Project Report: Perceptions of AI in Hiring

Rebecca Gelles
University of Maryland
College Park, Maryland
rgelles@cs.umd.edu

Duncan McElfresh
University of Maryland
College Park, Maryland
dmcelfre@math.umd.edu

Anjali Mittu
University of Maryland
College Park, Maryland
amittu@umd.edu

ACM Reference format:
Rebecca Gelles, Duncan McElfresh, and Anjali Mittu. 2018. Project Report: Perceptions of AI in Hiring. In *Proceedings of , College Park, MD, October 2018,* 16 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The primary goal of this study is to answer the question of what it means for a ranking algorithm to be fair in the context of hiring. This question has become more significant in recent years, as many employers now use Applicant Tracking Systems (ATS) to recruit, screen, and rank job applicants. ATS often include a wide range of tools for tasks including parsing resumes, administering custom application forms, managing applicant data, screening and sorting applicants, and communicating with applicants. New developments in AI and machine learning promise to improve human resource management. For example, recent papers promise to rank applicants based on personality traits [11], predict employee performance [6], predict employee turnover [18], and select corporate directors [10].

Employers have a legal, moral, and practical obligation to use hiring practices that are fair and just [2]. Automated procedures in hiring, such as those offered by ATS, make it difficult to assess fairness and justice. For example, suppose that an employer has 10,000 applicants: an ATS filters this list to 100 candidates, and then ranks these candidates. How do we know this filtering and ranking is fair? What does “fair" mean in this context? What principles of fairness and justice should new algorithms and AI systems uphold? As computer scientists, we approach this problem from the perspective of algorithmic ethics.

To examine these questions, we applied the work of Lee [16] to a closely related question. In the original study, the researchers compared peoples’ perceptions of human and algorithmic decisions. They considered four types of managerial decisions, which included work assignment, work scheduling, hiring, and performance evaluation. They constructed scenarios in which a decision was made, and the decision-maker was described as either algorithmic or human. For each condition, they measured perceptions of trust, perceptions of fairness, and emotional response.

1.1 Our Study

In contrast to previous work, we narrowed our focus to only one application of algorithmic decision-making which we considered particularly important: ranking job applicants. Given that a significant body of evidence had already been amassed comparing human decision-makers to computer-based ones, our study instead focused entirely on algorithmic decision-makers, delving into the question of what participants’ perceptions are of different kinds of algorithms rather than of algorithms in contrast to humans. To examine this question, we varied two properties of the algorithm, complexity (simple vs. complex) and transparency (transparent vs. opaque), to see how these changed participants’ perceptions. We aimed to measure how these two factors – complexity and transparency – impact respondents’ perceptions of a job-applicant sorting algorithm. Rather than including some situations which are purely humans based, we focused on situations in which every scenario contains an algorithmic decision-maker. With this, we hoped to tease out participants’ feelings about how algorithmic decisions are made, rather than focusing, as prior work did, on their opinions on whether they should be made at all. Our study focused on asking questions related to each of the following: trust in the algorithm’s ability to make good-quality decisions, fairness of the decision, and emotional response to the situation.

Unlike previous work, our study addressed how the design and presentation of an algorithm impacts perceptions of it. This is an important question for anyone designing an algorithm with societal implications, and particularly for employers, whose hiring practices may be subject to public scrutiny.

Hypotheses. In our study, we evaluated the following hypotheses:

H1 Transparent algorithms are perceived as fairer than opaque algorithms, and are trusted more by participants.
H2 Simple algorithms are perceived as fairer than complex algorithms, but are less trusted by participants.
H3 Opaque algorithms elicit a more-negative response than transparent algorithms, regardless of complexity.

2 RELATED WORK

While our work focuses on users’ understanding of the details of the differences between particular kinds of machine learning algorithms, some of what underlies these judgments is the fear of the potential for bias within these algorithms, and the possibility that algorithm authors will not consider that issue carefully. This fear is not unfounded: Datta et al. demonstrated that Google job advertisements were showing very different postings to men versus to women [7], and Caliskan et al. show that names derived from resumes that are evaluated by machine learning algorithms are evaluated as more “pleasant” if they come from European-American candidates [5].

, *October 2018, College Park, MD*
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

But the level of bias can vary based on the algorithm used, and research has focused on how to reduce this bias. The measures designed to reduce bias often add complexity, and reduce transparency, as they can counter-intuitively require the incorporation of the protected categories that there is a potential for bias against into the model in order to avoid discriminating on the basis of them [24]. This adds another avenue for risk, because of the need to protect and store sensitive information on, for example, job applicants. There has been work on how to avoid these issues; a recent paper by Kilbertus et al. evaluated the efficacy of avoiding disparate impact in algorithms by incorporating encrypted sensitive attributes, so that algorithm creators would not have access to the data, and found reasonable success [14]. Algorithms of this type are complicated by nature, which may place a preference for the ability to understand algorithms as a layperson and a preference for fairness in algorithms into conflict.

