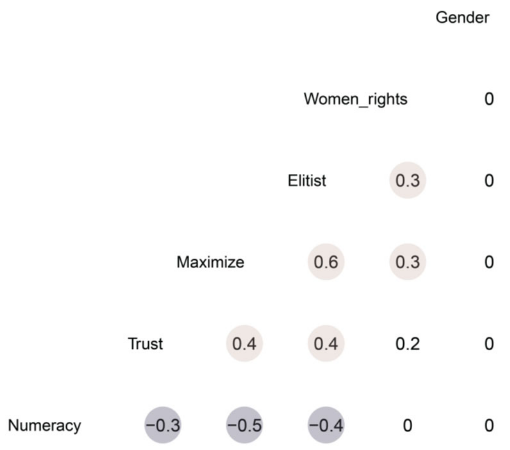
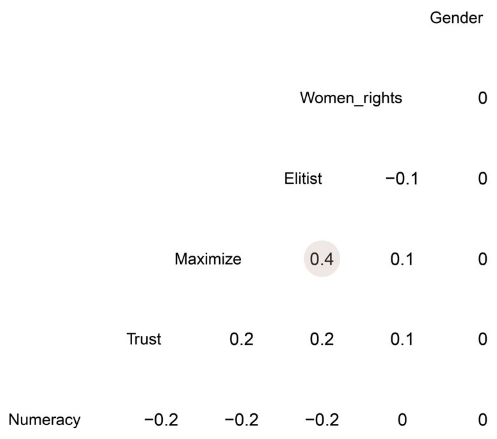


Figure 2 Correlations of the focal variable in the Mechanical Turk study



B Center on Regulation and Markets at BROOKINGS

The Center on Regulation and Markets at Brookings provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields. It creates and promotes independent economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets.

Questions about the research? Email communications@brookings.edu.
Be sure to include the title of this paper in your inquiry.

| 1775 Massachusetts Ave. NW, Washington, DC 20036 | 202.797.6000