People Perceive Algorithmic Assessments as Less Fair and Trustworthy Than Identical Human Assessments

LILLIO MOK, University of Toronto, Canada
SASHA NANDA, University of Toronto, Canada
ASHTON ANDERSON, University of Toronto, Canada

Algorithmic risk assessments are being deployed in an increasingly broad spectrum of domains including banking, medicine, and law enforcement. However, there is widespread concern about their fairness and trustworthiness, and people are also known to display algorithm aversion, preferring human assessments even when they are quantitatively worse. Thus, how does the framing of who made an assessment affect how people perceive its fairness? We investigate whether individual algorithmic assessments are perceived to be more or less accurate, fair, and interpretable than identical human assessments, and explore how these perceptions change when assessments are obviously biased against a subgroup. To this end, we conducted an online experiment that manipulated how biased risk assessments are in a loan repayment task, and reported the assessments as being made either by a statistical model or a human analyst. We find that predictions made by the model are consistently perceived as less fair and less interpretable than those made by the analyst despite being identical. Furthermore, biased predictive errors were more likely to widen this perception gap, with the algorithm being judged even more harshly for making a biased mistake. Our results illustrate that who makes risk assessments can influence perceptions of how acceptable those assessments are – even if they are *identically accurate* and *identically biased* against subgroups. Additional work is needed to determine whether and how decision aids should be presented to stakeholders so that the inherent fairness and interpretability of their recommendations, rather than their framing, determines how they are perceived.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Applied computing** → *Law, social and behavioral sciences.*

Additional Key Words and Phrases: algorithm aversion, fairness, bias, risk assessment

ACM Reference Format:
Lillio Mok, Sasha Nanda, and Ashton Anderson. 2023. People Perceive Algorithmic Assessments as Less Fair and Trustworthy Than Identical Human Assessments. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 309 (October 2023), 26 pages. https://doi.org/10.1145/3610100

1 INTRODUCTION

Advances in algorithmic systems could help bring about positive societal change, but their use as decision aids has sparked concerns about the fairness and trustworthiness of automated assessments. On the one hand, algorithmic predictions show promise in informing domestic violence arraignment decisions [13], Alzheimer's Disease risk assessments [21], and credit risk assessments via models that help determine the likelihood of loan defaults [66, 74]. On the other hand, a large body of

Authors' addresses: Lillio Mok, lillio@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Sasha Nanda, sasha.nanda@mail.utoronto.ca, University of Toronto, Toronto, Ontario, Canada; Ashton Anderson, ashton@cs.toronto.edu, University of Toronto, Toronto, Ontario, Canada.

2573-0142/2023/10-ART309 $15.00
https://doi.org/10.1145/3610100

literature has examined many shortcomings of algorithmic systems in terms of how they can exacerbate biases and fail to engender trust [43–45, 47, 57, 78, 79, 89, 95, 102].

As a result, recent work has investigated how people perceive the opportunities and risks of algorithmic predictions. In terms of the benefits provided by algorithmic predictions for computation-intensive tasks [1, 13, 85], people are *averse* to algorithmic advice even when it offers substantial performance improvements [33, 34, 53], although they occasionally over-rely on algorithms for quantitative tasks requiring expertise [49, 70]. In terms of societal risks, people are often concerned with the *fairness* and *interpretability* of algorithmic systems [16, 63, 64, 67, 75, 84, 89], although these perceptions too are often malleable [89] and misaligned [63, 64] with quantitative notions of fairness [72, 102]. These tensions between varying, often-misaligned perceptions of algorithms' performance and their societal desirability are especially salient when they are used to assess risk in high-stakes situations, such as criminal justice [43, 56], healthcare allocation [78], and creditworthiness [44]. Understanding how people perceive algorithmic assessments is therefore a key step towards ensuring productive, fair, and beneficial human-AI decision-making.

To what extent are assessments perceived differently based on the agent – algorithm or human – to whom they are attributed? Unfilled gaps remain in the literature on this question. Research on aversion often examines how much people behaviorally rely on individual algorithmic predictions in specific task scenarios, but does not consider attitudinal perceptions of fairness and trustworthiness [33, 34]. Other studies do analyze these stated attitudes, but typically either investigate general algorithm use without considering individual predictions [5, 63], do not reveal performance information to participants [5, 16, 63], or do not manipulate framing algorithmic assessments as human-made [16, 60, 69].

Thus, it remains unknown whether an algorithm's **perceived societal shortcomings**, such as unfairness, depend on the **individual assessments** it makes **compared to identical human assessments**. Without direct comparisons of identical assessments from humans and algorithms alongside performance information, it is difficult to understand if algorithmic predictions are considered unusable for e.g. humor detection [22] because people think that algorithms inherently should not be used for the task. Instead, people may avoid algorithms only because they speculate that algorithms would fail to achieve identical, human-like performance. We therefore seek to fill this gap in the present work. Our central research question is: (**RQ1**) *Are algorithmic assessments perceived to be more or less accurate, fair, and interpretable than identical human assessments?* In other words, we aim to compare how quantitatively-equivalent assessments are perceived depending only on whether an algorithm or human is thought to have made them.

Beyond how a risk assessment is being made, people are also known to react differently depending on *what* assessments are being made [33, 106]. This latter question has become increasingly salient with revelations of systematic social inequities in the outcomes of algorithms used in high-stake scenarios [20, 24, 78]. In addition to problematic algorithms (c.f. [107]), the data generated by people [51, 81] and communities [79] used to train these algorithms often contribute towards unjust outcomes that may harm various disadvantaged subgroups. However, the role that biased outcomes play in affecting how people perceive different types of predictor, algorithm or human, remains unclear. To investigate this, we pose a secondary research question: (**RQ2**) *To what extent do predictive biases impact the difference in how people evaluate algorithmic and human assessments?*

To address these research questions, we conducted an online 2×2 factorial experiment in which we asked participants to evaluate loan repayment risk assessments. Participants were shown the profiles of individual loan applicants and risk assessments that were attributed either to an algorithmic or a human predictor (first factor). These assessments were further either unbiased or biased against female applicants (second factor). Between these 4 conditions, participants saw otherwise *identical* loan applicant scenarios with *identical* outcomes. For each of the individual

loan applicants they considered, we asked participants to evaluate whether the predictor was accurate, fair, interpretable, and trustworthy in real life. Participants were also asked at the end of the experiment for their perceptions of the risk assessor after seeing all the loan applicants.

Summary of results. Across our study, we find that algorithmic risk assessments were systematically perceived to be less accurate, fair, and interpretable than identical assessments made by a human. These effects occurred as soon as participants saw a single erroneous assessment, and were amplified when errors arose as a result of systematic biases against disadvantaged subgroups. In comparison, participants evaluated assessments from a biased human and an unbiased algorithm similarly. In their open-ended responses to our questions, participants were also substantially more likely to praise the human than the algorithm for what they thought to be fair and accurate predictions; conversely, participants also expressed more apprehension over an algorithm they perceived as unfair than a human assessor.

Our study thus contributes causal evidence that different agents making identical predictions in identical circumstances can yield drastically different perceptions of how just the prediction is. This complements existing work on whether *any use at all* of algorithmic decision aids in various high-stakes domains is thought to be acceptable without performance information [5, 63]. Our results suggest that people perceive algorithms as unsuitable for certain tasks not only because the lack of this information leads them to speculate algorithms as yielding less accurate and less fair results than humans. Instead, people appear to place substantial emphasis on the algorithmic assessor itself even when its outcomes are identical to a human's. This provides support for the use of frameworks that do not focus only on outcomes, like procedural justice [65, 68, 92], to explain how people evaluate the suitability of an algorithm for making risk assessments.

2 BACKGROUND

Our work draws from two bodies of existing research on how people perceive algorithmic predictions. Firstly, the literature on *algorithm aversion* suggests that, depending on how quantitative and socially high-stakes predictive tasks are, people are more likely to place faith in human predictions over those made by algorithms. Secondly, we also consider the literature on how algorithms are evaluated for *fairness and interpretability*, particularly through the lens of observer perceptions. We then apply insights from this existing work to the domain of algorithmic risk assessments, which pervasively involve quantitative computation and are often used to guide socially impactful decisions.

Algorithm Aversion. There are many tasks at which algorithms have objectively better performance than humans, who have numerous cognitive biases [46] and difficulties recalling simple quantities [42, 73]. Technological developments across many fields requiring complex calculations, ranging from criminal justice [13, 56] to healthcare [21], are therefore often coupled to algorithmic improvements [62, 85]. Despite this, there is a growing body of work indicating that people are averse to the use of algorithms, even in heavily quantitative, objective scenarios in which they excel [6, 30, 33, 34, 90]. One explanation is that people are more sensitive to flaws in algorithmic predictions and thus will asymmetrically prefer humans when seeing machine-made errors [33], especially when these errors are made early [76]. Furthermore, people are less averse to algorithms when they are perceived as autonomous, accurate, inherently capable, or human-trained [53]. Similarly, others find that perceptions of potential algorithmic risks, such as unfairness and privacy concerns, could hinder otherwise positive enthusiasm about adopting algorithms [5]. Recent evidence also suggests that the level of expertise attributed to decision-makers could further determine whether people are averse to or appreciate algorithmic decisions [47, 49], especially because people calibrate their reliance on algorithms by inferring their expertise in the task at hand [110].