While our paper is exploring new territory by examining the details of what kinds of algorithmic decision-making users prefer, it is not the first to examine user opinions on algorithmic decision-making as compared to other types of decision-makers. One particularly relevant example is a recent paper by Lee [16], which studies perceptions of both algorithmic and human decision-makers. The authors composed four scenarios related to work assignment, scheduling, hiring, and performance evaluation. The researchers compared algorithmic and human decision-makers in each scenario, finding that respondents trusted algorithms to make better decisions than humans, in cases that require mechanical skills. In cases that require "human" (non-mechanical) skills, respondents trusted algorithms less than humans, and, surprisingly, responded with more negative emotions to algorithmic decisions than human decisions. Perceptions also depend on how the algorithmic decision-maker is presented. Recent work by Binns et al. [3] addressed how different explanation styles affect perceptions of algorithmic decision makers. The authors considered several decision-making scenarios, including hiring, and measured perceptions of fairness and justice of each. While the results were inconclusive, the authors found that when respondents encountered multiple explanation styles, respondents perceive some styles as more fair than others.

A related strain of research studies perceptions of bias in algorithmic systems. Woodruff et al. [22] conducted a workshop with participants from marginalized populations to discuss algorithmic bias. The researchers found that participants largely indicated that perceptions of bias would affect their trust of tech companies and products. Complementing this work, Grgić-Hlača et al. [12] investigated the underlying factors causing perceptions of bias. The researchers considered an algorithm for predicting criminal recidivism risk. They constructed scenarios in which an algorithm used certain features of an individual to determine recidivism risk, studying perceptions of fairness in how the algorithm was implemented.

As our work relates to the ability of laypeople to understand and trust the underlying features of machine learning techniques, interpretable machine learning models are an area that can aid in this effort, and help us define what transparency means in the context of machine learning. There has been recent development of tools that add interpretations to ML models: LIME [20], Gestalt [19], and ModelTracker [1]. Interpretability of ML models can be understood in two different ways: understanding how the model works or having the model explain the result [17]. The former, which is more relevant for this work, can be thought of as the transparency of the model. A model is fully transparent if the whole model can be understood at once, if each of the individual parts of the model can be understood, or if the algorithm can be understood [17].

While it is well understood that humans perceive information differently, it has been shown that any form of explanation helps to improve understanding of the model [20]. Studies have shown that there are many benefits to making your model interpretable [13], [20], [9]. Interpretations can be used to increase learning and understanding of the problem, promote safety and ethics, optimize the model to the correct criteria, and understand the trade-offs in the model [9]. Some suggest that interpretability and transparency could be the solution to the inability of humans to trust ML models [15] [23] [9]. This is because understanding the model helps to optimize and confirm the level of reliability, fairness, and trust in the model. However, other studies suggest model explanations could decrease trust in the model depending on the level of detail in explanation [4]. If the examiner felt that there was not enough details in the explanation to capture the complexity of the model, they were less likely to trust the model.

In our study, we modeled our interpretations on the LIME tool because LIME focuses on interpretations at the level of the individual prediction. This matches with the scenarios we used, modeled from Binns et al. [3] and Lee [16], which were also at this level. This method of explanation has been shown to improve understanding of the model for experts and novices in ML [20]. It has not yet been shown whether the LIME method of interpretation increases the perceived fairness or trust of the model compared to a model without interpretation.

3 METHOD

To evaluate participants' perceptions of fairness, trust, and emotional response in algorithms used for hiring, we relied heavily on the work of Lee [16] and Binns et al. [3]. However, as each of their work was focused on a broader set of scenarios than just hiring, and involved comparing participants' responses to human versus algorithmic decision-makers, their techniques required adaptation for our work. Nevertheless, we incorporated their core design principles, like the use of a survey providing participants with hypothetical scenarios involving named third parties who might go through the hiring processes described, and follow-up questions on Likert scales, in order to structure, collect, and analyze our data.

