The Impact of Features Used by Algorithms on Perceptions of Fairness

Andrew Estornell[1†], **Tina Zhang**[2†], **Sanmay Das**[3], **Chien-Ju Ho**[1], **Brendan Juba**[1] and **Yevgeniy Vorobeychik**[1]

[1]Washington University in Saint Louis
[2]Amherst College
[3]George Mason University

Abstract

We investigate perceptions of fairness in the choice of features that algorithms use about individuals in a simulated gigwork employment experiment. First, a collection of experimental participants (the selectors) were asked to recommend an algorithm for making employment decisions. Second, a different collection of participants (the workers) were told about the setup, and a subset were ostensibly selected by the algorithm to perform an image labeling task. For both selector and worker participants, algorithmic choices differed principally in the inclusion of features that were non-volitional, and either directly relevant to the task, or for which relevance is not evident except for these features resulting in higher accuracy. We find that the selectors had a clear predilection for the more accurate algorithms, which they also judged as more fair. Worker sentiments were considerably more nuanced. Workers who were hired were largely indifferent among the algorithms. In contrast, workers who were not hired exhibited considerably more positive sentiments for algorithms that included non-volitional but relevant features. However, workers with disadvantaged values of non-volitional features exhibited more negative sentiment towards their use than the average, although the extent of this appears to depend considerably on the nature of such features.

1 Introduction

Systems relying algorithms for decision making are increasingly pervasive, and have significantly impacted the information and opportunities that people receive, with examples ranging from housing opportunities through Facebook's advertisements [Ali *et al.*, 2019], job opportunities through LinkedIn's talent search [Geyik *et al.*, 2019], to gig work employment on crowdsourcing markets [Hannák *et al.*, 2017]. This trend has necessitated careful investigations into both the fairness and efficacy of these systems, particularly in the context of vulnerable communities.

The scope of such investigations is two-fold: defining and formalizing what it means for an algorithmic decision-making system to be fair, as well as designing systems with algorithmic procedures or outcomes that adhere to these definitions of fairness. This line of research has lead to numerous conceptual frameworks for understanding algorithmic fairness, such as group fairness [Hardt *et al.*, 2016; Agarwal *et al.*, 2018; Kusner *et al.*, 2017], which aims to ensure that algorithmic decisions do not result in inequitable impacts on certain groups (e.g. historically marginalized communities), and individual fairness [Dwork *et al.*, 2012], which aims to ensure that similar decisions are made for similar individuals. Taking a broader perspective on fairness and justice considerations across a variety of domains, concerns of *procedural justice* aim to ensure that the decision-making procedures and institutions are perceived as *fair* by affected individuals [Thibaut and Walker, 1975; Lemons and Jones, 2001; Lee *et al.*, 2019]. Procedural justice has in turn received some recent attention in the context of algorithmic decision making [Binns *et al.*, 2018; Vaccaro *et al.*, 2019; Lee *et al.*, 2019; Wang *et al.*, 2020; Woodruff *et al.*, 2018]. While there has been considerable theoretical and legal discussion about the fairness of using certain types of features (e.g. race or gender) in decision-making [Fiss, 1970; Sánchez-Monedero *et al.*, 2020; Merritt and Reskin, 1997], a question that has received somewhat less attention in the literature is how the choice of features used by algorithms influences *human perceptions of fairness*. Existing work in this area includes that of Grgić-Hlača *et al.* [2018b], which considered aggregate opinions regarding the fairness of using specific features in specific decision contexts in the design of algorithms, balancing feature fairness and efficiency. We take up this thread by considering perceptions of feature fairness, as well as overall sentiments, from different stakeholders in an employment context.

Specifically, we designed a human subjects experiment in which participants were split into two roles: *selectors*, who are asked to choose which hiring algorithm we should use, and (prospective) *workers*, who are then hired, or not, via the chosen hiring algorithm. The central question in the experiment is how the choice of which features an algorithm uses impacts both the decisions and the sentiments of human participants *in both of these roles*. We systematically study this by viewing features along two dimensions: *volitionality* (a feature is a result of something that the individual can readily control, e.g., academic performance) and *relevance* to the task at hand. Relevance, in turn, can take two forms: *direct relevance*, when a feature is relevant to the task as naturally understood by people, e.g., debt in the context of lending decisions, and *implied relevance*, when a feature is not facially relevant, but nevertheless leads to higher accuracy through non-obvious channels. We create three algorithmic options centered around these issues, ordered by increasing accuracy: **Algorithm 1** uses features that are both volitional and directly relevant. **Algorithm 2** adds several non-volitional but directly relevant features to those in Algorithm 1. **Algorithm 3** adds features to Algorithm 2 that are neither volitional nor directly relevant, but which improve accuracy. We explain to participants that the selected hiring algorithm will be used to decide whether a particular individual (worker participant) would be hired to label dog breeds in a series of 10 images.

Selectors are asked to choose between two of these three algorithms, chosen at random, which they recommend to be used in making the above decision. Workers, in turn, first provide information that is used to construct features, and then are chosen (or not) for the image labeling task. Regardless of whether they are chosen, all workers are asked about their sentiments regarding the task, including perceptions of fairness. Finally, workers are asked to split a $1 bonus between themselves and their selector counterpart who chose the algorithm in their treatment; this was essentially a *dictator game* in which the worker played the dictator role [Güth *et al.*, 1982]. Our goal in this design was to elicit both explicit sentiments (via survey questions) as well as any implicit sentiments that do not directly emerge from survey responses (the tendency of workers to share a fraction of their bonus).