Algorithm aversion may also depend on characteristics of the task: they are thought to be more reliable for objective tasks, whereas subjective tasks require humanness that algorithms do not possess [22]. For instance, algorithmic advice about song popularity is more acceptable than advice about specialist, scientific opinion [70], whereas algorithms are less trusted to detect humor and to plan weddings [22]. In the realm of recommendations, people are indifferent between news articles automatically recommended based on personal consumption patterns and those curated by journalists [99]. In some mathematical situations, people follow algorithmic suggestions even if they are incorrect [96]. Beyond tasks that involve only singular agents, people also varyingly avoid algorithms in multi-agent settings. Under cooperative circumstances, people's receptiveness of e.g. conversational agents depends on how anthropomorphically and competently they are represented [52, 55]. In adversarial contexts, people evade algorithms both when they provides beneficial advice against an adversarial human [39] and when they are themselves the adversary [38].

Encouraging the adoption of algorithms in situations where they objectively outperform humans is thus essential to obtaining better aggregate outcomes [85]. Both the framing of the decision-maker, via its expertise, human involvement, and autonomy [49, 53, 110], and the framing of the task, via its objectivity and risks [5, 22], are likely to change whether people trust algorithmic decisions. Furthermore, algorithms are more acceptable when people are afforded more control [34], given information about previous algorithm use [2], and shown how algorithms learn over time [11]. This body of literature on algorithm aversion and appreciation thus highlights the myriad factors influencing the adoption of algorithms for tasks in which they outperform humans.

Fairness. Although research on algorithm aversion reveals the complex reasoning behind adopting algorithms, it typically focuses on situations in which performance is the key desideratum. In contrast, many other factors determine whether decisions are desirable, such as whether they are fair, explainable, and just [16, 44, 60, 106]. From a theoretical standpoint, many mathematical metrics allow for algorithmic fairness to be directly quantified [72, 102]. For example, an algorithm may be considered fair only if its predictions achieve statistical parity across subgroups [24], if (dis)similar individuals are given (dis)similar outcomes [36], or if decisions are insensitive to minor counterfactual data changes [59]. Indeed, fairness problems may also be inherent to the data from which algorithms are developed, be the data social [79], lingual [17], or visual [20], and be it used for healthcare [78] or criminal justice [56, 57]. Thus, meeting appropriate fairness metrics is often thought to be a necessary (although insufficient) requirement for adopting algorithms. On the one hand, algorithms and data should withstand litmus tests against these metrics in theoretical and laboratory settings (c.f. [24, 56, 102]); on the other, models deployed in the field need to be audited *in situ* for fairness [20, 78].

Nonetheless, empirical research has shown that theoretical fairness definitions are often inconsistent both with each other and with *human perceptions* [24, 25, 58, 64] – akin to how performance and *perceived* performance are misaligned in the algorithm aversion literature. For example, perceptions of different fairness metrics often depend on whether protected attributes are salient [89]. People also rate algorithms as more fair when outcomes favor themselves, even in the presence of obvious biases against demographic groups [106]. Mathematical metrics could further entirely miss altruistic values that are difficult to quantify [64]. It thus stands to reason that combining fairness metrics with human intervention may reconcile inconsistencies with both. And yet, results suggest that human-plus-algorithm processes may not only fail to correct, but even exacerbate, unfair and inaccurate decisions [43–45]. Considering these results alongside algorithm aversion [5], one may therefore suspect that any use of algorithms at all may be considered unfair in certain scenarios. Evidence for this has been found in managerial tasks [63], in which the use of an algorithm, let alone its outcomes, for tasks involving more human-like qualities is considered less fair.

These findings illustrate the need to incorporate other understandings of fairness that do not focus solely on outcomes, such as procedural justice [65], philosophical accounts [15], and judicial discourse [8]. For one, the amount of control that humans have across a decision-making procedure plays a key part in determining whether that procedure is socially acceptable [23, 34, 65]. Similarly, other procedural characteristics such as democratic representation and outcome alignment with constituent preferences are desiderata for legitimizing platform governance [80, 97], as is the ability for stakeholders to contest decisions [3, 71].

A critical aspect of procedurally-just algorithmic decision-making is the extent to which algorithms can be made *interpretable* [65, 109], which expands on the idea that outcomes alone do not determine the suitability of a decision maker (c.f. [16, 65]). Empirically, interpretability is often studied by manipulating the complexity and black-box nature of algorithms [83], showing the features used visually [60] or numerically [111], or by providing *post hoc* explanations of algorithmic decisions [16, 61, 75, 88, 105]. However, despite increasing trust placed in algorithms via prediction calibration (i.e. deferring to an algorithm's suggestion [60, 111]), improved algorithmic interpretability leads neither to consistently improved performance in human-plus-algorithm settings [60, 69, 83, 111] nor better evaluations of an algorithm's correctness [83, 84]. Nonetheless, uninterpretable predictions are still *perceived* to be less understandable [67], just [16], and potentially even less fair [35].

Algorithmic Risk Assessment. Given the myriad factors influencing how acceptable algorithms are perceived to be, many of these phenomena are closely related to the use of algorithms in making high-stakes assessments of risk. For instance, automated risk assessment tools have been increasingly considered for adoption in the realm of criminal justice [28, 29, 31], leading to studies arguing variably that these systems may harm [4] or improve [40, 56] the fairness of judicial decisions. In healthcare, quantifying patient risk also plays a critical role in the distribution of treatment resources [9, 48, 103], algorithms for which are also known to exacerbate social inequities in existing care management practices [78, 104]. Financially, computational risk estimation is used pervasively to evaluate creditworthiness and allocate assets [27, 91], with scholars positing that issues with algorithmic justice can be addressed through recourse [100, 102].

Three aspects of risk assessments make it an area in which algorithm aversion and fairness concerns are particularly relevant. Firstly, risk measurement is inherently a quantitative, computation-intensive task that has traditionally been regarded as yielding objectively correct or incorrect outcomes for individual predictions [43, 44]. It therefore shares many characteristics with similarly quantitative and computational forecasting tasks, such as academic success prediction, in which algorithm-averse behaviors have been observed [33, 34]. Secondly, risk assessments are often used in situations like criminal justice where their aggregated predictions have the potential for serious societal implications, such as unfairness towards ethnic minorities [4]. Questions of fairness and justice are therefore especially salient because algorithmic decision aids can transform latent social inequities into tangible harms by assessing risk unfairly [56, 78, 104]. In turn, one may thus expect observers to be averse to risk assessments made by algorithms lacking the procedural transparency and accountability to contestation that humans have [3, 65, 71]. Thirdly, risk assessments in vastly different domains share a fundamental mechanism, i.e. computationally quantifying risk from historical and possibly biased data, and are ecologically pervasive [43]. This makes them an ideal object of study at the intersection of algorithm aversion and fairness.

Relation to this work. Taken together, the bodies of research on algorithm aversion, fairness, and interpretability illustrate the importance of empirically understanding human perceptions of algorithms, beyond the objective measurement of computational performance or the adherence to desirable mathematical properties. These perceptions are notably relevant for algorithmic risk

assessments, the high-stakes, computational nature of which lends to distinct concerns about their acceptability as aids for societal decisions. We thus situate our research question within key characteristics of the existing empirical work below.

(1) **Who Predicts vs. What and Why Assessments are Made.** The literature on algorithm aversion and appreciation is broadly concerned with how people compare predictions from human and algorithmic agents by manipulating *who* makes the prediction [22, 33, 34, 53, 70]. In contrast, the work on fairness and interpretability is focused on different mathematical or perceived aspects of algorithms and datasets, i.e. *what* or *why* decisions are made [44, 56, 60, 60, 69, 72, 78, 83, 84, 102, 106, 111]. Our work compares how people perceive fairness and interpretability depending on whether assessments are made by a human or an algorithm.

(2) **Performance vs. Actions vs. Perceptions.** Research on both algorithm aversion and fairness measures outcomes in terms of objective performance (such as accuracy), participant actions (such as choice calibration), and subjective perceptions (such as stated fairness and trust). The former two categories are often studied together with monetary incentives, and typically measure how well human-plus-algorithm predictions agree with known ground truths [33, 34, 43, 44, 83]. On the other hand, perceptions are self reported. In work on algorithm aversion, measured perceptions are generally of confidence, suitability, performance, and preferences [5, 22, 33, 34, 99], whereas in fairness and interpretability research measured perceptions are unsurprisingly of fairness and interpretability themselves [16, 63, 64, 75, 89], as well as related constructs like trust [67], consistency [84], and justice [16]. Our work crosses both groups by measuring perceptions of performance, trust, fairness, and interpretability with attention incentives.

(3) **Concrete Situations vs. Generic Use.** Both bodies of research typically tackle two types of questions:

  (a) *How are algorithms evaluated for* ***individual, concrete predictions*** *they make under specific situations* ***with performance information****?* To address this question experimentally, studies often use tasks such as "The relevant features are $x_1, x_2, ..., x_n$ from which the algorithm predicts Y" (algorithm aversion, c.f. [33, 34]), or "Defendant A is predicted to have Y% flight risk based on their features $x_1, x_2, ..., x_n$" (fairness, c.f. [43, 44]). Our work studies this category of algorithm use.

  (b) *Should algorithms be* ***used at all*** *to make predictions in broad scenarios* ***without performance information****?* To address this question, task templates often look like "An [AI/human editor] decides about fitness recommendations" [5] or "Jayln works at the customer service center. Based on past call recordings, the (algorithm/manager) evaluates his performance" [63]. Because these tasks are used to study whether algorithms should be used at all, the actual features considered, concrete predictions made, and performance achieved by the model or human are typically not revealed.