3.1 Recruitment

As hiring is a process that affects most people at some point in their life, we were not heavily concerned about targeting a specific audience for our survey, but were instead interested in reaching a large population. For these reasons, we chose the platform Mechanical Turk for recruitment. We recruited exclusively participants who were United States residents over the age of 18. Mechanical Turk has known data biases, which were readily apparent in the participant pool we ended up recruiting [8]. Likely at least partially as a result of these biases, our participant pool ended up quite skewed: it is significantly younger, more educated, more white, and more

Category	Value	Total	Percent
Gender	Male	132	65%
	Female	70	35%
Age	18-29	81	40%
	30-39	87	43%
	40-49	22	11%
	50-59	13	6%
	60+	0	0%
Race/Ethicity	Black or African American	20	10%
	White	148	74%
	Hispanic of Latino	13	7%
	American Indian or Alaska Native	11	6%
	Asian, Native Hawaiian, or Pacific Islander	2	1%
	Muliracial/Other	11	4%
Education	Less than high school degree	1	0.5%
	High school graduate, diploma or the equivalent	27	13%
	Some college credit, no degree	36	18%
	Associate degree (2-year)	27	13%
	Bachelor's degree (4-year)	84	41%
	Master's degree	25	12%
	Doctorate degree	3	1%

Table 1: Participant demographic information including gender, age, ethnicity, and educational attainment.

male than the general population. Demographic data can be seen in Table 1.

In addition to demographic data, we also collected information from participants on their knowledge in three key areas related to the core concepts of the study: algorithms, artificial intelligence, and human resource management (in particular, hiring and firing). This allowed us to ensure that our sample was not drawn purely from participants who were experts in the topics examined in our study, or, conversely, from those who had no knowledge of them at all. As can be seen in Table 2, the vast majority of participants indicated they were moderately, slightly, or not knowledgeable – these accounted for 74%, 76%, and 74% of each of the three categories, respectively. There are still approximately a quarter of participants in each category with a higher degree of knowledge, which should ensure that our results represent opinions of both the general population and those with some level of technical experience in the relevant areas.

3.2 Survey Design

As described above, we had four different algorithmic-decision-making conditions: simple and transparent, simple and opaque, complex and transparent, and complex and opaque. We wrote four scenarios, one for each of these conditions, using the scenarios from Lee [16] and Binns et al. [3] as a guide.

In order to avoid ordering effects, which have been shown to be quite significant in scenarios like this, we used a between-subjects design [23]. Each participant was only shown one scenario, representing one of the four conditions, and then asked follow-up questions on that scenario. Each of the four scenarios can be found in Appendix A.

For each scenario, we asked questions on a 7-point Likert-style scale, related to various perceptions of the algorithm, again using Lee [16] and Binns et al. [3] as a guide. We asked questions focused on three different areas of perception: fairness, trust, and emotional response. Specific questions in these areas can be found in Appendix A, and explanations of how these questions were compiled into more general evaluations of these three categories may be found in Section 4. We constructed the survey in Qualtrics, and connected it to Mechanical Turk's native survey tool.

3.3 Ethics

The study was approved by the University of Maryland Institutional Review Board. There are no known risks for the participants and the study is not targeting any vulnerable populations. We asked participants to complete an online consent form, which they were also offered a printable copy of, at the beginning of the survey, and any participants who failed to do so were disqualified from participating. All of the sensitive data that we collected, including MTurk IDs, IP addresses, cookies, and demographic data, was stored securely in a password-protected computer or in a reliable, security-conscious third party service.

3.4 Limitations

To eliminate ordering effects, we used a between-subjects design. However, between-subjects designs have their own limitations: in particular, this means that each of our scenarios were evaluated by different participants, with different backgrounds and levels of knowledge on the topics we studied. This is an inherent problem in studies of this nature with no easy solution. If we had found significance in any of our results, it would likely have been worthwhile to also run the tests considering each of our demographic categories and knowledge answers as covariants, in order to ensure that we were actually demonstrating a true effect, and not one caused by a confound.

We used MTurk to recruit participants, which allowed us to recruit from a large population of varying ages, incomes, and

Category	Value	Total	Percent
Algorithms	Extremely knowledgeable	12	6%
	Very knowledgeable	40	19%
	Moderately knowledgeable	79	38%
	Slightly knowledgeable	58	28%
	Not knowledgeable at all	17	8%
Artificial intelligence	Extremely knowledgeable	14	7%
	Very knowledgeable	36	17%
	Moderately knowledgeable	79	38%
	Slightly knowledgeable	67	33%
	Not knowledgeable at all	10	5%
Human resource management	Extremely knowledgeable	21	10%
	Very knowledgeable	31	15%
	Moderately knowledgeable	46	22%
	Slightly knowledgeable	54	26%
	Not knowledgeable at all	53	26%

Table 2: Self-reported participant knowledge about algorithms, artificial intelligence, and human resource management.

genders[21]. Still, Mturk's population also has known variation from the general population [8]. Turkers are generally younger, more educated, and have lower incomes than the average American [21]. While we did not ask about income, in all other contexts the same was true of our resulting pool of participants, which may have limited the external validity of our results.