The experiment involved the use of deception when conveying the algorithm selection process as well as its possible deployment. Our central interest was in perceptions of fairness, rather than the task itself. Consequently, the choices of which workers to hire were in fact randomized and independent of worker features, despite workers being told that a specific algorithm was used to hire (or not hire) them. In addition, algorithm accuracies, were *design variables* that we created; no actual algorithms were developed or deployed. This experiment was approved by the IRB, subject to a detailed debriefing which was provided to both selectors and workers in the experiment. Throughout the experiment, we have received no complaints about our use of deception.

We found that the overall worker sentiment was quite positive. Workers who were hired expressed a more positive sentiment about the task than those not hired, as also observed in Wang *et al.* [2020]. Surprisingly, however, the fraction of hired workers sharing the final bonus was nearly identical to those not hired. Further, in contrast to Wang *et al.*, we find that the hiring decision is not necessarily the most influential factor in terms of worker sentiment; rather, in some cases having disadvantaged feature has a considerably stronger impact.

Interestingly, perceptions of relative fairness towards the three algorithms were quite different between selectors and workers. Selectors overwhelmingly chose Algorithm 3 over the others, and Algorithm 2 over Algorithm 1, and their fairness judgments generally aligned with this pattern.

Worker perceptions were more nuanced and influenced by contextual factors. Workers who were hired appeared essentially indifferent about which algorithm was used to make this decision. In contrast, those not hired expressed a strong preference towards Algorithm 2 and Algorithm 3 (which use non-volitional features) over Algorithm 1 (which uses only volitional features) when model accuracy was shown. However, there was not a clear preference between Algorithms 2 and 3 in this context. When accuracies were *not* shown, on the other hand, even workers who were not hired exhibited no significant *explicit* preference for any algorithm. However, in this case implicit sentiments were revealing: considerably fewer non-hired workers shared any of their final bonus with selectors when they believed that the algorithm used to make their hiring decision used features which were neither volitional nor directly relevant (Algorithm 3), compared to treatments involving Algorithms 1 and 2. On the other hand, more such workers shared a fraction of the bonus with selectors when the algorithm used non-volitional, but directly relevant features (i.e., Algorithm 2 was favored to Algorithm 1). Thus, in implicit sentiments, non-hired workers generally favored the algorithm using features that were clearly task-relevant, with volitionality being a secondary concern.

Our results thus reveal that neither selectors nor workers appear to view the non-volitionality of features used by the algorithm as inherently unfair. As such, both groups generally favored Algorithm 2 over Algorithm 1, if they favored any at all. On the other hand, the difference between selectors and workers appears to be due to the difference about judgments of feature *relevance*. Selectors seem to view an increase in accuracy as prima facie evidence of relevance. Workers, in contrast, appear to take special account of whether the relevance of features is direct and understandable (Algorithm 2), or solely evidenced by accuracy (Algorithm 3), which can be insufficient on its own to judge their use as fair.

Related Work: Common work in algorithmic fairness takes a computational perspective, focusing on defining what it means for an algorithm to be fair [Dwork *et al.*, 2012; Hardt *et al.*, 2016; Verma and Rubin, 2018; Kusner *et al.*, 2017; Buolamwini and Gebru, 2018; Mehrabi *et al.*, 2021; Hort *et al.*, 2022], auditing algorithms for bias [Washington, 2018; Buolamwini and Gebru, 2018; Wilson *et al.*, 2021], and designing algorithms which adhere to these definitions of fairness [Hardt *et al.*, 2016; Kusner *et al.*, 2017; Agarwal *et al.*, 2018]. This line of research does not seek to understand whether particular notions of fairness align with the expectations of individuals interacting with the algorithm. Moreover these definitions are framed over the outcomes of the algorithm, rather than the procedure use by the algorithm. Our work, in contrast, is focused on understanding perceptions of fairness in terms of procedural aspects of algorithmic decisions, in particular, the information (features) used by the algorithms.

Procedural justice, which motivates our work, is concerned with the design of the procedures or institutions with which individuals interact. While typical concepts in algorithmic fairness consider distributions of outcomes of algorithmic decisions, procedural justice is focused on the broader context within which such decisions take place, prioritizing considerations such as ensuring dignity of individuals, giving them a voice, as well as consistency and transparency of decisions [Tyler, 2006]. Procedural justice has been extensively studied in the context of criminal justice, employment, and promotion decisions [Fodchuk and Sidebotham, 2005; Houlden *et al.*, 1978; Lemons and Jones, 2001; Sunshine and Tyler, 2003; Thibaut and Walker, 1975; Tyler, 2003; Tyler and Huo, 2002; Tyler, 2006]. Such studies commonly demonstrate the significance of procedural justice in increasing social harmony, for example increasing overall satisfaction with decisions [Fodchuk and Sidebotham, 2005], likelihood of compliance with the outcome (e.g., arbitration) [Tyler, 2003; Tyler and Huo, 2002], satisfaction with one's employer, etc. Of particular significance in this line of study is the observation that individuals can maintain positive sentiment towards a system *despite unfavorable outcomes*, an inevitable consequence of scarcity of resources.