In summary, despite the rich body of empirical findings on algorithm aversion and fairness, we find little work that compares **perceptions** of **identical predictions** in identical risk assessment scenarios depending on **who** (i.e. algorithm or human) makes those predictions. Although this is closely related to two studies on the perceived fairness of *general* algorithmic use [5, 63], neither measures how people comparatively perceive fairness in identical, *actual* predictions made by algorithms versus humans. Without revealing the individual outcomes achieved and the assessors' performance information to study participants, one cannot tell if participants find algorithms to be more unfair because they have worse speculated performance or if they are thought to be inherently less fair than humans. Thus, our research question qualifies the extent to which the framing of risk assessments as natural or artificial can alter perceptions of their predicted outcomes in the

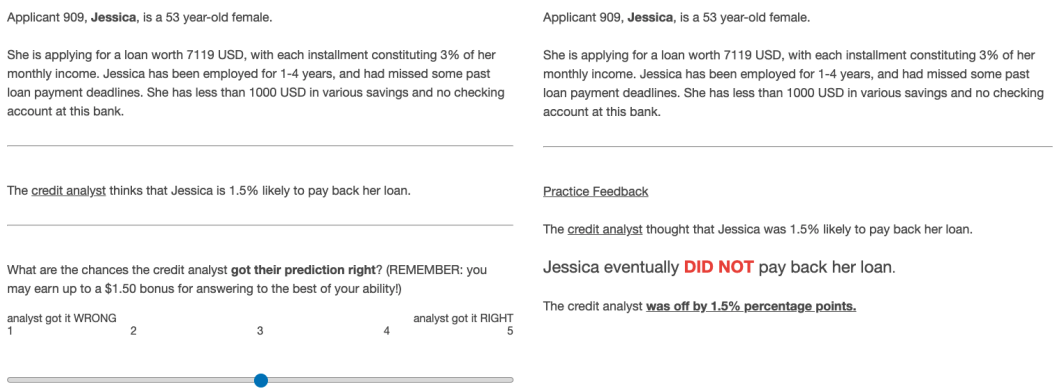

Fig. 1. *Left (**a**)*: Example task on first page showing a loan applicant profile and a prediction framed as being either from a statistical model or an analyst. Study participants are asked to guess the prediction's accuracy. *Right (**b**)*: Feedback about the prediction's actual performance is then shown to participants in the second page, along with in-task Likert scale questions on perceived fairness, interpretability, and real-world usability (Section 3.4). Each task consists of these two pages.

presence of performance information. This would, in turn, inform the design of better decision aid systems by identifying the factors that people consider when determining whether algorithmic systems are desirable and just.

3 METHOD

Our goal is to investigate the differences in how people perceive human and algorithmic predictors when they make the same risk assessment for the same scenarios. To address our research questions, we conduct a between-subjects, 2×2 factorial online experiment on Amazon Mechanical Turk ($n = 179$) in which participants are shown identical risk assessments that are framed as being made either by an algorithm or a human. In this section, we outline key aspects of our method: our task selection and construction, experimental manipulation, survey design, and outcomes measured.

3.1 Task Selection and Statistical Model

As with much of the literature presented in Section 2, our work is concerned with human- or machine-made predictions in response to a well-defined scenario with an objective, quantitative outcome. Following existing empirical work, we study risk assessment scenarios [24, 43–45, 56, 89, 102], which we operationalize by asking participants to respond to a loan compliance prediction task [44, 89]. Specifically, the prediction task asks: *given a set of features about a loan applicant, what are the chances that they will pay back the loan on time*? To this end, we use the German Credit Dataset [10, 102, 108], which describes one thousand deidentified loan applicants and whether they eventually repaid their loan[1]. This task is especially well-suited to our research question because its outcomes can be easily manipulated for gender biases (see Section 3.2) and also evaluated intuitively both on performance and fairness (see Section 3.4).

Data. We make two further modifications to this dataset. We first use an updated version that corrects for potential labelling errors in the previous dataset from existing work[2]. We then artificially de-bias the outcomes for gender by randomly reassigning binary gender labels to each row. In other

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[2] https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29

words, a coin is flipped for every participant that overwrites the existing gender label with a woman if it lands on a heads, and with a man if it lands on a tails. Loan applicants are therefore equally likely to be men or women independent of their other features. This is for multiple reasons. Firstly, this removes potential confounds in our experiment in which participants may react differently to loan scenarios based on an interaction between the predictor type (human versus algorithm) and gender-biased ground truths (e.g. that a certain gender pays loans back less frequently due to systemic social inequities, c.f. [78, 79]). Secondly, this allows us to vary the gender biases in our predictive agent's outputs for our research question, with less interference from obvious gender biases inherent to the dataset. Thirdly, this overcomes gender labelling errors in the original dataset[3]. In the altered dataset, there are neither gender differences in the outcome (loan compliance; $p > 0.05$ with unpaired t-tests) nor features (e.g. loan duration, loan amount, employment duration, checking account status, credit history, or age; all $p > 0.05$).

Statistical Model. We then trained gradient boosted trees [41] to predict loan repayment on the modified dataset, while excluding demographic variables (age, gender, and marital status) as is standard in the literature [44]. We evaluate the model's performance using the Brier score [44], for which we obtain a score of 0.17 on a 20% hold-out dataset[4]. To reaffirm the steps we took to minimize gender biases in the dataset, we also obtain a disparate impact ratio of 1.01, an average odds difference of 0.005, and an equal opportunity difference of 0.005 across the entire updated dataset [10, 102]. In other words, men and women were equally likely to be given a loan, had essentially the same true and false positive rates, and specifically had less than a percentage point's difference in true positive rates. Thus, predictions for loan compliance are virtually identical between the genders in the de-biased dataset.

We then construct loan applicant profiles using the variables in the dataset, actual loan repayment outcome, and model-predicted outcomes. We select age, gender, and 5 of the top 10 most predictive features from our model that are easy to fit into a short text description. These include the loan amount, percentage of income taken, employment status, credit history, and known bank accounts open. Following similar work on racial bias [14, 37, 98], we further generate obviously gendered names for the applicants by selecting the most popular and heavily gender-skewed names from a USA national names database[5]. An example applicant profile is shown in Figure 1.

Reintroducing Gender Biases to Predictions. We construct two versions of the model built on the gender de-biased data. We first create a *gender-biased* version by forcing negative predictions for all positive instances of female applicants. In other words, every female applicant who eventually paid back their loan is predicted as having a high risk of nonpayment. We set false negatives to have a mean predicted 5% chance of repayment with up to 2% noise to mimic natural predictions. Although financial institutions are often concerned with minimizing false positives, individuals are more concerned with fewer false negatives or overly-harsh predictions in loan scenarios [43].

[3]See link to updated dataset in previous footnote. Although this may impact previous work [102], our experiment by design requires ground truths with minimal gender biases and therefore uses artificial gender labels. We discuss the use of binary gender labels in Section 5.

[4]This score is equivalent to mean squared error between predicted probabilities and a binary actual outcome, and ranges from 0 to 1. For the model, we used the default settings available on the **scikit-learn** implementation. Note that our work is concerned more about resulting perceptions rather than developing novel, cutting-edge models; see e.g. [82] instead for improving predictive performance.

[5]https://www.ssa.gov/oact/babynames/limits.html; names are considered gender-skewed if more than 85% of people born with that name after 1970 fell into one gender. To avoid confounding from names that are ambiguously associated with ethnic minorities, we only consider the most popular names and check them against names in the literature on racial bias [14, 37, 98]. The exact names used are Amanda, Elizabeth, Emily, Jennifer, Sarah, Daniel, David, James, Joshua, and Matthew; one of Ashley, Jessica, Christopher, and Michael is randomly selected in the training example in Figure 1.

Applicant #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sex	M	M	M	M	M	F	F	F	F	F
Repaid Loan	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Predictions:										
Unbiased	TP	TP	FN	FN	TN	TP	TP	FN	FN	TN
Biased	TP	TP	TP	TP	TN	FN	FN	FN	FN	TN
Predictor:										
Human	*Analyst*	"	"	"	"	"	"	"	"	"
Algorithm	*Model*	"	"	"	"	"	"	"	"	"

Table 1. Experimental manipulations across the human vs. algorithm factor and the unbiased vs. biased factor. Blue cells represent correct predictions, i.e. true positives and negatives; red cells represent false negatives. Each participant sees the same 10 loan applicants in a randomized order.

Therefore, predictions with poor true positive parity and disfavoring (as opposed to favoring) certain subgroups should be perceived as less fair (c.f. [43, 78]).

We then construct an *unbiased* version with an identical error rate as the biased version. For the unbiased predictions, we take the same number of false negatives as the biased version and distribute them uniformly over the positive examples across both genders. In other words, both men and women applicants who paid their loans back are equally likely to be allocated a false negative. Empirically, we find that the equal opportunity difference to be 0.04 vs. 0.93 in the unbiased and biased versions respectively, indicating an extreme worsening of the true positive rate for women when they are the only recipients of artificial false negatives – i.e., that they would be incorrectly rejected for a loan[6] [10, 102]. By using our updated dataset, we can attribute this gender inequality to the predictions made as opposed to biases inherent to the original dataset – without altering the other variables in the dataset [102].