Our study involves hiring decisions, which are only one algorithmic decision-making context. This thus makes it difficult to generalize from this topic to other contexts. In order to ensure that our examination of this particular topic was as strongly grounded as possible, we relied on the work of others. In fact, rather than developing our own concepts of trust, fairness, and emotional response, we employed the following two previous papers as a guide: [16] and Binns et al. [3]. Both papers give strong arguments for the why their techniques aptly measure these concepts in their survey design, so we relied on their strategies when designing our own.

4 RESULTS & ANALYSIS

4.1 Fairness and Trust

Participants were asked questions on their perceptions of the fairness and trustworthiness of each scenario. This was similar to the approach in Lee [16]. Participants rated how fair and how trustworthy the hiring situation was. See Appendix A for details on the questions.

Fairness. The perception of fairness for each scenario is related to our H1 and H2. The result of this question is shown in Figure 1. To test these hypotheses in regards to fairness we used the following null hypotheses.

$H1_{a0}$ Transparent and opaque algorithms produce the same distribution of fairness.

$H2_{a0}$ Simple and complex algorithms produce the same distribution of fairness.

For $H1_{a0}$ we grouped responses by transparency so that we could test each pair of differing condition. We ended up with four pairs of conditions. These were "simple/opaque" vs. "simple/transparent," "complex/opaque" vs. "complex/transparent," "simple/opaque" vs. "complex/transparent," and "complex/opaque" vs. "simple/transparent." We use the Mann-Whitney U test for each pair. We found that for every pair of conditions, we cannot reject $H1_{a0}$ ($p > 0.05$).

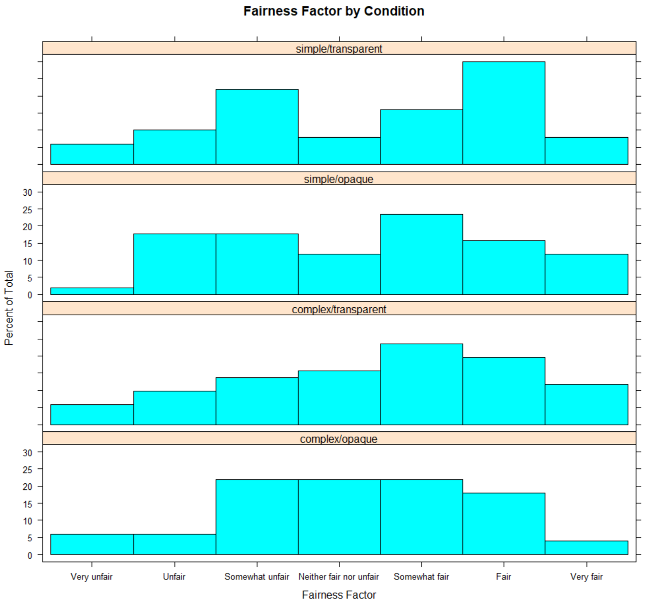

Figure 1: Histogram of fairness, by condition.

We used similar pairing for $H2_{a0}$; however, we grouped by complexity instead of transparency. We ended up with the following pairs: "simple/opaque" vs. "complex/opaque," "simple/transparent" vs. "complex/transparent," "simple/opaque" vs. "complex/transparent," and "complex/opaque" vs. "simple/transparent." Again, we used the Mann-Whitney U test for each pair and found that for every pair of conditions, we cannot reject $H2_{a0}$ ($p > 0.05$).

Trust. Similar to fairness, the perception of trust for each scenario is related to our H1 and H2. The result of this question is shown in Figure 2. To test these hypotheses in regards to fairness we used the following null hypotheses.

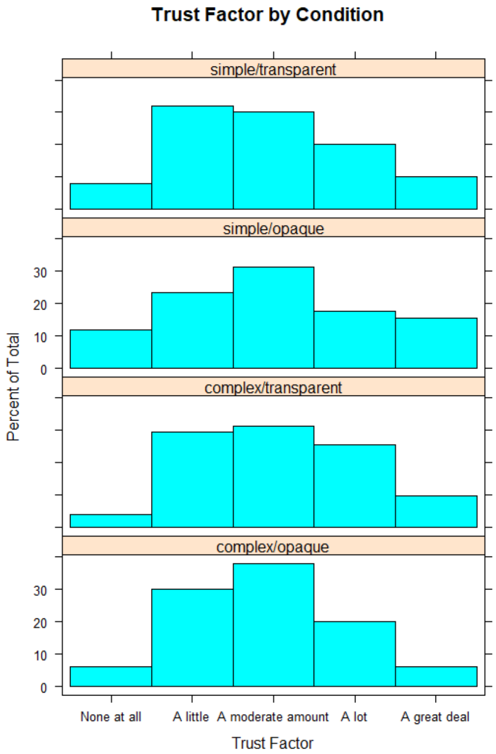

Figure 2: Histogram of trust, by condition.