Although procedural justice is relatively under-explored in an algorithmic context, this issue has received some recent attention, with scholars investigating the trust, transparency, and accountability, of algorithmic decision-making systems [Binns *et al.*, 2018; Vaccaro *et al.*, 2019; Lee *et al.*, 2019; Wang *et al.*, 2020; Woodruff *et al.*, 2018]. Of particular relevance to our work are recent studies which have investigated the ways in which features chosen impact perceptions of fairness [Grgić-Hlača *et al.*, 2018b; Albach and Wright, 2021; Pierson, 2017; Grgic-Hlaca *et al.*, 2018a]. However, these works consider perceptions by those *outside* the decision-making process. In contrast, we consider the issue of fairness associated with features used by the algorithm in a human subjects experiment of a simulated employment scenario, in which judgments about fairness are made by the individuals who believe themselves to be directly affected by the algorithmic decision. In addition, we elicit fairness perceptions from two other perspectives: those with no stake in the process (pilot survey) and those tasked with selecting the hiring algorithm. Measuring perceptions from three differing perspectives is motivated by work on egocentric notions of fairness [Thompson and Loewenstein, 1992; Gelfand *et al.*, 2002; Greenberg, 1983] which demonstrate that one's role in the process can impact their perspective on fairness.

2 Experimental Design

2.1 Experiment Overview

We investigate judgments about the fairness of features used in an algorithm using a simulated employment experiment. All participants are told that the goal is to select a subset of workers to label a series of images of dogs with their corresponding breeds, and we were deploying an algorithm to make such a selection, from a menu of algorithms with differing sets of features and accuracy. Thus, while all participants were paid, those selected for the task received a pay specific to the task, in addition to all other payments. The stated rationale for this was deceptive by design: in fact, no algorithm was ever designed or used, and workers were selected for the task uniformly at random. Per the IRB-approved protocol, we debriefed all participants after the experiment in full detail.

Our experiment divided participants into two groups: *selectors* ($n = 114$), who were asked to choose between two algorithms in service of our stated (rather than actual) goal described above, and *workers* ($n = 1404$),[1] each paired with an algorithm that—in the way it was described to them—was used to decide whether they were selected for the task after extracting the features from them.

We conceptually categorize features along two dimensions: volitionality (whether it can be readily changed by the individual) and relevance (whether it is relevant to the task, in this case, labeling images). Additionally, we consider relevance from two vantage points: direct relevance, when the relevance of a feature to a task is evident, and implied relevance, when the feature increases accuracy (suggesting relevance), but it is not clear what the mechanism is through which it does so. As the nature of both volitionality and direct relevance is in part subjective, we used a pilot experiment to evaluate human judgments of both of these for a collection of features, as described presently. We used the results of this pilot to choose representative features that were directly relevant but non-volitional, and neither directly relevant nor volitional. At the end of the experiment, each participant was asked to opine on their perceptions of fairness, whether the decision made by the algorithm was justified, and whether they were satisfied overall. Finally, each worker participant was given a bonus, a part of which they are allowed to share with a selector who—according to our description—chose the algorithm that made the hiring decision impacting them (in fact, we did not pair participants directly; so we paid out the total amount of such bonus shares divided evenly among all selector participants).

For our experiment, we recruited a total of 1568 participants from Amazon Mechanical Turk, restricting location to be in the United States. We excluded incomplete responses from our analysis, and paid participants whether or not their data has been excluded. Since all our hypotheses are one-sided pairwise comparisons unless explicitly mentioned otherwise, we test for significance using one-sided t-tests when data is numerical and one-sided proportion z-tests when data is binary. When testing multiple pairwise comparisons, we use Tukey‘s range test to correct the corresponding p-values. We use TOST (two one-sided tests) with margin ε, to test for approximate equivalence [Lakens, 2017; Wellek, 2010], further details are provided in Section B of the supplement. Next, we provide further details for the main parts of the experiment; the complete set of experiment surveys is provided in Section F of the Supplement.

2.2 Pilot Survey

In order to select features that align well with the common meanings of volitional and relevant pertinent to our task, we first ran a pilot survey from 50 people (residing in the US) on Amazon Mechanical Turk. In this survey, we elicited volitionality and relevance information about the following features: *eyesight, age, race, employment, income, arrest record, history of substance abuse, zipcode, tobacco use, city and state of birth, parent's occupation,parent's income, and parent's*

[1] During the course of our experiment we made a single change to the worker survey, namely updating the language of the dictator game to more explicitly clarify that workers keep the remainder of the $1 which they did not give to selectors. Of the 1404 workers, 928 were given surveys with the updated language. When analyzing the $1 shares given to selectors we use only those workers who received surveys with updated language.

tobacco use. For each feature, we stated a hypothetical situation described as follows: *"Suppose we wish to develop a machine learning algorithm for hiring Amazon Mechanical Turk workers to provide labels for photographs"*. We then asked the participant's opinion on (a) whether this information is relevant to the hiring decision (relevance), (b) whether the individual has control over this characteristic (volitionality), and (c) whether it is fair to use particular input features in machine learning algorithms when hiring workers for this task (fairness); each scored on a 5-point Likert scale. Full details of the pilot survey are in Appendix C of the supplement.