3.2 Experimental Manipulation

Our study incorporates a 2×2 factorial experimental design that manipulates the agent to whom risk assessments are attributed (i.e. a human versus an algorithm) and the extent to which predictions made are biased against protected subgroups (i.e. women, in our case). Each factor modifies the task described in Section 3.1 and Figure 1 as follows:

- **Assessor Type: Human Credit Analyst vs. Statistical Model**. As the main manipulation for answering our research questions, the predictor type factor replaces all occurrences of the predicting agent with either *"credit analyst"* or *"statistical model"* across our entire experiment. The pronoun we present for each predictor is respectively *"they"* and *"it"*.
- **Predictive Biases: Unbiased vs. Gender-Biased**. Aside from the tutorial in our experiment, all incentivized tasks will show predictions that are either gender-unbiased or gender-biased using the method described in Section 3.1, as shown in Table 1. Specifically, there are two levels:

(1) Predictions are *unbiased*: loan assessments for both men and women are equally likely to be false negatives.

[6]We further validate this by measuring the *calibration difference* between men and women, which is considered a desideratum for risk assessment tools [12, 26, 40]. We group repayment probabilities from the unbiased model into 5 bins and count the people who actually repaid their loans per bin. We find that the unbiased model is empirically well-calibrated for both groups with a mean calibration difference of 0.007 across the bins. In contrast, the biased model is extremely poorly calibrated with all females receiving low probability of repayment, therefore yielding a calibration difference of -0.463 in the first of two probability bins. No female fell in the upper probability bin.

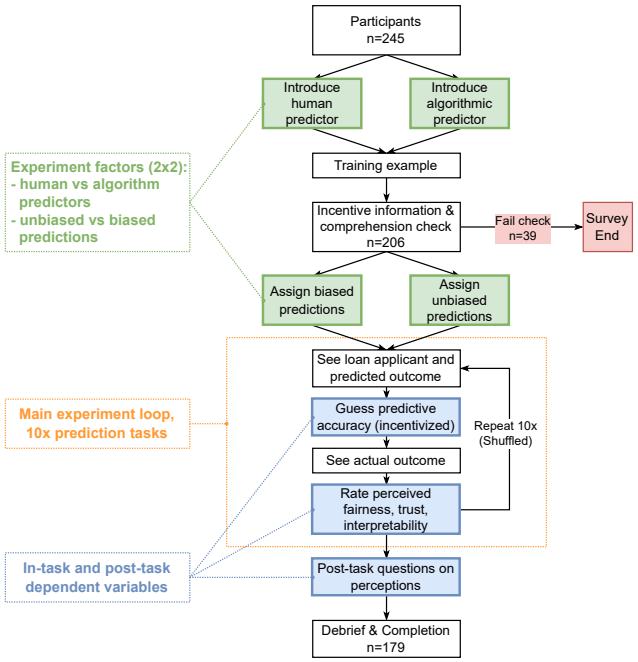

Fig. 2. Survey flow and experimental design.

(2) Predictions are *biased*: all loan assessments for women and no loan assessments for men are false negatives.

A summary of these factors is shown in Table 1. Besides our manipulations, every participant sees *identical* loan applicants shuffled into a random order. Only the predictor and its risk assessment varies between conditions.

3.3 Experiment Design and Workflow

Having constructed various loan applicant profiles for each of our experimental conditions, we now outline the rest of our online experiment implemented through Qualtrics shown in Figure 2. There are three phases in our design: an introduction and training example, a main experiment loop, and a post-task set of questions with debrief.

Introduction, training example, incentives. Participants are told that they will see a series of loan applicants receiving credit at a bank, and then introduced to the risk predictor – either a credit analyst or a statistical model – that will assess the likelihood of loan repayment. They are also informed that they will be tasked with evaluating the assessor.

Participants are then given an overview of the task plus a training example, which we randomly select from the four best-predicted applicants by Brier score (applicants are either male or female and either will or will not pay their loans back). One such example is shown in Figure 1a, which contains both an applicant profile plus a risk assessment from the predictor. We do not introduce any predictive errors in the training task. Participants are then asked to guess the predictor's accuracy in the example, after which feedback on the predictor's actual performance is shown (Figure 1b). They are prompted to respond to in-task evaluations on the feedback page, which we outline in Section 3.4.

Category	In-Task Question	Post-Task Question
Performance	What are the chances the [analyst/model] got [their/its] prediction right?	How accurate do you think the [analyst/model] was at guessing whether people paid their loans back?
Fairness	Do you think the [analyst/model]'s prediction is fair?	How fair do you think the [analyst/model]'s predictions were?
Interpretability	Is the [analyst/model]'s prediction understandable?	How much did you understand the [analyst/model]'s predictions?
Real-life adoption	Would you trust the [analyst/model] to decide who gets a loan in real life?	Think about banks you've used. Do you think your bank should employ this [analyst/model] to approve and reject loan applicants?
Open-ended	-	Do you have any thoughts about the [analyst/model] you saw?

Table 2. In-task and post-task questions asked to participants. All answers scored on Likert 5-point scales; open-ended question was only asked once at the end of the survey with a free-form text box.

After the training example, participants are shown the incentive structure: they can earn up to an additional $0.75 USD ($1.50) if they are in the top 10% (1%) performing participants, with total pay determined in Section 3.5. This improves participant attentiveness and data quality (c.f. [33, 34, 43]), and is evaluated based on their responses to the performance question in every task in the main loop as depicted in Figure 1a. To continue to the experiment loop, participants must then answer a comprehension check about the incentives correctly. This helps both response quality and participant salience of who the risk assessor is, i.e. analyst or model, and the incentives for answering to the best of their ability [33].

Main experiment loop. The main loop consists of ten "official" tasks (orange box in Figure 2) identical in format to the training example in Figure 1. For each question, participants are presented with the risk assessment scenario description, as well as either the statistical model or the credit analyst's prediction. They then judge how accurate the risk assessment is by answering the in-task performance question (Table 2 and Figure 1a), before being shown the ground truth. Finally, they answer the remaining in-task questions on the feedback page.

To select ten tasks, we choose the four positive applicants and one negative applicant for each gender with the best Brier scores in Section 3.1, excluding the applicants chosen for the training example. The individual manipulations are shown in Table 1. Participants in the human condition will be shown a "credit analyst" making risk assessments, while those in the algorithm condition will see a "statistical model". Those seeing an unbiased predictor will receive 2 false negatives for each gender, while those seeing a biased predictor will see 4 false negatives for women only. The ordering of applicants is again randomized to reduce anchoring and learning effects.

Post-task questions and debrief. After completing the previous phases, participants are asked a series of questions to measure outcomes for our research question. A full description is provided below in Section 3.4. We further ask participants for their voluntary disclosure of their gender, age, and education status, detailed in Section 3.5. Finally, they are debriefed about the experiment and told that the data, predictions, and predictor are fictional in line with existing work [106].

3.4 Measured Outcomes

In accordance with our research questions, we measure the extent to which people perceive the risk assessment predictor as being accurate, fair, interpretable, and trustworthy in real life by adapting

	Factor 1: Predictor		Factor 2: Bias	
Condition	Analyst	Model	Biased	Unbiased
Participants	96	84	88	92
Age	44.2 ± 9.9	41.9 ± 10.9	44.8 ± 10.3	41.6 ± 10.4
Years Post-Secondary	3.3 ± 1.9	2.9 ± 1.8	3.1 ± 1.9	3.1 ± 1.8
Is Female	35.4%	57.1%	48.9%	42.4%
USA Residence	100%	100%	100%	100%
HITs Completed	≥50	≥50	≥50	≥50
HIT Approval	95%	95%	95%	95%

Table 3. Demographics by experimental condition.

questions from existing work [5, 63–65, 84, 106]. Note that, in comparison to the interpretability of *post hoc* explanations [61, 105], this notion of interpretability is more closely related to how models are thought to be understandable based on their transparency [60, 83, 111]. In our case, we manipulate how the model is framed as human or algorithmic, as opposed to which and how features are visible. These outcomes are measured across two sets of questions during each ("in-task") and after all ("post-task") of the official tasks. There are thus 10 in-task questions and 1 post-task question per category, per participant. All questions are coded on 5-point Likert scales; in-task questions referred to the current task and ground-truth feedback, whereas post-task questions ask about participants' overall perceptions. For example, participants answer on a scale of 1 (the predictor got it wrong) to 5 (the predictor got it right) in response to the in-task performance question, which is incentivized for attentiveness. These questions are tabulated in Table 2.

3.5 Participants

We recruited 245 participants from MTurk, of which 39 failed the comprehension checks and 27 exited before finishing the experiment. The remaining $n = 179$ successfully completed the task. We required participants to be MTurk Masters residing in the USA and with at least a 95% approval rating over more than 50 HITs. Participants received $2 USD for participating in the experiment plus up to an additional $1.50 according to the incentive structure in Section 3.3, which we determined via the length of pilot experiments and the median USA minimum wage. In the final version that we launched, participants who completed the task took a median of 10 minutes (Q1 of 8 and Q3 of 15 minutes).

Demographics. In addition to restricting our pool to skilled MTurkers from the USA, we also asked participants to report their gender, age, and years spent in post-secondary education in line with existing work [63]. We aim, firstly, to help contextualize our studies with respect to participant demographics that may not be perfectly representative of the broader populace. Secondly, we want to probe for potential pre-existing attitudes towards algorithms, known to be correlated with age and education level [5, 70], without anchoring participants on explicit questions about algorithmic knowledge. The distribution of these demographics across both of our experimental factors are shown in Table 3.

We note that participant demographics are comparable across our experiment with the exception of gender in the human-algorithm factor. This appears to be a chance occurrence as our survey randomizes participants into conditions before any demographics are collected in the debrief. One may therefore be concerned that our results for the human-algorithm factor may be confounded by differences in how different genders respond to the questions in Table 2. To check for this, we test whether self-identified females and males ($n = 1$ person identified as non-binary) responded

differently to the in- and post-task questions differently. We find no significant differences between their responses ($p > 0.1$ for all questions using bootstrapped Welch tests), suggesting that differences in the responses for the human-algorithm factor do not arise from underlying gender disparities. Additionally, we find no noticeable correlational effects between age, years of education, and participant responses ($-0.12 < r < 0.10$, $p > 0.2$ for all questions).