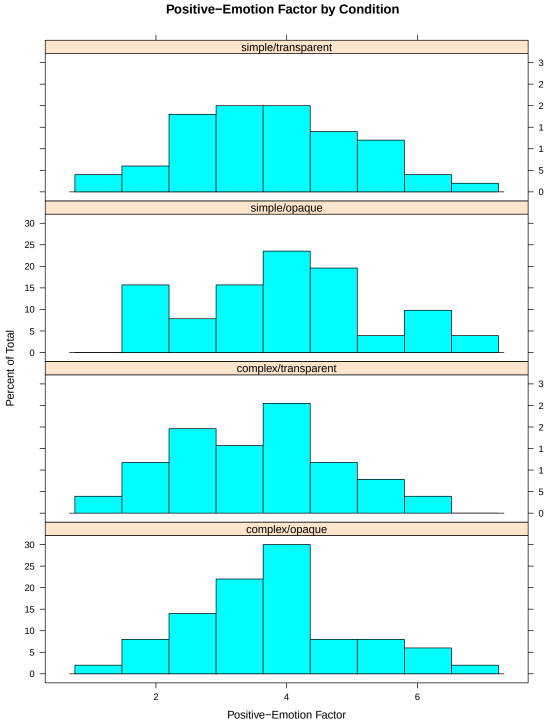

Figure 3: Histogram of emotional response factor, by condition.

$H1_{b0}$ Transparent and opaque algorithms produce the same distribution of trust.

$H2_{b0}$ Simple and complex algorithms produce the same distribution of trust.

We grouped responses in the same manner as the analysis on fairness. For $H1_{b0}$ we looked at groupings by transparency. We used the Mann-Whitney U test for each pair and found that for every pair of conditions, we cannot reject $H1_{b0}$ ($p > 0.05$). For $H2_{b0}$ we looked at groupings by complexity. Using the Mann-Whitney U test for each pair, we found that for every pair of conditions, we cannot reject $H2_{b0}$ ($p > 0.05$).

Due to the results of our analysis on fairness and trust, we are not able to confirm either H1 or H2.

4.2 Emotional Response

We asked several questions to understand how participants believe how the named subject ("Alex") of the AI hiring decision would feel in each scenario. Our approach mirrors that of Lee [16]. Questions included how much participants agreed that the hiring process would make Alex feel happy, joyful, proud, disappointed, angry, and frustrated. All questions were on a 7-point Likert scale of agreement; see Appendix A for details.

Emotional Response Scale. As in Lee [16], we aggregated these six questions into a single factor (positive-emotional-response): first we flipped the polarity of the positive-emotion questions (happy, joyful, proud), so all emotional responses had the same polarity; then, we average the answers to all questions, such that a higher answer corresponds to a more-positive emotional response. This positive-emotional-response scale is consistent – in our collected responses, we calculate a Cronbach's α of 0.88 (Lee reported $\alpha = 0.9$.)

Testing H3. We use a set of statistical tests to check H3 is confirmed in our experiments. The goal of these tests was to rule out a variety of null hypotheses. The first of these was the following:

H0-a Transparent conditions and opaque conditions produce the same distribution of emotional response.

To test H0-a, we grouped emotional responses into two samples: a) for conditions (simple/opaque) and (complex/opaque), and b) for (simple/transparent) and (complex/transparent). We used a Mann-Whitney U test to determine whether these samples have the same distribution, and found that we could not reject H0-a ($p > 0.05$).

Next, we tested a slightly different null hypothesis, comparing each pair of conditions (X and Y):

H0-b Condition X and Y produce the same distribution of emotional response.

We again used the Mann-Whitney U test for each pair of our four conditions (6 pairs in total). We found that for every pair of conditions, we could not reject H0-b ($p > 0.05$).

We also performed F-tests comparing variances for both H0-a and H0-b, and again found we could not reject the null hypotheses. In light of these tests, we could not confirm H3. However there were some nuances in these data that warranted closer inspection.

Exploratory Analysis. Despite our inability to reject the null hypothesis H3, some surprising features of the emotional response data caused us to examine it more closely. Figure 3 shows the distribution of emotional response for each condition.

There are subtle, yet visible, differences in these conditions–perhaps due to difference in responses to individual emotional-response questions. In particular, there are significant differences between the negative emotional response questions. Surprisingly, we find that transparent conditions elicit a more negative emotional response than opaque conditions.

