# My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning

Category: Theoretical and Empirical

**Abstract**—Machine learning technology has become ubiquitous in our society, but, unfortunately, often exhibits bias. For example, facial recognition software, systems used by judges in setting bail and sentencing, and banking and investment systems have all been shown to exhibit racist or sexist behavior. Thus, principle-based approaches to developing and deploying machine learning technology so that it aligns with ethical standards have become increasingly important. In particular, disparate stakeholders need to interact with and make informed decisions about machine learning models. Visualization technology can empower such decisions by supporting stakeholders in understanding and evaluating trade-offs between, for example, accuracy and fairness of models. This raises important open questions, including, "Can visualization design choices affect a stakeholder's perception of model bias and willingness to adopt the model?" This paper aims to empirically answer that question and to study how a layperson's trust in machine learning models is affected by the model's intrinsic characteristics (e.g., performance and fairness) and extrinsic factors (e.g., visualization design choices). Through a series of controlled, crowd-sourced experiments with more than 1,500 participants, we identify a set of strategies people follow in deciding which models to trust. Our results show that when making decisions, the visual design choices significantly affect people's likelihood to prioritize fairness over model performance. For example, participants value fairness more when it is explained using text than as a bar chart, and being explicitly told a model is biased has a bigger impact than showing past biased performance. We test the generalizability of our results by comparing the effect of multiple textual and visual design choices and offer potential explanations of the cognitive mechanisms behind the difference in fairness perception and trust. Our research guides design considerations to support future work developing visualization systems for models, and raises several concerns about how representation can be abused to misguide people.

**Index Terms**—machine learning, fairness, bias, trust, visual design, gender, human-subjects studies

## 1 Introduction

Data-driven systems that use machine learning (ML) are ubiquitous in today's society, spanning high-impact domains such as healthcare [41], banking [7, 55], hiring [60], and the criminal justice system [6]. Unfortunately, such systems can be unsafe and biased, including racist and sexist, which erodes people's trust. For example, IBM Watson recommended potentially fatal cancer treatments [58], cancer diagnosis systems have exhibited lower detection rates for people of color [76], software used by courts in setting bail have been found to have racial bias [6], and facial recognition systems routinely discriminate against women and people of color [16]. Such issues have led to legal bans of some types of ML systems [62, 63]. The U.S. Executive Office of the President has identified bias in software as a major concern for civil rights [24] and one of the ten principles Satya Nadella, the CEO of Microsoft, has laid out for artificial intelligence is that "AI must guard against bias, ensuring proper, and representative research so that the wrong heuristics cannot be used to discriminate" [49]. While extensive work focuses on reducing bias in ML algorithms [2, 8, 30, 45, 47, 66, 79, 80, 83], such methods often result in compromises, for example, sacrificing system accuracy for fairness, or requiring more expensive data or computational resources, thereby necessitating human involvement and complex decision making.

Visualization is one powerful strategy to inform such decision making [17, 38]. But, visualization design choices often affect how people reason [73], compare data values [27], infer about people [36], and draw causal conclusions [75]. Moreover, a model's perceived fairness and one's trust in a visualization are affected by visualization design [23, 46]. Similarly, the way ML models are described can impact people's trust in those models [77] and how they perceive model fairness [67]. As more data scientists and other stakeholders use visualization to support reasoning about ML models [17], the visualization community must study the effects of visualization on how people reason about ML models, including perceptions of model fairness and trustworthiness.

This paper addresses this underexplored space by empirically assessing how visual design, model performance and fairness, and user characteristics affect laypeople's trust in ML models.

To this end, and inspired by trust games from behavioral economics [20, 31, 84], we performed a series of experiments. We showed participants pairs of investment models (one fair and one biased), and they selected the model that they would invest (i.e., entrust) their money with. The commitment to invest serves as a proxy for trust, and the frequency of investment toward the fair model encodes the relationship between perceived model performance and fairness. We focused on the demographic parity aspect of fairness, i.e., the difference in positive outcomes across protected groups [21, 28], and considered gender-based biases.[1] This trust game-based instrument allowed us to analyze how people's trust in models can be shaped by visualization design choices, the models' performance and fairness, and user characteristics.

Figure 1 summarizes the seven research questions underpinning our experiments and their respective findings. Through detailed statistical analyses, we generate psychometric functions describing trade-offs in men's and women's perceived trustworthiness of a model based on its fairness and accuracy, and across visual representations and stakeholder-model relationships. We complement our statistical analyses with qualitative analyses of participants' self-reported reasoning strategies.