We observed that age and race are judged as the least volitional features, while income, substance abuse history, arrest record, zipcode, employment history, and tobacco use are judged as highly volitional. For our task, eyesight and age are perceived as the most relevant features; both are also judged to be among the least volitional. Thus, in the main experiment, eyesight and age represent features that are relevant, but not volitional. We can also note that parent's income and occupation are among the least relevant and volitional features; we chose these to represent features which are not relevant and not volitional in the experiments. These observations align with prior work [Grgić-Hlača *et al.*, 2018b].

Perhaps the most surprising finding in our pilot survey is that judgments of fairness depend strongly on perceived task relevance of a feature, whereas *volitionality appears to play no role*. Specifically, we fit linear regression of fairness against relevance and volitionality. The coefficient corresponding to relevance is ∼0.9 ($p < 0.001$), while the coefficient corresponding to volitionality is ∼0.0 ($p > 0.3$). As we shall see below, this anticipates our findings in the main experiment.

2.3 Main Experiment

We now describe the design of our main experiment. Recall that participants were divided into two groups: *selectors*, who chose which hiring algorithm is to be used, and *workers*, who were told that a particular algorithm was used to determine whether they are hired or not. Next, we describe the main elements of the experimental procedure.

At the core of the experiment were three algorithms that differed along two dimensions: 1) the choice of features used and 2) accuracy. The details about the three algorithms (we simply refer to them as Algorithm 1, 2, and 3) are given in Table 1. As this table demonstrates, Algorithm 1 includes only features that are both directly relevant to the task and volitional (in the sense that they measure something prospective workers have significant control over, in our case, knowledge of dog breeds). Algorithm 2 adds two features (eyesight and age) that are deemed non-volitional but relevant (based on the pilot survey), while Algorithm 3 adds two more features (parent's occupation and income) that are generally viewed as neither relevant nor volitional. To avoid complicating the scenario with legal considerations, we deliberately excluded features such as race and gender. This experiment has two treatments on accuracy differences among algorithms: *small* (∼5%) and large (∼10%). For both treatments Algorithm 3 has a fixed accuracy of 94.7%.

	Features	Acc (T1)	Acc (T2)
Alg 1	Performance	88.4%	73.0%
Alg 2	Performance, eyesight, age	91.6%	81.9%
Alg 3	Performance, eyesight, age, parent's occupation/income	94.7%	94.7%

Table 1: Algorithms and accuracy shown to Selector and Workers. Performance is measured on image labeling, while other features are self-reported. T1 and T2 refer to treatments that vary accuracy.

Selector Procedure In our first set of experiments, we recruited 120 participants to the role of selector. Each selector was shown two of the three algorithms above (enabling a direct pairwise comparison), presenting both the features used and associated accuracies (based on two accuracy treatments). At this point, we screened their understanding of the algorithms by having them answer three validation questions, and only moved them forward if all three were answered correctly. We then explained to them that we wish to use one of the two presented algorithms to hire individuals using Amazon Mechanical Turk to label breeds for a collection of dog images. At this point, we asked them to recommend one of the two algorithms for us to use in hiring. After selectors made their recommendation, we asked them which of the two algorithms was more fair. At the end, we asked the selector participants to provide reasons for their recommendation and fairness judgments. Finally, we presented a detailed briefing that explained the deceptive elements of the design, and the actual experiment. In addition to the pair of algorithms presented, we systematically varied two design aspects of the selector survey: 1) small (∼ 5%) vs. large (∼ 10%) difference in accuracy between Algorithms 1 and Algorithm 2, and Algorithm 2 and Algorithm 3; and 2) whether we included an explicit cue in the description of the selector task emphasizing the importance of fairness. Further details are in Appendix E. Each selector was paid $0.5 (not including the bonus shares described below), and median task completion time was 4.8 minutes.

Worker Procedure Our second experiment involved prospective workers, done independently from the selector experiment. In this setting, each worker was randomly assigned to one of three treatments, each corresponding to an algorithm. We then provided background information about the task (similar to that for selectors), and presented them with all three algorithmic options, highlighting the actual algorithm ostensibly chosen for the task by the selector (who we said was a person we recruited using Amazon Mechanical Turk). We randomly divided workers into three treatment groups: those shown accuracy with 1) *small* differences and 2) *large* differences, and 3) those not shown accuracy information at all. Just as selectors, workers only proceed to the next step of the process if they correctly answer three validation questions ensuring that they have understood the task. Full details are in Appendix A.

Next, we elicit from each worker the full set of features that we tell them will be used by algorithm to make a hiring decision; workers are not told *how* these features will be used. To obtain features about ability to accurately label breeds of dog images, we ask each worker to label breeds for 10 dog images, for which they are paid $0.5; we do not tell them their efficacy at the end of this task. All other features are self-reported, and only elicited if the chosen algorithm is said to require them. Next, we introduce a small artificial time delay during which we say that the algorithm is making a hiring recommendation. In reality, the hiring decision itself randomly splits workers into the *hired* and *not hired* treatment groups. Any worker who is hired is asked to label an additional 3 images and receives an additional $0.5 bonus.