To examine this further, we created a negative-emotional-factor, using only the negative emotional-response questions; this factor is consistent ($\alpha = 0.87$). Using a Mann-Whitney U test, we found that both the complex/transparent and simple/opaque condition pair and the complex/transparent and complex/opaque condition pair produced a different distribution of the negative-emotional-response factor.

Next, we fit a regression model on all conditions, using "transparent" and "opaque" as binary variables; we found that "transparent" is the only significant predictor ($\beta = -0.52, p < 0.05$).[1] These results indicate that more transparent algorithms may in fact elicit a more-negative emotional response than opaque algorithms, and this effect may be exacerbated by the complexity of the algorithm; this warrants further study. Given this exploration, the following hypotheses could be more relevant for future work, although qualitative research is likely also needed:

H4 Transparency in complex decision-making algorithms elicits a more-negative emotional response to decisions made by these algorithms.

H5 Transparency in simple decision-making algorithms does not impact the emotional response to these algorithms.

5 DISCUSSION

The survey and analysis conducted for this study were unable to draw any firm conclusions, as the study was designed to test three hypotheses about perceptions of hiring algorithms, and in each case tested we were unable to disprove the null hypothesis. Without additional information, we cannot confirm exactly what this means, as there are a variety of possibilities. We will examine each of these options in turn, and consider how future work could help identify which is most likely.

One option is simply that the effect we were looking to find had too small an effect size to be found given our sample size. Our sample size was quite reasonable, so if this were the case, such an effect, if it existed, would be fairly small. Given that not a single one of our hypotheses produced a statistically significant result, simply running a larger version of the same study is likely not the right choice without future work providing more evidence pointing to this conclusion.

Another possibility is that the hypotheses considered in this study do not actually match up to peoples' baseline opinions. Since our hypotheses were one-tailed, if our assumptions were incorrect we would not have identified a statistically significant result in the opposite direction. For example, if participants in our study had believed, instead of hypothesis H1, that opaque algorithms were fairer and more trustworthy than transparent algorithms, this result would not be found. The possibility that our basic assumptions about the directionality of participants' opinions may have been off seems to be at least partially backed up by the exploratory analysis performed on the emotional response data.

Similar to this possibility is the option that participants' perceptions of trust and fairness and emotional responses are not significantly influenced in either direction by the complexity and opacity of hiring algorithms. While previous work has demonstrated that these perceptions are influenced by whether a human or an algorithm is the decision-maker, we may have extrapolated too far from this work in assuming that the questions we aimed to answer were ones the general public would have opinions on.

In order to determine which of these possibilities may have contributed to our lack of results, one reasonable option would be to transition away from quantitative, deductive research and perform an inductive, qualitative, study on the topic of people's perceptions of algorithmic decision-making. This study would have a focus not on comparisons to human decision-makers, but on asking participants open-ended questions about what they perceived to be fair ways for algorithms to make decisions, when they would find decisions made by computer algorithms to be trustworthy, and how being evaluated by a computer would make them feel. By asking the questions in an inductive format, researchers would be able to build hypotheses that could then be formally tested in later quantitative work, with less chance of making errors in the directionality of hypotheses.

6 CONCLUSION

In our study, we aimed to measure participants' perceptions of trust, fairness, and emotional response when comparing algorithmic decision-makers used for hiring across two different axes, complexity and transparency. We tested three hypotheses based on perceptions of these situation, deriving our initial hypotheses from the previous work that existed in the field as well as our own intuitions. To test these hypotheses, we based our work off of two previous studies which tested very similar topics, but focused on comparing human decision-makers to algorithmic decision-makers, and tested a wide range of topics rather than focusing on hiring decisions [16] [3]. These studies guided our choices of study structure, questions design, and analysis technique.

After performing our study and follow-up analysis, we were unable to confirm any of the hypotheses we identified in our initial work. We put forth a number of possible explanations for this lack of confirmation, and suggested future work that could allow the exploration of which of these explanations is most likely to be correct, as well as providing relevant information to future researchers in this field. We believe that this future work could be valuable in helping creators of algorithm-based decision makers make better choices about how to create algorithms that people will find trustworthy and fair.

REFERENCES

[1] 2015. *ModelTracker: Redesigning Performance Analysis Tools for Machine Learning.* ACM âĂŞ Association for Computing Machinery.

[2] G Stoney Alder and Joseph Gilbert. 2006. Achieving ethics and fairness in hiring: Going beyond the law. *Journal of Business Ethics* 68, 4 (2006), 449–464.

[1] In this model, a negative β indicates a more negative emotional response

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[4] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[6] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–27.