By exploring ML trust and fairness in a particular context, i.e., investment in the presence of gender bias, we synthesize five key insights and contributions of broad relevance to ML fairness visualization, and attempt to empower decision making. First, we provide empirical evidence that visualization design choices significantly impact people's prioritization of fairness over performance, influencing trust, as evidenced through detailed comparisons of bar charts, text conditions, scatterplots, and tables. Second, we demonstrate that men and women weigh accuracy-fairness trade-offs differently when provided with identical visual stimuli. Third, we show that an individual's relationship to the model (whether the model's outcome affects the individual or someone else) and explicit warnings of bias can impact trust more significantly than other factors. Fourth, we identify a set of strategies laypeople use when reasoning about ML models. And fifth, we translate our findings into a series of design recommendations for practitioners developing ML fairness visualizations and visual analytics tools.

[1] Gender is not binary. In this paper, we focus on bias against men and women. Future work should explore broader gender implications.

| | |
|---|---|
| RQ1: Do accuracy and fairness affect men's and women's trust differently? | Yes. Women trusted the fairer model more often than men, while men prioritize performance more. When the model's bias disadvantages their gender, the bias threshold that causes people to choose the more fair model was lower for women than men. |
| RQ2: Does making the decision on behalf of a client vs. oneself affect trust? | Yes. Participants tolerated more bias when deciding for themselves than when deciding on behalf of a client. |
| RQ3: Does model performance magnitude affect how much bias affects trust? | Slightly. For models with lower performance, participants trusted the fair model slightly less often, prioritizing performance slightly more. |
| RQ4: Does describing the models' history using textual and visual representations affect trust? | Yes. Participants trusted the fair model more often when its history was described using text than bar charts. Participants behaved similarly within multiple different textual and graphical representations, including orientation and color. Showing visualizations with more information led participants to trust the fair model more. |
| RQ5: Do demographics and personal characteristics affect people's behavior? | Yes, willingness to trust, behavioral inhibition and activation scores, and cognitive reflection test scores are all associated with differences in behavior when choosing a model. |
| RQ6: Does explicitly labeling a model as unfair (whether or not it is) affect trust? | Yes. Participants were less likely to select the model labeled as biased, even if that model was actually more fair. |
| RQ7: What strategies do laypeople follow in selecting which model to trust? | We identified seven such strategies. Some participants explicitly quantify and avoid a model's bias, while others ignore bias and rely on average performance instead. Others still prefer the model that historically preferentially treated others like them. |

Fig. 1. Our study answers seven research questions to understand people's trust in ML models.

## 2 Related Work

A rich body of research has studied visualization of ML models to support analyses [18, 35, 78]. Model fairness has emerged as a particularly important aspect of ML models to visualize [4, 17, 29, 38, 44, 48, 68, 71, 72, 81]. Our work empirically explores how design choices and model properties impact a person's trust in ML models, thus also contributing to understanding the design space of ML model visualization.

A growing interest in assessments of model fairness, and potential remediation of bias, has led to a broad array of toolkits and visual analytics systems. Microsoft's Fairlearn [11] and IBM's AI Fairness 360 [9, 37] implement several fairness metrics and learning algorithms that attempt to enforce fairness, and visualize a model's fairness and accuracy. Fairkit-learn [38] goes further by also visualizing the Pareto optimal frontier of a set of models with respect to model quality metrics, including fairness and accuracy. FairSight [4], FairVis [17], and SliceTeller [81] are visual analytics system that, too, incorporate model fairness in supporting decision making. The What-If Tool [70, 71] enables non-programmers to visualize datasets and perform counterfactual analysis and observe the effects of data changes on a TensorFlow model. It uses color encoding to overlay prediction correctness on scatterplots, histograms, and text-based representations of model statistics. Discrilens [68] visually explores model bias using causal modeling, developing novel set-based visualizations and D-Bias [29] is an interactive visualization tool for bias identification and mitigation for tabular datasets through causal models. FairRankVis [72] supports bias assessment of ranking algorithms. Visual design effectiveness depends on the task [72], user needs [71], user goals [68], and workflows [4]. Our work is complementary to each of these tools as insights into how visual design choices affect users' trust can potentially improve the tools' effectiveness. Importantly, too much information can overload participants and result in low quality decisions [54], so design choices can significantly affect people's ability to make good decisions.

People's perception of whether a model is fair depends on how information is represented: conveying data as scatterplots leads to a lower perception of fairness than using text [67]. Meanwhile, users' perception of fairness is affected by their characteristics (e.g., demographics, education, and computer literacy), the model's actual fairness, the textual representation describing the model, and the model's transparency and development process [69]. Our study focuses on assessing users' trust, beyond the perception of fairness.