This preliminary analysis illustrates that any significant results in our factorial experiment are extremely unlikely to arise from participant demographics and pre-existing attitudes towards algorithms. On the one hand, because we use a between-subject design, participants are equally likely to be allocated any of the 4 experimental conditions regardless of their attitudes. On the other, their demographics have no empirical association with how they responded to task questions.

4 RESULTS

We now detail the results obtained from our experiment, which we group into three subsections. Firstly, we illustrate that risk assessments framed as algorithm-made are systematically judged more harshly than those framed as human-made, i.e. that there is a *human-algorithm perception gap*. Secondly, we show that biased errors play an amplifying role in this relationship by negatively shifting perceptions to different degrees, depending on the predictor. Finally, we reinforce our results with a qualitative analysis of the responses to the experiment's open-ended question.

4.1 Algorithmic Assessments Are Judged More Harshly

To address our primary research question **RQ1**, we measure the outcomes for the post-task questions in Table 2 aggregated across participants in the credit analyst and statistical model conditions. For in-task questions, we take a per-participant average before aggregating responses in each of these conditions. We then estimate the effect of seeing an algorithmic predictor by normalizing with the mean and standard deviation of participants who saw a human predictor, which we plot in Figure 3.

We observe that participants who saw a statistical model's risk assessment were consistently more likely to judge its predictions negatively than those who saw a human's risk assessment. The model condition induced a negative effect across each perceived construct, as shown in Figure 3. In the post-task questions, participants rated the algorithm as having worse performance as a whole ($p < 0.05, d = -0.270$)[7], as well as being less fair ($p < 0.05, d = -0.329$), less interpretable ($p < 0.01, d = -0.407$), and less desirable for making real-life decisions ($p < 0.01, d = -0.532$). The substantial difference in perceived interpretability is especially surprising, given that we never manipulated any available information in our loan applicant tasks (Figure 1) even when altering biases in the risk assessments.

We further find evidence of the human-algorithm perception gap in the in-task questions. Three questions had significant, negative effects, with participants perceiving the algorithm as less accurate ($p < 0.05, d = -0.408$), less fair ($p < 0.05, d = -0.318$), and less trustworthy in real life ($p < 0.05, d = -0.340$). Although it was statistically insignificant ($p > 0.05$), being shown algorithmic assessments still had a negative effect on the assessments' perceived interpretability ($d = -0.306$). Thus, our results point consistently towards participants forming a more negative impression of credit assessments made by an algorithm than those made by a human analyst.

To probe the in-task questions further, we additionally measure how participants' in-task perceptions change as they see erroneous risk assessments. We aggregate participants' responses to each question before the first error (i.e. a false negative predicting they will default on a loan), between the first and second errors, and so on over all 4 errors in the experiment (see Section 3.2).

[7]Unless otherwise specified, p-values are derived from comparing bootstrapped Welch t statistics between experimental conditions [86], across the 4 measured categories of perceptions.

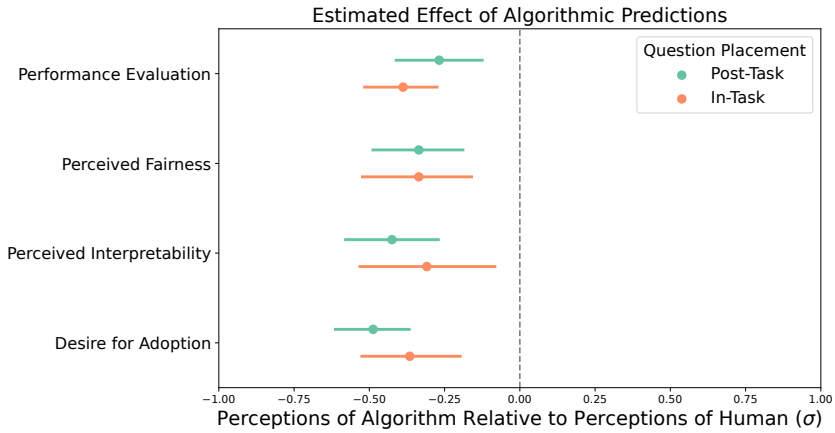

Fig. 3. Estimated effect of seeing predictions from a statistical model over a human analyst. Variables are defined in Table 2, and outcomes normalized so that the human-predictor condition has a standard deviation of one. Error bars show a 95% bootstrapped confidence interval.

Since the loan applicant ordering is randomized, each interval contained 2 questions in expectation. Figure 4 illustrates the in-task responses aggregated in this way for the analyst versus model condition; Figure 7 in Appendix A further splits this by the bias condition. We find perceptions to be similar in the model and analyst groups before exposure to any errors. The largest negative shift in perceptions occur universally after participants are exposed to their first error, echoing work on the importance of first impressions in helping develop trust in intelligent systems [76]. After the first error, in-task responses then separate over time: perceptions of the algorithmic assessor trend slightly negatively in comparison to perceptions of the human assessor. Since the order in which participants observe loan applicants is randomly shuffled across participants, this gap is unlikely due to artifacts such as learning effects. Instead, these shifts are likely due to seeing the consecutive errors we imposed on the risk assessments.

4.2 The Amplifying Effect of Erroneous and Biased Predictions

Based on the patterns in Figure 4, we further investigate the extent to which errors lead to shifts in participant perceptions of the risk assessor. This is further split by the biased-unbiased condition in Figure 7 in Appendix A. We focus on the effects of the first error for two key reasons. Firstly, all responses suffered from a systematic drop after the first error, illustrating the gap in participants' perceptions after they are shown flawed predictions. Secondly, the separation in responses between the algorithmic predictor and human predictor begins after the first error, suggesting that it plays a critical role in how participants' evaluations of the human and the algorithm diverge.

We first compare the responses in Figure 4 before and after the first error, the latter of which includes responses collected in the same task containing the first error. We find that, before the first error, there are no statistically-significant differences between how people viewed the algorithmic and human assessments ($p > 0.05, -0.132 < d < -0.034$). After the first error, however, effect sizes grew and became statistically significant. Perceived performance ($p < 0.05, d = -0.399$), fairness ($p < 0.05, d = -0.339$), and real-world trust ($p < 0.05, d = -0.353$) were all lower for participants who saw the algorithmic predictor compared to those who saw the human analyst. Although statistically borderline ($p = 0.086$), participants also comparatively thought the statistical model was less understandable than the human after the first error ($d = -0.301$). This again reinforces the

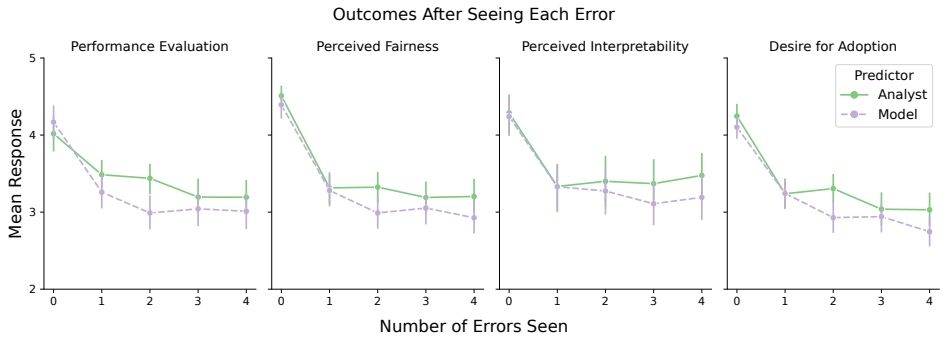

Fig. 4. Average in-task perceived performance, trust, fairness, and interpretability as a function of the number of errors participants have seen so far. Empirically, participants saw responded to a mean of two questions per perception category between errors.

findings in Figure 3, and further echoes existing work on how erroneous algorithms may lead to overly-harsh judgments of their predictions [33].

Does this effect depend on the *kinds* of errors that algorithmic and human predictors make – especially those that strongly disfavor certain subgroups? We thus extend our analysis of the first error to include our secondary **RQ2** by considering the biases underlying risk assessments. To this end, we calculate a before-after delta in the responses for each participant, formally $\Delta_i = q_{ai} - q_{bi}$ for participant i and in-task question q that is answered before (b) or after (a) seeing the first error. Since participant responses all systematically shifted negatively after the first error in Figure 4, we expect the deltas to be negative. Figure 5 plots the per-participant deltas, separated by whether they interacted with a human or algorithmic risk assessor, and whether they were exposed to unbiased or biased errors.

We find that exposure to biased false-negatives (and therefore biased true-positives, see Table 1) led to a statistically-significant widening of the gap between human-made and algorithm-made predictions for **every question**. The algorithm is punished substantially more for making a biased error through its perceived performance ($p < 0.01, d = -0.709$), fairness ($p < 0.05, d = -0.430$), interpretability ($p < 0.05, d = -0.846$), and desirability for real-life use ($p < 0.05, d = -0.464$). In comparison, when participants saw an unbiased error, there were no statistically-significant differences between perceptions of the human and the algorithmic predictor ($p > 0.05$ with $-0.350 < d < -0.108$). This effect suggests that the human-algorithm perception gap is closely tied to the biases in their predictive errors. The algorithmic predictor is judged much more harshly than the human predictor when both make the same biased mistakes, whereas judgment is less asymmetric when they make unbiased mistakes.