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

[8] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, 135–143. https://doi.org/10.1145/3159652.3159661

[9] F. Doshi-Velez and B. Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints* (Feb. 2017). arXiv:stat.ML/1702.08608

[10] Isil Erel, Léa H Stern, Chenhao Tan, and Michael S Weisbach. 2018. *Selecting Directors Using Machine Learning*. Technical Report. National Bureau of Economic Research.

[11] Evanthia Faliagka, Athanasios Tsakalidis, and Giannis Tzimas. 2012. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet research* 22, 5 (2012), 551–568.

[12] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *arXiv preprint arXiv:1802.09548* (2018).

[13] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models. *Decis. Support Syst.* 51, 1 (April 2011), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

[14] Niki Kilbertus, Adrià Gascón, Matt J Kusner, Michael Veale, Krishna P Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. *arXiv preprint arXiv:1806.03281* (2018).

[15] Been Kim. 2015. *Interactive and Interpretable Machine Learning Models for Human Machine Collaboration*. Ph.D. Thesis. MIT, Cambridge, MA.

[16] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[17] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[18] Jiamin Liu, Yuxi Long, Ming Fang, Renjie He, Tao Wang, and Guosheng Chen. 2018. Analyzing Employee Turnover Based on Job Skills. In *Proceedings of the International Conference on Data Processing and Applications*. ACM, 16–21.

[19] Kayur Patel, Naomi Bancroft, Steven Drucker, James Fogarty, Andrew J. Ko, and James A. and Landay. 2010. Gestalt: Integrated Support for Implementation and Analysis in Machine Learning, In Proceedings of the 23rd annual ACM symposium on User interface software and technology.

[20] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[21] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, 2863–2872. https://doi.org/10.1145/1753846.1753873

[22] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.

[23] L Richard Ye and Paul E Johnson. 1995. The impact of explanation facilities on user acceptance of expert systems advice. *Mis Quarterly* (1995), 157–172.

[24] Indrė Žliobaitė and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (2016), 183–201.

A SURVEY TEXT

Below is the text of our survey. Note that each participant was only shown one scenario (i.e., only one of the four sections of the “Scenario Block” is shown to each participant.)

UNIVERSITY OF MARYLAND

Intro Block

Consent Form

Project Title	Hiring Decisions Survey
Purpose of the Study	*This research is being conducted by Anjali Mittu, Duncan McElfresh, Rebecca Gelles, and Michelle Mazurek at the University of Maryland, College Park. We are inviting you to participate in this research project because you meet our requirements. The purpose of this research project is to better understand attitudes about fairness in the use of automation and AI for HR purposes.*
Procedures	*The procedures involve the following steps:* *1) You will complete a survey regarding a specific hiring scenario involving computers.* *2) We will ask you some demographic questions.* *The entire process should take 10 minutes or less.*
Potential Risks and Discomforts	*There are no known risks to participants. We will collect some personally identifiable information (MTurk IDs, IP addresses, and cookies) to prevent repeat attempts, but this information will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.*
Potential Benefits	*There are no direct benefits to participating in this study. We hope that, in the future, other people might benefit from this study through improved understanding of what users consider to be a fair use and explanation of computation in and surrounding the hiring process.*
Confidentiality	*All survey answers will be collected and analyzed anonymously; survey answers will be stored in a password-protected server. To prevent duplicate participation, we will collect participants' MTurk ID, IP address, and use cookies.* *Any potential loss of confidentiality will be minimized by storing this data in a password-protected server, but it will not be associated with specific survey answers or data.*

	If we write a report or article about this research, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the University of Maryland, College Park or governmental authorities if you or someone else is in danger or if we are required to do so by law.
Compensation	*You will receive $1.20, and you will be responsible for any taxes assessed on the compensation.* *You are only allowed to participate once -- if we find you participating a second time, you will not be paid again.*
Right to Withdraw and Questions	*Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.* *If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:* ***Michelle Mazurek*** ***3421 A.V.Williams Building*** ***University of Maryland*** ***College Park, MD 20742*** ***301 405 6463*** mmazurek@cs.umd.edu

I am age 18 or older.

○ Yes

○ No

I have read this consent form or had it read to me.

○ Yes

○ No

I voluntarily agree to participate in this research and I want to continue with the survey.

○ Yes

○ No

We encourage you to print a copy of this consent form for your records.

Consent form 3

Before we begin, please verify that the Amazon Mechanical Turk ID shown below is your ID. If it is your ID, please click next. If it is not your ID, please enter your ID in the text field and then click Next.

MTurk ID: ${e://Field/MID}

Scenario Block

Please read the below situation and answer the following questions.