Our study employs a trust game [10] to measure users' trust and gain insights into human reasoning. Trust games define trust as occurring when a trustor gives resources to a trustee with no enforceable commitment from the trustee [20]. For example, a trustor can lend their car to a trustee, knowing the risk that the trustee may not give it back. This formulation of trust and the use of trust games are common in behavioral economics [31, 84]. A typical trust game involves two anonymously paired participants: the trustor starts with money, some of which they may chose to give to the trustee. The experimenter triples the transferred money and the trustee can then return some portion back to the trustor [10]. The trustee can be simulated by a computer [5]. Giving more money indicates more trust; while altruistic behavior can also explain giving more money, altruism is typically not the cause of trust-like behavior [15]. Trust games can limit the two participants' decisions to a "trust" or "do not trust" decision, ensuring payoffs reward mutual trust but penalize asymmetric trust [43]. The day-trader investment task [82] is an interactive trust game of multiple trials, with participants taking part in a variant of a prisoner's dilemma task, testing group cooperation and trust [40]. Face-to-face interaction results in more trust than between players who never met [82]. Our study's trust game uses is a single round; the trustor selects between two models with which to invest money based on the models' history of returns.

The current state of visualization research across social sciences, computer science, and behavioral economics involves a variety of trust definitions that are not always tested for validity and can be inconsistent; using rigorously tested metrics to minimize bias and ensure repeatability can help [23]. One of the most common approaches to trust measurement in visualization research is asking participants to self-report on the Likert scale how much they trust a visualization or believe in its accuracy [22]. But self-reported measures are not always accurate and interpretations of scales can vary; trust games are a less subjective measure of trust [23]. Another approach to trust measurement is asking participants to select the best option and rate the trustworthiness of a visualization, but trustworthiness ratings often do not predict the selections, suggesting that self-reporting may not accurately measure trust [74]. For these reasons, our study uses trust games to more accurately measure trust.

Of the two closest papers to our work, one examined how ML model accuracy affects trust, showing that a difference between stated and observed accuracy reduced trust [77]. The other studied how participants of various races perceived models that discriminate against white and black people and found that human judges inspired more trust than models [34]. Our study focuses on understating the effect of accuracy, gender bias, and visual design on trust.

Finally, fairness is domain-specific [21, 66], and many definitions are mutually incompatible [26]. Our study uses one common definition, demographic parity, which requires the distributions of model predictions to be similar for the sensitive groups [21, 28].

## 3 Experiment 1: People's Trust in ML Models

We presented participants with pairs of ML models and asked them to select one model from each pair to invest with. We varied characteristics of these models, such as fairness and performance. We operationalized performance as the rate of return on investment and fairness as the difference between the return for men and women. In each pair, one model was generally more fair but had a lower average return than the other.

### 3.1 Study Design

The independent variables we examined were the visual representation (bar chart or textual description), the scenario (invest for oneself or on the behalf of a client), average model performance (low or high), the difference in average returns between the fair and biased model (biased model returns 10% more or 20% more, on average). We adopted a mixed-subject design for this experiment and counterbalanced our conditions such that half of the participants saw one level of every between-subject variable. We describe the manipulations for these variables next.

**Visual Representation (between-subject)**: We used two types of representations to communicate ML model performance: a textual description and a bar chart (orange boxes in Figure 4).

**Scenario (between-subject)**: We used two scenarios: in one, the participants made the investment on their own behalf, and in the other, on behalf of a client whose gender was unspecified.

**Model Performance (between-subject)**: We created two conditions of model performance based on the average return on investment from the fair model. The fair model either had high performance (return on investment of 50%) or low performance (10%). We chose 50 as substantial enough for the participants to make an informed decision between the fair or biased model, and 10 to allow for a condition where one gender lost money with the biased model, which returned less money than it was given.

**Difference in Average Return (within-subject)**: For each tier of model performance (high vs. low), we manipulated the differences between the average return on investment for the fair and biased model to be either small (10%) or large (20%). For example, for the small-difference condition, the high-performing fair model returned 50% to both men and women, while the competing biased model returned 40% to men and 80% to women, so its average rate of return of 60% is 10% higher than the fair model. We adopted a pseudo-staircasing method to generate the specific return values for the biased models. In perception research, staircasing methods involve increasing or decreasing the discriminability of a presented stimulus depending on the participants' response [50, 57]. Staircasing allows researchers to identify the just-noticeable difference between two intensities of a stimulus. In our study, we pit a fair model against a biased model that has a higher average return to observe the threshold for people to be willing to choose the biased model despite its bias. We vary the model's degree of bias by increasing or decreasing the differences in return to men and women. Tables 1–4 in the SM [1] show all the conditions we tested.

For all the experiments, participants completed several psychometric tests to evaluate the potential impacts of individual cognitive and personality differences on model selection:

**Cognitive Reflection Test** contains three quantitative questions shown to be correlated with quantitative reasoning ability [25]. A good CRT performance suggests that the participant took the survey seriously and possesses decent quantitative reasoning skills.

**Rotter's Interpersonal Trust Inventory** survey [59] consists of 25 statements that measure an individual's tendency to trust others.