Finally, we elicit sentiments about the worker's experience. First, workers are asked to respond to a short survey that elicits their explicit sentiments about the experience in three ways, for which workers are paid $0.2. We ask 1) whether they felt that the procedure used to make the hiring decision was *fair*, 2) whether they felt that the hiring decision in their case was *justified*, and 3) whether they were *satisfied* with their experience. These three aspects capture for us *explicit* sentiments towards the task. Their choices for each sentiment are provided on a 5-point Likert scale, with 1 indicating strong disagreement, 3 indicating neutral sentiment, and 5 indicating strong agreement. Second, we capture implicit sentiments by giving each worker a final $1 bonus, and asking if they would be willing to share a fraction of this bonus with the selector who (they were told) recommended the algorithm used to hire, or not hire, them. This is effectively a well-known *dictator* game in behavioral economics [Camerer, 2011; Eckel and Grossman, 1996]. We measure whether workers *share* a nonzero fraction of the $1 (a decision with direct economic impact on themselves) as a means of capturing their implicit sentiments towards the hiring process.

After the survey we provide a detailed debriefing, describing the experiment and deceptive elements that were used. The median task completion time for workers was 8.4 minutes.

3 Results

Selector Perspective We begin with our analysis of the selector recommendations and fairness judgments. *We hypothesize that selectors will focus on the accuracy of an algorithm in each pairwise comparison; thus we expect that Algorithm 2 would be preferred to Algorithm 1 (H1), and Algorithm 3 to Algorithm 2 (H2).*

Our results support both H1 and H2: *selectors preferred to recommend Algorithm 2 to 1 ($p < 0.001$), and Algorithm 3 to 2 ($p < 0.001$).* Moreover, we find that their recommendations were largely, though not fully, consistent with their judgments of relative fairness of the three algorithms: *by a relatively large margin, Algorithm 3 more fair than 2, and was also judged more fair than 1 ($p < 0.001$ for both comparisons).* While Algorithm 2 was deemed more fair on average than Algorithm 1, this comparison only yielded $p = 0.1$, and is therefore inconclusive. Both of these observations can be gleaned from Figure 1. Across all algorithms selectors' recommendation and perception of most fair have a correlation 0.56 ($p < 0.001$). Thus, both recommendations and fairness judgments of selectors align closely with displayed accuracy of the algorithm. Moreover, when fairness judgments do clash with accuracy, recommendations follow the latter, as we can see in the difference between recommendations and fairness judgments for Algorithm 1 (Figure 1).

This general observation is further supported by qualitative data provided by the selectors in the form of an open-ended response rationalizing their recommendations and fairness judgments. We group this data into five categories: 1) *performance* (i.e., the algorithm has better performance), 2) *more features* (i.e., the algorithm used more features than other algorithms), 3) *relevance* (i.e., the algorithm used features that are task-relevant), 4) *other* (another reason), and 5) *uninformative* (no meaningful explanation provided; $\sim 35\%$ of responses). Full details are provided in Appendix E.

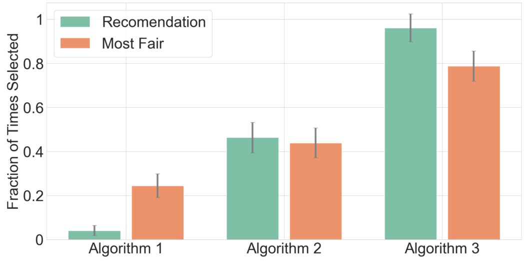

Figure 1: Frequency at which an algorithm was recommended for use, or perceived to be the most fair, scaled by how many times that algorithm was shown to selectors. Algorithm 3 is recommended more frequently, and perceived as more fair, than Algorithms 1 and 2 ($p < 0.001$); Algorithm 2 is recommended more frequently than Algorithm 1 ($p < 0.001$). Error bars represent standard errors.

We also consider the impact of different levels of relative Algorithm accuracies, and of the addition of a fairness cue compared to the accuracy-only framing. We found no statistically significant difference between selectors' recommendations and perceptions of fairness across these treatments. Full details are provided in Appendix E.

In summary, the primary consideration for selector's decision is model performance. Given the framing of the selector task, this is not in itself surprising. However, what *is* surprising is that fairness judgments were closely aligned with recommendations, and based primarily on efficacy judgments, with neither volitionality nor direct relevance of features having much impact.

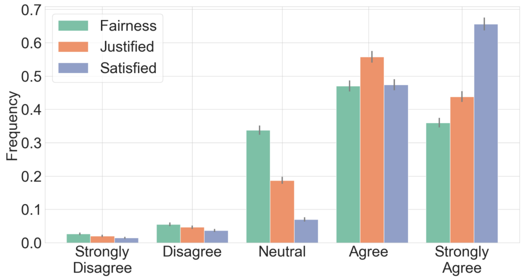

Figure 2: Distribution of worker sentiments.