In tandem with our findings in Section 4.1, these results again highlight the differences in how people perceive human-made and algorithm-made risk assessments. Algorithmic predictions are evaluated more harshly than identical ones made by humans under the same circumstances and with the same information. When algorithmic predictions are erroneous, they are punished even more for displaying underlying biases against demographic subgroups – whereas participants judge humans less harshly for making the same biased mistakes. In fact, we find that a *biased human* and an *unbiased algorithm* are perceived comparably across all 4 post-task questions in Table 2 ($p > 0.05, -0.174 < d < 0.151$). We even observe that a biased human appears to be perceived as slightly more acceptable than an unbiased human across all 4 constructs after seeing the first error ($0.132 < d < 0.340$), although this is statistically borderline ($0.02 < p < 0.22$). These

Fig. 5. The shift in responses to in-task questions after seeing a single error, split by the different conditions to which participants are randomized. A negative shift indicates that participants viewed the risk assessments more harshly after the first error.

findings potentially suggest that human assessors are perhaps expected to make biased mistakes; we speculate on potential implications below in Section 5.

4.3 Open-Ended Responses

We further explore whether the gap in perceptions of humans and algorithms is reflected in responses to our open-ended question in Table 2. To this end, one author performed iterative, open coding on the responses to identify emergent analytical themes. Another author then coded the responses independently, after which both coders discursively identified four agglomerated themes: "poor performance" (mentions of general poor performance or drastic errors), "unfair" (mentions of bias against subgroups or sexism), "uninterpretable" (mentions of unintelligible reasoning or outcomes), and "positive thoughts" (thoughts opposing the other 3 themes as well as general praise). Themes are non-exclusive and may occur together in a given response. The resulting labels had strong inter-coder agreement (Cohen's $\kappa = 0.79$). We plot the number of responses with agreement from both coders in Figure 6 as a percentage of how many responses received labels in each condition for the model-human experimental factor[8]. In total, 34 out of 46 non-empty responses for the model and 44 out of 52 for the analyst received labels with agreement from the coders, representing an aggregate labelled response rate of 44%. The remaining responses contained little information, such as "no comment". Although this response rate is fairly low, we present these qualitative responses as motivating examples of the nuances in our participants' perceptions of the risk assessor. Information about quoted participants is available in Table 4.

Fairness and interpretability. We first note that the open-ended responses reinforce our quantitative results illustrating that the model is perceived as less fair than the analyst. The statistical model is thought to be unfair against women substantially more often than the analyst, with e.g. P149[9] stating *"It seemed to be most strongly biased regarding gender"*. P78 went further, positing that the *"disparity between the model with respect to women in comparison to men was striking. In no circumstance did the model predict a woman to pay back a loan regardless of their situation (e.g., employment, salary, loan history). And with very few exceptions, the model predicted men to pay back the loan. To me this suggests a strong gender bias"*. P166 was blunt about the model's biases: *"It was a pretty sexist"*. Compared to the 8 (24%) responses from participants in the model condition, only 4 (9%) from those in the analyst condition mentioned unfairness – of which 3 were tenuous: *"seemed*

[8]For additional information, we split responses by the unbiased-biased factor in Figure 8 (Appendix A).
[9]Participant identifiers are assigned before the comprehension check and thus may exceed $n = 179$.

Participant	Predictor	Biases	Participant	Predictor	Biases
4	Analyst	Biased	146	Model	Biased
8	Model	Biased	147	Analyst	Unbiased
16	Analyst	Biased	148	Model	Unbiased
19	Analyst	Biased	149	Model	Unbiased
20	Model	Biased	159	Model	Unbiased
34	Analyst	Unbiased	161	Analyst	Biased
35	Model	Unbiased	166	Model	Biased
38	Analyst	Unbiased	167	Analyst	Unbiased
60	Model	Biased	174	Analyst	Biased
64	Model	Biased	182	Model	Biased
65	Analyst	Biased	186	Analyst	Biased
78	Model	Biased	189	Model	Biased
84	Analyst	Biased	197	Analyst	Unbiased
99	Analyst	Biased	198	Analyst	Unbiased
106	Analyst	Biased	262	Model	Biased
115	Analyst	Biased	277	Analyst	Unbiased

Table 4. Participants with featured responses in Section 4.3 and their assigned experimental conditions.

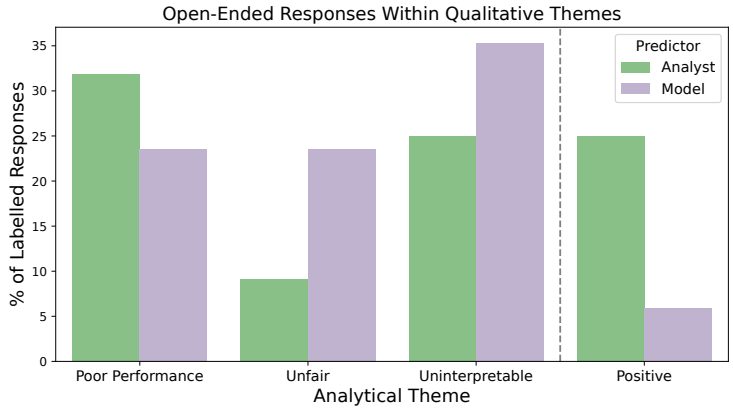

Fig. 6. Number of open-ended responses mentioning each analytical theme on which both coders agree, normalized by the total number of responses per condition that received at least one analytical theme.

a bit biased against women" (P84), *"I think they had some bias in certain cases but I am not totally sure why. It might have been based on age or gender?"* (P147), and *I think maybe the analyst didn't year [sic] women applicants fairly"* (P4).

We find similar discrepancies in the perceived interpretability of the model versus analyst, though participants were less assured in their criticisms because of their epistemic uncertainties. With regards to the model, 12 (35%) participants stated that the model's predictions were hard to understand, such as *"I wasn't sure exactly how it came to its conclusions, as it wasn't really clear."* (P159), *"I think that it was rather confusing"* (P182), and *"I can't figure out what made it get wrong the people that it did"* (P189). The 11 (25%) participants who mentioned they could not understand the analyst also had very similar responses, e.g. *"I could not figure out how the analysis [sic] was calculating its decisions"* (P115), but this represented a much smaller fraction of responses from those in the analyst condition.

Thus, responses containing thoughts about fairness and interpretability echo our quantitative results, with the participants being more likely to perceive the model in a negative way than the analyst. This is especially noticeable for those discussing fairness and sexism in the model, reflecting

our findings in Section 4.2 that biased errors amplify the perception gap between the analyst and the model.

Performance. Nonetheless, we also note that participants were more likely to explicitly ascribe poor performance to a human analyst (32%) than an algorithmic model (24%) in Figure 6. For instance, P16 stated *"I feel like the credit analyst wasn't very accurate. It frequently got things very wrong, even when the income and credit history of the person did not suggest they would fail to pay back the loan"*. Similarly, P16, P98, and P161 noted harshness in the analyst like *"I thought they were too harsh many of the times"* (P98), while P19, P34, P38, P167, P186, and P277 pointed out the inconsistency between the analyst's generally acceptable performance and occasional outliers e.g. *"I think at time it was right on, while other times it was way off"* (P186). Both are perhaps unsurprising given that our experimental conditions, by design, all incorporate obvious false negatives for 4 out of 10 loan applicants (see Table 1). Despite the relatively fewer participants in the model condition mentioning performance, similar sub-themes of harshness (P60, P146) and inconsistency (P8, P20, P64) also emerged in their responses.

There is therefore an apparent gap between our quantitative results about performance and participants' open-ended responses. We find that this could partially be because stated perceptions of the model's interpretability may override stated perceptions of its performance. While responses about not understanding the analyst were often interlaced with considerations of performance, e.g. *"The credit analyst seemed to be all over the place. I couldn't find any consistent patterns in the way their decisions were being made"* (P106), responses regarding the model's uninterpretability were short and essentially never mentioned performance like P189 above. Furthermore, the boundary delineating responses about performance and responses about interpretability was also blurred for coders – the poor-performance theme had the lowest inter-coder agreement ($\kappa = 0.67$ vs. $\kappa = 0.86$ for the 3 other themes collectively). Thus, the elevated uncertainty surrounding uninterpretable algorithmic predictions may in part account for why poor performance is mentioned more distinctly by people who saw the human analyst.

Praise. The clearest result we found in our qualitative analysis is the asymmetric praise given to the analyst ($n = 11$, 25%) versus the model ($n = 2$, 6%). Given that we only used false negatives in our experimental manipulation (Table 1), we had expected responses to be purely critical of either random or sexist loan denials. Instead, many praised the analyst, e.g. *"I think the analyst was fair and made the best decisions based on the information given"* (P197). In fact, two participants who saw the analyst provide biased predictions against females explicitly said the analyst was fair: *"I thought they were generally fair and accurate"* (P65) and *"I think he/she was very fair"* (P174).

In contrast, the two positive comments for the algorithm were hesitant, with P148 claiming *"The model was correct most of the time"* and P8 stating *"It seemed to be making very sound judgements most of the time, but occasionally very different from what I thought."* Criticism of the algorithm was unambiguous and harsh: *"the statistical model is not ready to be used, it cannot be trusted in real life"* (P262) and *"I thought it was bad. This is why humans are better. Why let models take our jobs?"* (P35). By comparison, the strongest critique we received about the analyst was more about job function rather than trust: *"he should go back to being a teller"* (P99).