**Behavioral Inhibition System (BIS)** measures sensitivity to punishment and motivation to inhibit behavior that results in negative outcomes. **Behavioral Activation System (BAS)** measures sensitivity to reward and motivation to encourage seeking the achievement of goals.

### 3.2 Procedure

We deployed the study using Qualtrics [56] and distributed it via Prolific.co [51]. The survey started with a consent form. Next, the participants completed the Cognitive Reflection Test (CRT) and the Interpersonal Trust Inventory. They then read a brief introduction to the survey and performed several rounds of a sample of our investment trust game. They were told each model's rates of return depended on the model's accuracy: The more accurate the model was, the higher average returns on investment it produced. One of the models provided the same rates of return on investment for both men and women. The other model was biased and provided a higher return rate to one gender, but its average return rate was higher than that of the fair model. We counterbalanced the gender towards which the model is biased, such that half of the time the model was unfair to men, and the other half to women. We neither explicitly informed participants of the model's average return nor bias, but showed the returns for men and women for each model in either text or bar chart form. Figure 2 shows a bar-chart version example.

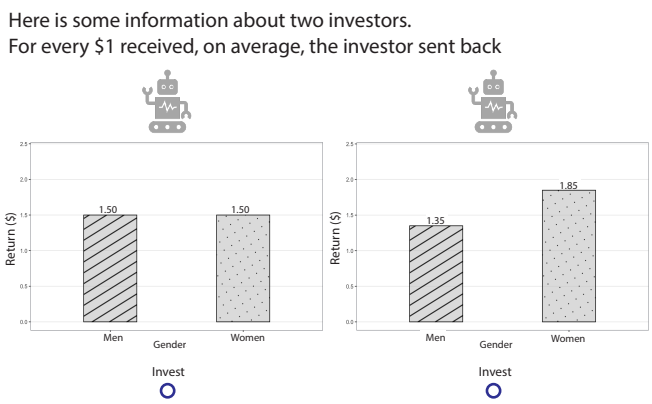


Fig. 2. An example question using the bar chart representation.

Next, participants completed 48 rounds of the trust game, covering the conditions outlined in Section 3.1. The rounds were presented in a random order. We recorded participants' investment choices for each pair of models. We also provided participants a free-response text box to explain their reasoning for 16 of the rounds, randomly distributed throughout the survey. See Section 7 for more details.

We included six attention checks throughout the survey. One attention check was related to their visual reasoning skills (look at a bar chart and select the tallest bar). The second tested their basic mathematical reasoning skills (how much money would they get in return if they invest $10 and the model returns $1.50 for every $1). The other four attention checks showed participants pairs of fair models, one providing a higher return than the other; to pass, the participants needed to select the higher-return model. We excluded participants who failed at least one attention check from our analysis.

At the end of the survey, participants completed the BIS/BAS inventory and reported their demographic information including age, gender, race/ethnicity, income, and education. Finally, participants reported how much effort they put into completing the study. They were assured that the answer to this question would not affect their compensation and were encouraged to answer honestly.

### 3.3 Participants

We recruited 1,347 participants for this experiment using Prolific.co [51]. Participants were compensated at $12 USD per hour. We filtered for participants who were fluent in English, over 18 years old, and reside in the United States. After filtering responses to remove attention check failures and nonsense, we were left with 1,326 participants (599 women, 608 men, 118 non-binary or indicated to prefer to self-describe, $M_{age}$ = 37.0, $SD_{age}$ = 13.2).

### 3.4 Results

We used R for all statistical analyses. We make public our data and R scripts: `https://osf.io/er5a3/?view_only=ce6801454c35476780d8591056f7c450` (hereon referred to as SM) [1]. We constructed two models to find statistically significant effects of our studied variables and their interactions:

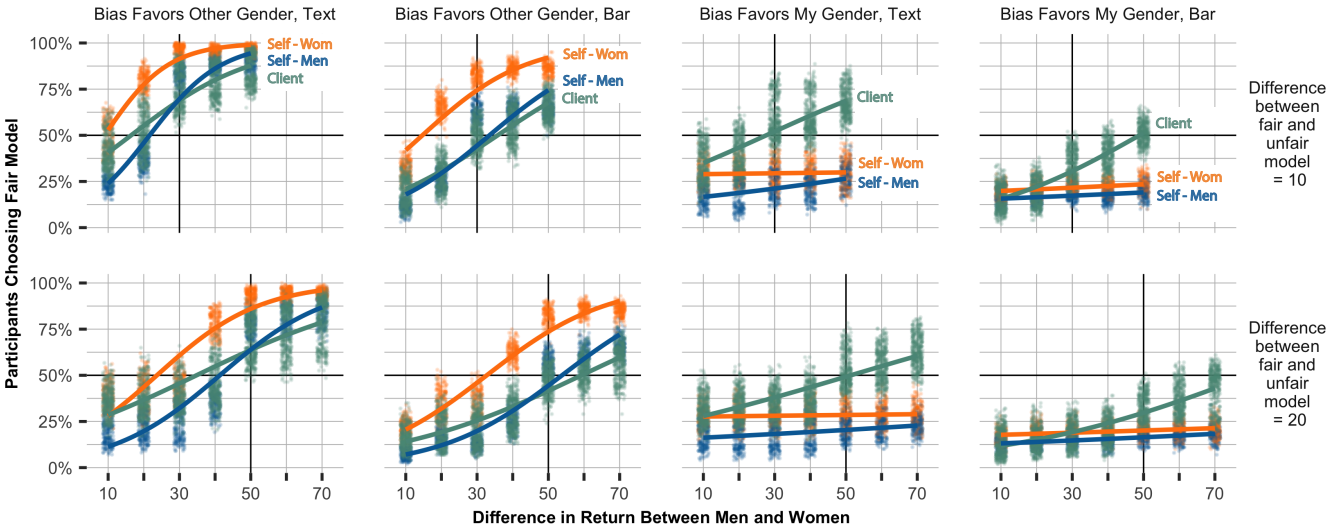


Fig. 3. Mean subset plots and logistic regression lines for bootstrapped results data. Data where participants are investing on behalf of a client is labeled "Client" (teal), and data where they are investing on their own behalf is separated by gender and labelled "Self - Men" (blue) and "Self - Wom" (orange). The X-axis represents the difference between returns to men and women by the biased model, and the Y-axis values represent the percentage of participants who chose the fair model for each bootstrapped data set. Subplots are separated by representation type (bar vs. text) along with whether the biased model returns more to the participant's gender or the other gender.

**Logistic Regression:** The first was a logistic linear model on the combined data from all studies from Section 3.1, filtered for responses from those identifying as men and those identifying as women. (Section 3.5 discusses results for those identifying as non-binary, other, or preferring not to disclose.) The dependent variable was the participant's choice to use the fair (1) or the unfair (0) model. The linear predictor formula included a set of predictors needed to answer research questions RQ1–RQ4, as well as all of their second-order interactions — gender, direction of bias (consistent or inconsistent with participant gender), scenario (whether or not participants were investing on behalf of a client), the return of the fair model and the maximum difference between returns for the biased model (used to describe model performance), and representation type (bar chart or text). We also included CRT scores, trust scores, BIS/BAS scores, and all collected demographics, including age, income, race/ethnicity, and education. Finally, this model included second-order interactions between gender and scores as well as gender and demographics.

**Bootstrap Re-Sampling:** To approximate a measure for uncertainty in the data, we performed bootstrap re-sampling and fit a linear model to the resulting data. We sampled the response data from those identifying as men and those identifying as women separately. For each of these two genders, we took sample size $N$ with replacement, where $N$ was the number of responses from that gender. We did this 100 times and aggregated the results across each of the above predictors, excluding all scores and demographics aside from gender. This aggregation provided us with percentage values of participants choosing the fair model, which we then used as the dependent variable for the model. The formula included the main effects and all second-order interactions of the variables across which we aggregated.

### 3.4.1 RQ1: Effects on Men's and Women's Trust

Figure 3 shows how men and women behaved when investing on their own behalf when using text and bar chart model representations.

On average, everyone was more likely to choose the model that gave their gender the higher return, which tends to be the biased model. However, women chose the fair model about 1.5 times more often than men ($\eta^2_{part} = 4.88 \times 10^{-4}$, $p < 0.001$). Overall, 48.9% of women chose the fair model, while 35.8% of men did ($SE = 1.25 \times 10^{-3}$).

When participants invested on their own behalf and the biased model favored their own gender, 17.2% of men and 26.4% of women chose the fair model. When the model was biased against their own gender, 49.7% of men and 68.7% of women chose the fair model. This pattern continued when participants chose on behalf of a client. When the biased model favored the participants' gender, 34.8% of men and 41.8% of women chose the fair model. When the bias was against, 41.5% of men and 58.5% of women did so.

We identified another asymmetry between women's and men's behavior. Recall that the biased model always generated higher average returns, so there are trials where a gender receives the same return from both the fair and the biased models, while the other gender gets an even higher return from the biased model. In these scenarios (e.g., fair: 50% to men and women, biased: 70% to men and 50% to women), men (58.8%) were more likely to choose the biased model than women (30.6%), even if their own gender was being discriminated against.

### 3.4.2 RQ2: Effect of Choosing For Yourself vs. a Client

We compare how trust in ML models changes when participants invest not for themselves, but on behalf of a client of an unspecified gender. We refer to these two conditions as two investment scenarios.