Alex applies for an engineering position on a job search website by submitting their resume and personal statement. The website lists skills that are required for the job. Each time an application is submitted, a computer model reviews the application. The computer model produces a score for each application by looking at the following factors:

- Keywords in the resume selected by the hiring manager
- Education
- Past experience

The website mentions that a computer will be evaluating the application but does not mention what factors will be considered significant. If Alex's score is high enough, they are called back for an interview.

Alex applies for an engineering position on a job search website by submitting their resume and personal statement. The website lists skills that are required for the job. Each time an application is submitted, a computer model reviews the application. The computer model produces a score for each application by looking at the following factors:

- Keywords in the resume selected by the hiring manager
- Education
- Past experience
- Similarities to applications from current high-performing employees
- Prediction of personality traits from wording of resume and cover letter

The website mentions that a computer will be evaluating the application but does not mention what factors will be considered significant. If Alex’s score is high enough, they are called back for an interview.

Alex applies for an engineering position on a job search website by submitting their resume and personal statement. The website lists skills that are required for the job. Each time an application is submitted, a computer model reviews the application. The computer model produces a score for each application by looking at the following factors:

- Keywords in the resume selected by the hiring manager
- Education
- Past experience

The website mentions that a computer will be evaluating the application and mentions what factors will be considered significant. If Alex’s score is high enough, they are called back for an interview.

Alex applies for an engineering position on a job search website by submitting their resume and personal statement. The website lists skills that are required for the job. Each time an application is submitted, a computer model reviews the application. The computer model produces a score for each application by looking at the following factors:

- Keywords in the resume selected by the hiring manager
- Education
- Past experience
- Similarities to applications from current high-performing employees
- Prediction of personality traits from wording of resume and cover letter

The website mentions that a computer will be evaluating the application and mentions what factors will be considered significant. If Alex’s score is high enough, they are called back for an interview.

To what extent do you understand this hiring process?

○ Completely understand

○ Mostly understand

○ Moderately understand

○ Slightly understand

○ Do not understand at all

To what extent do you think that this is an appropriate hiring process for an engineering position?

○ Extremely appropriate

○ Moderately appropriate

○ Slightly appropriate

○ Neither appropriate nor inappropriate

○ Slightly inappropriate

○ Moderately inappropriate

○ Extremely inappropriate

How likely is it that real companies are using a process like this one?

○ Extremely likely

○ Moderately likely

○ Slightly likely

○ Neither likely nor unlikely

○ Slightly unlikely

○ Moderately unlikely

○ Extremely unlikely

How much do you trust this algorithm to select good quality applicants for the company?

○ A great deal

○ A lot

○ A moderate amount

○ A little

○ None at all

How fair or unfair is it for Alex that the algorithm makes this hiring decision?

○ Very fair

○ Fair

○ Somewhat fair

○ Neither fair nor unfair

○ Somewhat unfair

○ Unfair

○ Very unfair

How much do you agree or disagree that the hiring process would make Alex feel:

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
Happy?	○	○	○	○	○	○	○
Joyful?	○	○	○	○	○	○	○
Proud?	○	○	○	○	○	○	○
Disappointed?	○	○	○	○	○	○	○
Angry?	○	○	○	○	○	○	○
Frustrated?	○	○	○	○	○	○	○

Demographics

What is your age?

Please specify the gender with which you most closely identify

○ Male
○ Female
○ Other
○ Prefer not to answer

Please specify your ethnicity (you may choose multiple options)

□ White
□ Hispanic or Latino
□ Black or African American
□ American Indian or Alaska Native
□ Asian, Native Hawaiian or Pacific Islander
□ Other

What is your country of residence?

What is your knowledge of algorithms?

○ Extremely knowledgeable

○ Very knowledgeable

○ Moderately knowledgeable

○ Slightly knowledgeable

○ Not knowledgeable at all

What is your knowledge of artificial intelligence?

○ Extremely knowledgeable

○ Very knowledgeable

○ Moderately knowledgeable

○ Slightly knowledgeable

○ Not knowledgeable at all

How much experience do you have with human resource management (hiring, firing)?

○ Extremely knowledgeable

○ Very knowledgeable

○ Moderately knowledgeable

○ Slightly knowledgeable

○ Not knowledgeable at all

Please specify the highest degree or level of school you have completed:

○ Less than high school degree

○ High school graduate, diploma or the equivalent (for example: GED)

○ Some college credit, no degree

○ Associate degree (2-year)

○ Bachelor’s degree (4-year)

○ Master's degree

○ Doctorate degree

○ Prefer not to answer

Block 3

Please make note of the following code. You will input it through Mechanical Turk to indicate your completion of the study. Then click the button on the bottom of the page to submit your answers. You will not receive credit unless you click this button.

Code: ${e://Field/RandomID}

Powered by Qualtrics