Overall, our qualitative results again paint a picture of the divergent perspectives around risk assessments based on who makes the assessments. On the one hand, participants were more likely to explicitly describe algorithmic predictions as less fair and less interpretable. On the other, they were evidently more willing to praise the human analyst, despite it making identical predictions in identical scenarios. Strikingly, while several participants highlighted the biased algorithm's gender bias and even sexism, others actually praised the biased analyst for being seemingly fair. This

further reinforces our finding that biased errors amplify the gap between how algorithmic and human risk assessments are evaluated favorably by observers.

5 DISCUSSION

Our results reveal a common pattern in how people perceive risk assessments made by humans and algorithms. Despite yielding *identical* risk assessments, algorithmic predictions are perceived to be less fair, less interpretable, less accurate, and less trustworthy for real-life decisions than human predictions (**RQ1**). These effects materialize even after seeing just a single erroneous risk assessment, and are exacerbated depending on whether the error is a result of systematic predictive biases (**RQ2**) – to the extent that biased human assessments and unbiased algorithmic assessments are viewed comparably. Participants' asymmetric perceptions of human and algorithmic agents are reflected further in their open-ended responses, with far fewer expressing positive sentiments or praise regarding algorithmic assessments. In other words, our findings illustrate a *human-algorithm gap* in how negatively risk assessments are perceived when they treat subgroups inequitably.

These findings contribute to the bodies of empirical literature on algorithm aversion and fairness. On the one hand, an algorithms' mathematical properties and whether its outcomes are fair, understandable, and accurate can drastically alter whether people find it to be acceptable [5, 43, 44, 60, 64, 69]. On the other, an algorithm, *qua* algorithm, may not be trusted even for tasks in which it quantifiably performs better than humans [22, 33, 34, 53]. Our work illustrates that in identical, high-stakes risk assessment scenarios with the same information and outcomes, algorithms are still judged more harshly than humans – both in terms of objective performance, and subjective desiderata like fairness.

Our results thus indicate that more work is needed to understand how the framing of *who* makes a risk assessment can alter its acceptability. Existing research on fairness has identified the role of decision favorability [106], information on protected variables [89], different types of explanation [16] on perceptions of fairness. Studies on interpretability also show that feature visualizations [60], manipulating model transparency [83], and different types of explanations [67, 69, 75, 84, 111] can change whether people feel they understand algorithms. In contrast, our study suggests that the ways in which a prediction- or decision-making agent is framed can influence fairness and interpretability (alongside general suitability for risk assessment), even before people are shown different characteristics of what and how individual decisions are made.

The intervening effects of biases on the human-algorithm perception gap, as shown in Figure 5, further hint at how people perceive errors asymmetrically. Work on algorithm aversion has, for example, identified that errors in algorithmic predictions are possibly more salient to observers, leading to skewed preferences for predictions from humans [32, 33]. Our results extend this phenomena by providing evidence that a single error from a *biased* model can further widen the gap between human-made and machine-made predictions, not only of perceived performance but also of perceived fairness and interpretability. We additionally find patterns suggesting that a biased human and an unbiased algorithm are perceived comparably, and even that a biased human appears slightly more acceptable than an unbiased one.

One speculative reason is therefore that mistakes in human-made assessments are expected to be biased, such that errors made by machines are held to higher, non-human standards – not only for imperfect performance [33] but also for opaqueness [109] and, in our case, inequitable treatment of subgroups. With the systematic societal inequities across data-driven and algorithmic decision-making systems [8, 20, 78, 79], how, then, can we separate the perceptions of who makes decisions from the perceived unfairness of imperfect decisions themselves? After all, human predictions with clear inequities may be more acceptable than machine predictions with substantially fewer biases but identical error rates – a hypothesis for which our results provide some evidence. More work is

therefore needed to help people distinguish between "who" and "what" when evaluating flawed risk assessments against societal values like fairness and interpretability.

Beyond the framing of algorithmic predictions and the biases of individual predictors, our work further contributes more broadly to the discussion of how justice in algorithmic systems should be conceptualized. The differences in how predictions are perceived in identical situations and outcomes empirically highlights the importance of justice frameworks that are not purely outcomes-based or distributive [19, 36, 77, 87, 102]. Approaches like procedural justice [68, 92] may be poised to better capture user considerations of decision-making characteristics that exist in parallel to outcomes. For instance, the transparency of decision-making processes are not necessarily coupled to equitable outcome distributions, but can still influence perceptions of fairness [16, 65, 83, 84]. This is echoed by the interpretability effects we find in Figure 3 despite providing identical information to participants. Similarly, the amount of control human stakeholders can exert over both decision-making processes and outcomes also contribute to the acceptability of how decisions are made [23, 34, 65]. Beyond tasks like risk assessments with individually "objective", quantifiable outcomes, procedural aspects of more subjective decision-making scenario like content moderation [80] and discussion mediation [64] can also explain how people perceive the legitimacy of decisions. In context of these findings, our work therefore suggests that a human agent may be perceived as being more just by virtue of it – and its errors – being procedurally more suitable to risk assessments compared to an algorithm. Nonetheless, while procedural justice may *explain* the empirical degree to which observers find risk assessments to be acceptable, it may not necessarily *prescribe* how risk assessments can be made objectively more ethically desirable. We leave discussions of what constitutes moral decision-making processes to future work.

5.1 Limitations and Opportunities

Although we find clear differences in how human and algorithmic risk assessments are judged, our work has several limitations. Our method employs a narrow operationalization of "bias" relative to the plethora of ways in which algorithmic outcomes can be unfair [15, 24, 79, 102]. Even within the realm of gender equity, our study of biases through the lens of binary gender labels does not consider non-binary identities [50, 54, 93]. Although similar binarizing approaches are often used in related work [18, 45, 56, 78, 89], more work needs to be done on how inequitable human and algorithmic predictions across non-binary genders can elicit different fairness perceptions. For our current research questions, we believe our method adequately operationalizes how certain subgroups can be systematically disadvantaged by computational systems.

Another limitation is our study of risk assessments in isolated scenarios for predicting loan repayment with high-skilled crowdworkers. While this enables us to remove potential noise and interaction effects with other factors, such as the extent to which humans are involved in creating the statistical model [53], it also does not fully capture the real-life intricacies of decision-making based on risk assessments [78, 94, 95]. The physical deployment of risk-assessments as decision aids, let alone actual as actual decision makers, is fraught with complex ethical concerns [7, 78, 95], which may in turn amplify the human-algorithm perception gap in our experiment. Furthermore, beyond risk assessments, algorithms are used for an ever-expanding range of tasks as described in Section 2: tasks may be more subjective [22, 80], less socially impactful [76], require varying levels of expertise [49, 76, 110], and involve multiple agents that are cooperative [52, 55] or competitive [38, 39]. Follow up work is therefore needed to understand how well our results generalize to other situations in which algorithms are used to supplement or replace human judgment but do not have all the distinct characteristics of risk assessments.

More broadly, our work illustrates the effects of attributing predictions to different types of agents, but does not identify concrete ways of closing the human-algorithm perception gap. On the

assumption that predictions ought to be evaluated mainly for how they distribute outcomes [19, 77, 87, 102], attributing predictions to an artificial agent may distract from the merits or flaws that would otherwise be salient in a natural agent's predictions. Thus, finding concrete ways of reducing this human-algorithm perception gap is necessary to help stakeholders more clearly evaluate predictors for their predictions, as opposed to who or what the predictors are.

Nonetheless, our results also open multiple pathways for exploring how risk assessment systems can be designed to reduce this gap. For one, the literature we surveyed largely presents human and algorithmic predictions either to different observers separately in a between-subject manner [33, 34], to the same observers but as conceptual points of comparison in within-subject studies [63], or only as algorithmic aids in the decision-making pipeline [43]. It remains untested if algorithmic risk assessments can be *supplemented* with historical human predictions, e.g. by showing performance information to observers demonstrating that the algorithm is fairer towards protected subgroups than typical human experts. There are also promising procedural approaches that give observers more control over algorithms [23, 34, 65], though these have rarely been applied to risk assessments, and avenues of recourse over outcomes [71, 100, 101], though these have not been evaluated for their empirical, perceived desirability. Furthermore, because of the apparent salience of first errors in our study and in related work [76], interventions for reconciling perceptions of humans and algorithms should be targeted at erroneous predictions. This could take the form of a feedback loop after errors if they occur on a short time scale [110], or by identifying likely errors in low-confidence predictions [111].

5.2 Conclusion

Algorithmic risk assessments are increasingly deployed to predict averse outcomes in domains ranging from healthcare to criminal justice. In light of growing concerns over their potential for exacerbating existing social inequities in high-stakes decision making processes, many strands of research have attempted to develop ways of improving the inherent fairness of these automated systems. We focus on a more fundamental question: whether identical assessments would be perceived the same way if they were made by a human. Even though they are presented with the same information and yield the same outcomes, we find that human predictions are systematically thought to be more accurate, fair, understandable, and deployable in real-world systems. This perception gap materializes after just a single erroneous risk assessment and is exacerbated in the presence of identical biases against protected subgroups. Our work thus illustrates the need for a deeper interrogation of how quantitative fairness criteria and human perceptions should be balanced to build prosocial assessment systems.

REFERENCES

[1] Pankaj Ajit. 2016. Prediction of employee turnover in organizations using machine learning algorithms. *algorithms* 4, 5 (2016), C5.

[2] Veronika Alexander, Collin Blinder, and Paul J Zak. 2018. Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior* 89 (2018), 279–288.

[3] Ali Alkhatib and Michael Bernstein. 2019. Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.