Participants were on average 3.25 times more likely to choose the fair model when investing on behalf of a client than themselves ($\eta^2_{part} = 7.74 \times 10^{-5}$, $p < 0.001$). They chose the fair model more often on clients' behalf when their gender received a higher return (38.3% vs. 21.8%), but less often on clients' behalf when their gender received a lower return (50.0% vs. 59.2%). But their gender significantly interacted with their tendency to choose the fair model depending on the scenario ($\eta^2_{part} = -2.30 \times 10^{-1}$, $p < 0.001$). Overall, 33.5% of men and 47.6% of women participants choosing on their own behalf chose the fair model, while 38.2% of men and 50.2% of women choosing on behalf of a client did so. Participants also became more likely to choose the fair model on behalf of a client as the bias of the biased model increased (Figure 3).

Additionally, in conditions where the fair model gives one gender the same amount as the biased model, participants were more likely to choose the biased one. Filtering for these conditions, estimated marginal means show that 33.7% of participants chose the fair model.

### 3.4.3 RQ3: Effect of Model Performance

We next considered how fairness-performance behavior trade-off changed when we varied the baseline performance of the fair model between high performance (50% return) and low performance (10% return). Participants were 1.41 times more likely to choose the fair

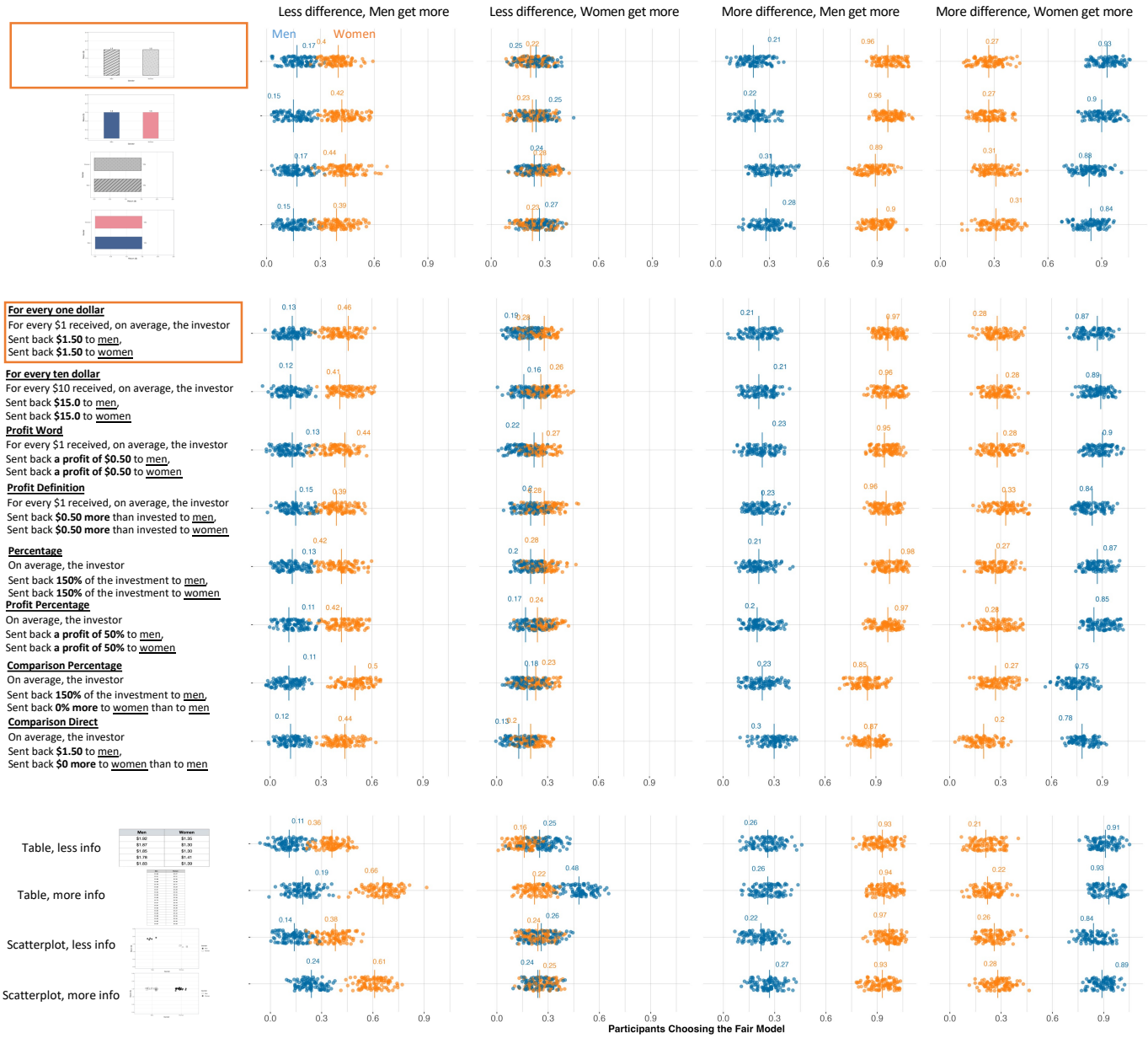


Fig. 4. The results along with the style of the 8 text, 4 bar chart, 2 table, and 2 scatterplot representations. The visualizations covered in orange box were used in Experiment 1. The dots in orange show the percentage women choosing the fair model, and the dots in blue represent men choosing the fair model.