Worker Perspective We begin by examining workers' general sentiments (perceptions of fairness, whether decision was justified, and overall satisfaction) towards the hiring procedure, aggregated over all treatments. We examine three hypotheses. *First, we expect that sentiments will be higher for workers who are hired than those who are not (H3). Second, we hypothesize that workers who are not hired exhibit less positive sentiments when placed in treatments involving the use of non-volitional* *features (H4). Third, we expect that such workers would also be less positively inclined towards the use of features that are not prima facie relevant to the task (H5).*

In general, worker sentiments are broadly positive, as shown in Figure 2. In particular, most participants agreed, or strongly agreed, with the statements that they were satisfied with the process and that the decision was justified. Fairness judgments were slightly more mixed, but again, very few expressed any negative sentiment on this measure either.

As we hypothesized (H3), being hired results in a more positive disposition towards whatever procedure was used in this decision in the case of *explicit sentiments*, as shown Figure 3 (left). The differences for each explicit sentiment (fairness, justified, and satisfied) between being hired and not hired are statistically significant ($p < 0.001$). Surprisingly, however, the fraction of workers sharing the final $1 bonus with selectors was insensitive to being hired (Figure 3, (right); $p < 0.001$ for approximate equality margin of 0.05). Thus, H3 is not supported in the case of implicit sentiments.

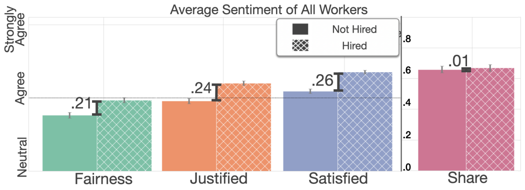

Figure 3: Average explicit sentiments (left) and fraction of workers sharing the $1 bonus during the dictator game (right) for hired vs. not hired workers. Sentiments for hired workers are greater than not hired workers ($p < 0.001$ for all three sentiments). In contrast, the fraction of workers sharing the $1 is approximately equal (margin $\varepsilon = 0.05$) between hired and not hired ($p < 0.001$).

As we show next, judgments of fairness, as well as other sentiments towards procedural issues, such as what information is used in algorithmic decisions, are highly contextual. The first context we consider is the tension between volitionality, direct relevance, and *implied* relevance, i.e., relevance which is not evident but implied by the increased accuracy of the algorithm. We study the impact of this tension on perceptions by considering three treatments: one where workers did not observe accuracy information, and two where they did (differing only in how large the accuracy differences were among the algorithms; 5% for small and 10% for large differences).

Recall that Algorithm 1 includes only features that are volitional and directly relevant, Algorithm 2 additionally includes features that are non-volitional, but still directly relevant, and Algorithm 3 also includes features that are neither, but exhibits a higher accuracy when this information was shown. We find that hired workers are approximately indifferent among the algorithms ($p < 0.05$ for margin $\varepsilon = 0.1$), whether accuracy is shown or not, both for explicit and implicit sentiments. Not so for workers who were not hired. As shown in Figure 4, non-hired workers who were shown model accuracy had a preference for the two algorithms which included non-volitional features, with a stronger preference for larger accuracy differences. This is precisely the opposite direction of the hypothesized impact in H4. However, explicit sentiments were similar for Algorithm 2 and Algorithm 3 when accuracy was shown (H5 is not supported). On the other hand, when information about accuracy was omitted, workers appeared nearly indifferent among the three algorithms (supporting neither H4 nor H5). Thus, our analysis of *explicit sentiments* does not support H4 in its original form, nor does it support H5.

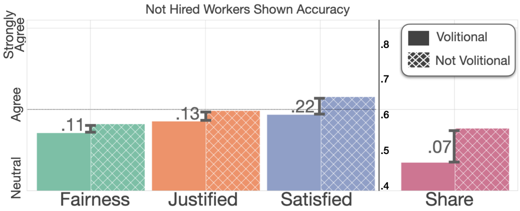

Figure 4: Average sentiment of not-hired workers shown model accuracy, partitioned by whether the hiring algorithm used only volitional features (solid) or used nonvolitional features (hatched). Sentiment differences are statistically significant for justified ($p < 0.05$) and satisfied ($p < 0.005$).

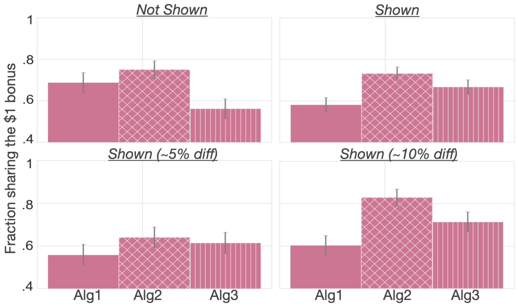

Figure 5: Fraction of not hired workers who share a nonzero amount of the bonus with selectors. These differences are significant for *Not Shown*: Alg2>Alg3 ($p < 0.005$), Alg1>Alg3 ($p < 0.05$), *Shown*: Alg2>Alg1 ($p < 0.005$), Alg3>Alg1 ($p < 0.05$), *Shown with* $\sim$*5%*: none, *Shown with* $\sim$*10%*: Alg2>Alg1 ($p < 0.001$), Alg2>Alg3 and Alg3>Alg1 ($p < 0.05$).