[5] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[6] Hal R Arkes, Robyn M Dawes, and Caryn Christensen. 1986. Factors influencing the use of a decision rule in a probabilistic task. *Organizational behavior and human decision processes* 37, 1 (1986), 93–110.

[7] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*.

PMLR, 62–76.
[8] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[9] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs* 33, 7 (2014), 1123–1131.
[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.
[11] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering* 63, 1 (2021), 55–68.
[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.
[13] Richard A Berk, Susan B Sorenson, and Geoffrey Barnes. 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies* 13, 1 (2016), 94–115.
[14] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.
[15] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.
[16] 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*. 1–14.
[17] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.
[18] Sylvie Borau, Tobias Otterbring, Sandra Laporte, and Samuel Fosso Wamba. 2021. The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing* 38, 7 (2021), 1052–1068.
[19] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.
[20] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[21] Ramon Casanova, Fang-Chi Hsu, Kaycee M Sink, Stephen R Rapp, Jeff D Williamson, Susan M Resnick, Mark A Espeland, and Alzheimer's Disease Neuroimaging Initiative. 2013. Alzheimer's disease risk assessment using large-scale machine learning methods. *PloS one* 8, 11 (2013), e77949.
[22] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.
[23] Lingwei Cheng and Alexandra Chouldechova. 2023. Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.
[24] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[25] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[26] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research* 25, 4 (2016), 1692–1706.
[27] Xolani Dastile, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91 (2020), 106263.
[28] Sarah Desmarais and Jay Singh. 2013. Risk assessment instruments validated and implemented in correctional settings in the United States. (2013).
[29] Sarah L Desmarais, Kiersten L Johnson, and Jay P Singh. 2018. Performance of recidivism risk assessment instruments in US correctional settings. *Handbook of recidivism risk/needs assessment tools* (2018), 1–29.
[30] Dalia L Diab, Shuang-Yueh Pui, Maya Yankelevich, and Scott Highhouse. 2011. Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment* 19, 2 (2011), 209–216.
[31] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 7.4 (2016), 1.
[32] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.

[33] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
[34] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.
[35] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.
[36] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[37] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics* 9, 2 (2017), 1–22.
[38] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For what it's worth: humans overwrite their economic self-interest to avoid bargaining with ai systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
[39] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.
[40] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation* 80 (2016), 38.
[41] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
[42] S Connor Gorber, Mark Tremblay, David Moher, and B Gorber. 2007. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review. *Obesity reviews* 8, 4 (2007), 307–326.
[43] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
[44] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[45] Ben Green and Yiling Chen. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
[46] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2015), 1–20.
[47] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.
[48] Clemens S Hong, Allison L Siegel, and Timothy G Ferris. 2014. Caring for high-need, high-cost patients: what makes for a successful care management program? (2014).
[49] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[50] Samantha Jaroszewski, Danielle Lottridge, Oliver L Haimson, and Katie Quehl. 2018. " Genderfluid" or" Attack Helicopter" Responsible HCI Research Practice with Non-binary Gender Variation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–15.
[51] IM Jawahar and Jonny Mattsson. 2005. Sexism and beautyism effects in selection as a function of self-monitoring level of decision maker. *Journal of Applied Psychology* 90, 3 (2005), 563.
[52] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great chain of agents: The role of metaphorical representation of agents in conversational crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.
[53] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. (2020).
[54] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.
[55] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
[56] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[57] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.
[58] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[59] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).
[60] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.
[61] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).
[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[63] 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.
[64] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.
[65] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
[66] Martin Leo, Suneel Sharma, and Koilakuntla Maddulety. 2019. Machine learning in banking risk management: A literature review. *Risks* 7, 1 (2019), 29.
[67] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
[68] E Allan Lind, Tom R Tyler, and Yuen J Huo. 1997. Procedural context and culture: Variation in the antecedents of procedural justice judgments. *Journal of personality and social psychology* 73, 4 (1997), 767.
[69] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.
[70] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[71] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
[72] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[73] Lillio Mok and Ashton Anderson. 2021. The complementary nature of perceived and actual time spent online in measuring digital well-being. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–27.
[74] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* 165 (2021), 113986.
[75] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).
[76] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[77] Robert Nozick. 1973. Distributive justice. *Philosophy & Public Affairs* (1973), 45–126.
[78] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[79] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.
[80] Christina A Pan, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–31.
[81] Louis A Penner, John F Dovidio, Tessa V West, Samuel L Gaertner, Terrance L Albrecht, Rhonda K Dailey, and Tsveti Markova. 2010. Aversive racism and medical interactions with Black patients: A field study. *Journal of experimental social psychology* 46, 2 (2010), 436–440.

[82] Paweł Pławiak, Moloud Abdar, Joanna Pławiak, Vladimir Makarenkov, and U Rajendra Acharya. 2020. DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring. *Information sciences* 516 (2020), 401–418.
[83] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.
[84] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[85] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine* 380, 14 (2019), 1347–1358.
[86] Jenő Reiczigel, Ildikó Zakariás, and Lajos Rózsa. 2005. A bootstrap test of stochastic equality of two populations. *The American Statistician* 59, 2 (2005), 156–161.
[87] John E Roemer. 1998. *Theories of distributive justice*. Harvard University Press.
[88] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.
[89] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[90] Victoria A Shaffer, C Adam Probst, Edgar C Merkle, Hal R Arkes, and Mitchell A Medow. 2013. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making* 33, 1 (2013), 108–118.
[91] Naeem Siddiqi. 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Vol. 3. John Wiley & Sons.
[92] Lawrence B Solum. 2004. Procedural justice. *S. CAl. l. reV.* 78 (2004), 181.
[93] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for HCI researchers. *Interactions* 26, 4 (2019), 62–65.
[94] Megan Stevenson. 2018. Assessing risk assessment in action. *Minn. L. Rev.* 103 (2018), 303.
[95] Megan T Stevenson and Jennifer L Doleac. 2021. Algorithmic risk assessment in the hands of humans. *Available at SSRN 3489440* (2021).
[96] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.
[97] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.
[98] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.
[99] Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. 2019. My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism* 7, 4 (2019), 447–469.
[100] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.
[101] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596* (2020).
[102] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.
[103] Christine Vogeli, Alexandra E Shields, Todd A Lee, Teresa B Gibson, William D Marder, Kevin B Weiss, and David Blumenthal. 2007. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. *Journal of general internal medicine* 22 (2007), 391–395.
[104] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. 2021. Mitigating bias in machine learning for medicine. *Communications medicine* 1, 1 (2021), 25.
[105] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[106] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. 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*. 1–14.
[107] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's new clothes. *Medium* (2017).
[108] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*. 1–6.

[109] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology* 32 (2019), 661–683.
[110] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.
[111] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

A ADDITIONAL INFORMATION

In this appendix, we provide two additional sets of information for readers. Figure 7 disaggregates Figure 4 further by both experimental factors; our observations about the gap from a single error remain the same in this variant. Figure 8 presents labelled open-ended responses in the bias experimental factor, in which we find mostly expected patterns about the unfairness of biases against protected subgroups. The discrepancies in responses about poor performance are again attributable to their coupling to thoughts about uninterpretability, leading to the lower inter-coder agreement detailed in Section 4.3.

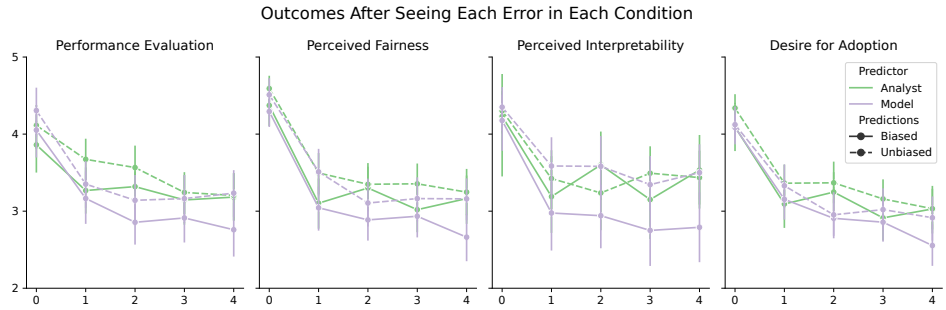

Fig. 7. Version of Figure 4 with in-task measures further split by the unbiased versus biased experimental manipulation.

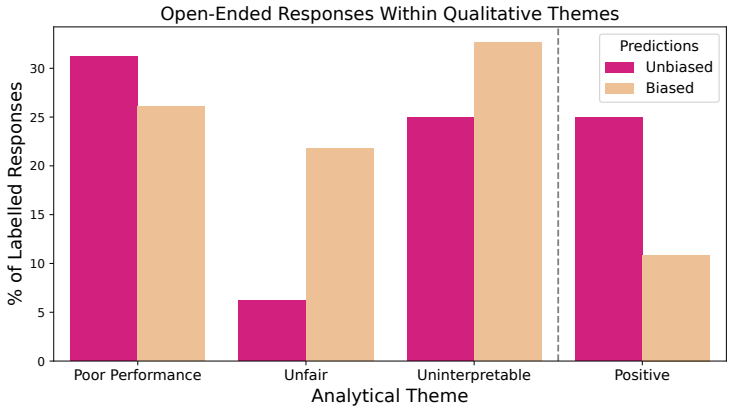

Fig. 8. Version of Figure 6 with responses split by the unbiased versus biased experimental factor, with normalizing denominators respectively of 32 and 46 labelled responses.

Received January 2023; revised April 2023; accepted July 2023