model when the average model performance was 50% compared to 10% ($\eta^2_{part} = 3.10 \times 10^{-5}$, $p < 0.001$). This suggests that people in general care about overall model performance and are willing to tolerate more bias if the returns are low. Participants are 1.29 times as likely to choose the fair model when the difference between the fair and biased model is smaller (i.e., a difference of 10% vs. 20% in average returns) ($\eta^2_{part} = 3.14 \times 10^{-4}$, $p < 0.001$).

Figure 3 shows that when participants are disadvantaged by the biased model and choosing on their own behalf, they tend to choose the fair model at least 50% of the time once the fair model begins to offer the same or a higher return than the biased model. When the bias is advantageous, both men and women have no tipping point when investing for themselves. When choosing on behalf of a client, there is a more gradual increase in participants choosing the fair model as the discrepancy between returns grows. This suggests that many participants have a point where they begin to prioritize fairness over average model performance when considering the returns to others, but do not acknowledge this trade-off when choosing for themselves.

#### 3.4.4 RQ4: Effect of Visual Representation

We next investigated the effect of visual representation of model information on fairness perception and trust. Figure 3 shows the main effect of visual presentation ($\eta^2_{part} = 9.52 \times 10^{-4}$): participants were 1.48 times more likely to choose the fair model when the model information was presented as text (49.9%) compared to bar charts (34.8%). We also found a significant interaction between visual representation and model bias ($\eta^2_{part} = -2.30 \times 10^{-1}$). Although overall participants were more likely to choose the fair model when their own gender was being discriminated against (and vice versa), the tendency to choose the fair model was stronger in the text (63.3%) condition, compared to the bar chart condition (45.9%). This remains true when participants' own gender was being favored by the biased model.

#### 3.4.5 RQ5: Demographics and Personal Characteristics

We next assessed whether different participant demographics, along with personal characteristics such as trust scores, are correlated with different decision-making patterns. We found a main effect of trust

scores: participants were more likely to choose the fair model ($\eta^2_{part} = 5.09 \times 10^{-5}$, $OR = 1.01$) if they scored higher on the trust inventory. Participants were also more likely to choose the fair model when they scored higher on the BIS ($\eta^2_{part} = 8.38 \times 10^{-5}$, $OR = 1.01$) and BAS drive ($\eta^2_{part} = 1.07 \times 10^{-5}$, $OR = 1.03$) inventories, but less likely when they scored higher on BAS reward ($\eta^2_{part} = 1.36 \times 10^{-5}$, $OR = 0.99$). For CRT, participants who scored higher CRT scores were less likely to pick the fair model ($\eta^2_{part} = 4.60 \times 10^{-4}$). We also found an effect of age, education, level of income, and race/ethnicity, but the effect sizes were negligible. Details can be found in the SM [1].

### 3.5 Non-Binary and Preferred-to-Self-Describe Participants

We received 119 responses from participants who did not identify as men or women. Among them, 100 participated in the scenario where they invested on their own behalf and the average return was 50% for the fair model, and 60% or 70% for the biased model. We share some anecdotal results on this limited dataset to provide preliminary insights into how non men and women reacted to our experimental set-up. Future work will more closely and systematically examine these effects for more in-depth insights.

We performed the same bootstrap re-sampling as above for this condition set but included those identifying as non-binary, who preferred to self-describe, or who preferred not to disclose. The direction of bias was unclear to this group as the model information displayed in the trust game does not include a history of performance for non men and women.

The only linear predictors were gender and maximum return difference. Both gender ($\eta^2_{part} = 2.09 \times 10^{-3}$, $p < 0.001$, $OR\frac{women}{men} = 1.78$, $OR\frac{non\text{-}binary}{men} = 2.67$) and maximum return difference ($\eta^2_{part} = 4.45 \times 10^{-4}$, $p < 0.001$, $OR\frac{70}{50} = 0.694$) had significant main effects (but no significant interaction). We found a significant difference between how participants who identified as non-binary interacted with participants who identified as men and women. Recall that the fair model in this batch of data always returned a profit of 50%. Of non-binary participants, 64.3% chose the fair model, more often than women (56.2%) and men (41.2%). However, when filtering the men's and women's responses to those where the gender doing the choosing was not advantaged by the bias, women chose the fair model most often (78.9%), followed by non-binary participants (64.3%), and men (56.0%).