Considering *implicit* sentiments—the fraction of workers who chose a non-zero share of the final $1 bonus to give the selector—offers rather surprising additional insight, which we can glean from Figure 5. When accuracy information is not shown (Figure 5, left), workers who were not hired had a distinct implicit dislike of Algorithm 3 (which utilizes features that are not facially relevant to the task), compared with either Algorithm 1 ($p < 0.05$) or Algorithm 2 ($p < 0.005$). Thus, without accuracy information to suggest implied relevance, the use of such facially irrelevant features is perceived as undesirable, providing support for H5. On the other hand, Algorithm 2 was slightly preferred to Algorithm 1, albeit not to a statistically significant degree; the use of directly relevant features appears to outweigh their non-volitionality. When accuracy information is shown, implicit sentiments towards Algorithm 3 increase, while those towards Algorithm 1 decrease correspondingly, with implied relevance now playing an important role. Nevertheless, some reservations about implied relevance appear to remain, with Algorithm 2 still preferred over Algorithm 3. In any case, H4 is not supported in its original form for the implicit sentiments.

Overall we observe that while hired workers are relatively indifferent among algorithms, the relative sentiments of those not hired are highly sensitive to context. Throughout, however, volitionality of features is consistently secondary to relevance (direct of implied). Explicit survey results do not yield a clear preference for including features that are not directly relevant, but result in higher accuracy. However, implicit sentiments suggest reservations about including such features.

Perceptions of Workers with Disadvantaged Features Our analysis so far has focused on overall sentiments. However, this does not account for the possibility that sentiments meaningfully differ between people who have different values of the non-volitional features. Recall that our design included three features that are non-volitional, and thereby present significant fairness concerns: eyesight, age, and parent's occupation/income. The former two (eyesight and age) are perceived as being intuitively relevant to the task (see Section 2.2), and the latter is not, but ostensibly increases accuracy in our design. We now consider to what extent the perceptions of individuals with disadvantaged values of these features differ from the population average. In particular, *we hypothesize that these individuals will tend to have less positive sentiments in treatments using such features than the rest, as their use may seem to them particularly unfair (H6).*

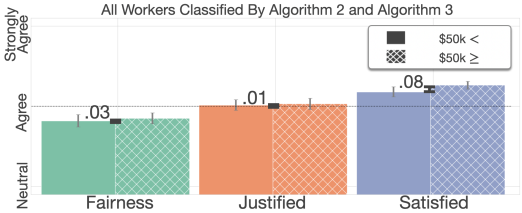

Figure 6: Explicit sentiments of workers divided by reported annual income. Only works classified with Algorithm 2 or Algorithm 3 reported their age. No sentiment difference is statistically significant.

Surprisingly, we find that H6 is not well supported in the case of parent's income and age. Specifically, participants with low-income parents exhibit only small difference in their sentiment compared to average (0.01-0.08, depending on the sentiment measure), and the difference is not statistically significant; see Figure 6. We find similar results in the case of age (Figure 12 in the Supplement).

A striking exception is eyesight. In this case, we find that workers reporting poor eyesight exhibit sentiments that are considerably lower than those reporting neutral or good eyesight, providing strong support for H6. In particular, the average sentiment difference between those with neutral or good eyesight, and those with poor eyesight was ~ 1.0 for each of the three sentiment types (this corresponds to a difference between "Agree" and "Strongly Agree", for example); see Figure 7. For example, these sentiment differences are larger than the differences between those of hired and not-hired workers, and that the sentiment difference between good-eyesight and poor-eyesight is even greater when only considering workers who are not hired. Each difference was statistically significant ($p < 0.005$) with the exception of the *fairness* sentiment if we only consider not hired workers ($p > .1$).

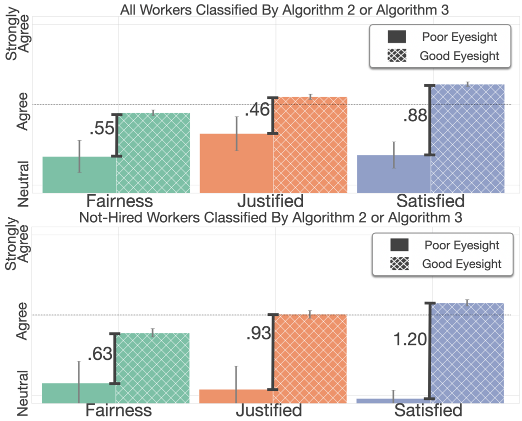

Figure 7: Sentiment of all workers (top), and not-hired workers (bottom), divided by reported eyesight; "Good Eyesight" indicates reports of neutral or better, while "Poor Eyesight" indicates reports of worse than neutral. Each sentiment difference is statistically significant at $p < 0.005$ level except fairness for not hired workers.

4 Discussion

The central takeaways from our analysis are two-fold. First, those in the managerial role of selecting workers were primarily focused on improving the accuracy of the selection algorithm, and considered that entirely fair along all dimensions. Second, negative sentiment about particular algorithms compared to others was limited to workers who were not hired; yet, preference was consistently for including features that are relevant even if they are not volitional.

Nevertheless, we now highlight important limitations of our study. First, as most human subjects experiments, it was low stakes. This has two motivations. First, it would be impractical to run an experiment of comparable complexity and size with significantly higher payments. Second, high payments can have an effect of implicit coercion, and would thereby pose a serious ethical concern. The consequence of small payments is that we do not know to what extent higher stakes would impact perceptions of fairness, and this is an important open question. More broadly, generalizability beyond our simple setting is an open issue. The key evidence that our results are likely to generalize is that they are broadly consistent with what we observe in the pilot survey as well, when we inquired about perceptions of volitionality and relevance of features abstractly: here we found near-perfect correlation between judgments of fairness and relevance, but volitionality is essentially uncorrelated with fairness. Indeed, our experiment suggests that the situation is more nuanced once real stakes are involved. Finally, while it may be tempting to draw simplistic conclusions that people do not care about volitionality, our results are in fact considerably more subtle, and this interpretation is unwarranted. Moreover, observations about general perceptions need not imply that our practice of algorithmic use must necessarily cater to these; ethical considerations may well transcend general perceptions—what is popular need not be the same as what is right.

Contribution Statement

Andrew Estornell and Tina Zhang contributed equally to this work†.

References

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[Albach and Wright, 2021] Michele Albach and James R Wright. The role of accuracy in algorithmic process fairness across multiple domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 29–49, 2021.

[Ali *et al.*, 2019] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30, 2019.

[Binns *et al.*, 2018] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. '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*, pages 1–14, 2018.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Camerer, 2011] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Eckel and Grossman, 1996] Catherine C Eckel and Philip J Grossman. Altruism in anonymous dictator games. *Games and economic behavior*, 16(2):181–191, 1996.

[Fiss, 1970] Owen M Fiss. A theory of fair employment laws. *U. Chi. L. Rev.*, 38:235, 1970.

[Fodchuk and Sidebotham, 2005] Katy Mohler Fodchuk and Eric J Sidebotham. Procedural justice in the selection process: a review of research and suggestions for practical applications. *The Psychologist-Manager Journal*, 8(2):105–120, 2005.

[Gelfand *et al.*, 2002] Michele J Gelfand, Marianne Higgins, Lisa H Nishii, Jana L Raver, Alexandria Dominguez, Fumio Murakami, Susumu Yamaguchi, and Midori Toyama. Culture and egocentric perceptions of fairness in conflict and negotiation. *Journal of Applied Psychology*, 87(5):833, 2002.

[Geyik *et al.*, 2019] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2221–2231, 2019.

[Greenberg, 1983] Jerald Greenberg. Overcoming egocentric bias in perceived fairness through self-awareness. *Social Psychology Quarterly*, pages 152–156, 1983.

[Grgic-Hlaca *et al.*, 2018a] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, pages 903–912, 2018.

[Grgić-Hlača *et al.*, 2018b] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Güth *et al.*, 1982] Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4):367–388, 1982.

[Hannák *et al.*, 2017] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933, 2017.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[Hort *et al.*, 2022] Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

[Houlden *et al.*, 1978] Pauline Houlden, Stephen LaTour, Laurens Walker, and John Thibaut. Preference for modes of dispute resolution as a function of process and decision control. *Journal of Experimental Social Psychology*, 14(1):13–30, 1978.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[Lakens, 2017] Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

[Lee *et al.*, 2019] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[Lemons and Jones, 2001] Mary A Lemons and Coy A Jones. Procedural justice in promotion decisions: using perceptions of fairness to build employee commitment. *Journal of managerial Psychology*, 16(4):268–281, 2001.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Merritt and Reskin, 1997] Deborah Jones Merritt and Barbara F Reskin. Sex, race, and credentials: The truth about affirmative action in law faculty hiring. *Colum. L. Rev.*, 97:199, 1997.

[Pierson, 2017] Emma Pierson. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*, 2017.

[Sánchez-Monedero *et al.*, 2020] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 458–468, 2020.

[Sunshine and Tyler, 2003] Jason Sunshine and Tom R Tyler. The role of procedural justice and legitimacy in shaping public support for policing. *Law & society review*, 37(3):513–548, 2003.

[Thibaut and Walker, 1975] John W Thibaut and Laurens Walker. *Procedural justice: A psychological analysis*. L. Erlbaum Associates, 1975.

[Thompson and Loewenstein, 1992] Leigh Thompson and George Loewenstein. Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes*, 51(2):176–197, 1992.

[Tyler and Huo, 2002] Tom R Tyler and Yuen J Huo. *Trust in the law: Encouraging public cooperation with the police and courts*. Russell Sage Foundation, 2002.

[Tyler, 2003] Tom R Tyler. Procedural justice, legitimacy, and the effective rule of law. *Crime and justice*, 30:283–357, 2003.

[Tyler, 2006] Tom R Tyler. Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol.*, 57:375–400, 2006.

[Vaccaro *et al.*, 2019] Kristen Vaccaro, Karrie Karahalios, Deirdre K Mulligan, Daniel Kluttz, and Tad Hirsch. Contestability in algorithmic systems. In *Conference companion publication of the 2019 on computer supported cooperative work and social computing*, pages 523–527, 2019.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[Wang *et al.*, 2020] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[Washington, 2018] Anne L Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17:131, 2018.

[Wellek, 2010] Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. CRC press, 2010.

[Wilson *et al.*, 2021] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

[Woodruff *et al.*, 2018] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.