Qualitatively, non-binary participants reported sometimes being indifferent to their choices, as they did not identify with either of the genders that received bias benefits (see Section 7 for more details). Some non-binary participants used their assigned birth sex to make decisions. Others mentioned being against sexism in general. We further discuss these topics in Section 8 and share our insights on how to better account for their experiences and capture their responses in visualization and related research.

## 4 Experiment 2: RQ4: Bar Styles and Text Phrasing

Results from Experiment 1 (Section 3.4) suggest that people perceived models as fairer when the model information was presented as text rather than a bar chart, which joins recent explorations that demonstrated presenting the same data in different visual formats alters the perception of algorithmic fairness [67]. We next test the robustness of our observations on bar charts and textual descriptions across multiple visual styles of bar charts and alternative phrasing of textual descriptions. This also enables us to make a fairer comparison between the overall effect of bar charts and textual descriptions.

### 4.1 Design and Procedure

To keep the length of the experiment manageable, we chose four conditions that had the biased model either being a little biased (giving one gender 55% and another gender 65% return on investment), or extremely biased (giving one gender 35% and the other 85% return), with the gender to which the model is biased counterbalanced. The return from the fair model is kept constant at 50 for all conditions.

We tested 8 textual and 4 bar chart representations (including the ones used in Experiment 1) — see Figure 4. Previous work [3] has demonstrated that semantic associations, including those related to discrimination and sensitive information, impact color selection when constructing visualizations. We were concerned that the inverse might hold true, specifically that visualizing bias using colors with known semantic association might alter participants' decision-making. To explore this possibility, we ran an experiment where returns for men and women were visualized using blue and pink, respectively (a traditional North American color convention for gender associated with childhood and adolescence). We detail our rationale for selecting the alternative textual descriptions in the SM [1].

### 4.2 Participants

We recruited 413 participants from Prolific.co [51]. After applying the same exclusion criteria as that in Experiment 1, we were left with 410 participants (195 women, 209 men, $M_{age}$ = 36.78, $SD_{age}$ = 13.14). Half were assigned to the text condition and interacted with the text variations, and the other half were assigned to the bar conditions and interacted with the bar variations. Although we exclusively recruited men and women, we ended up with some participants who identified as non-binary ($N = 4$) and some selected 'prefer to not disclose' ($N = 2$). Overall, 318 participants reported to have put in a lot of effort in the survey, 90 reported having put in some effort, and 2 reported putting in very little effort.

### 4.3 Quantitative Analysis

We performed bootstrap re-sampling for this data with the same approach as before. For each bootstrap sample, we calculated the percentage of participants in the sample who chose the fair model for each of the four conditions outlined in Section 4.1, across 8 text and 4 bar chart representations. We performed an ANOVA test on the bootstrapped samples, comparing the percentage of people that chose the fair model across visual representation types (text or bar chart), and gender.

We found a main effect of visual representation ($F(1,9504) = 458.373$, $p < 0.001$). Participants were slightly more likely to choose the fair model with bar representations (42.71% chose fair, $SE = 0.07$) than text descriptions (40.95% chose fair, $SE = 0.05$) with effect size $\eta^2 = -2.16 \times 10^{-17}$ ($\eta^2_{part} = -1.29 \times 10^{-15}$). For the bar designs, as shown in Figure 4, participants behaved similarly across design styles that varied in color pallet and orientation (vertical bar chart with pattern: 42.7%, $SE = 0.13$; vertical bar chart with colors: 42.7%, $SE = 0.13$; horizontal bar chart with pattern: 43.3%, $SE = 0.13$; horizontal bar chart with color: 42.2%, $SE = 0.13$). For the text descriptions, participants also behaved similarly across most alternative phrasings ( *every one dollar (exp1)*:40.5%, $SE = 0.13$; *every ten dollar*: 42.3%, $SE = 0.13$; *profit word*: 41.1%, $SE = 0.13$); *profit definition*; 42.2%, $SE = 0.13$; *percentage*: 41.8%, $SE = 0.13$; *profit percentage*; 42.7%, $SE = 0.13$; *comparison percentage*: 38.9%, $SE = 0.13$; *comparison direct*: 38.0%, $SE = 0.13$.

We also found an effect of the unfair conditions (small vs. large discrimination) on fairness perception and choice. Participants noticed the trade-offs between model performance and fairness ($F(3,9504) = 58769.274$, $p < 0.001$). They preferred the fair model compared to the biased model that more drastically discriminated against one gender, despite that biased model generating a higher average return (e.g., 35% returned to one gender and 85% to another). In these conditions, 57.95% of the participants chose the fair model. Participants were more tolerant of the biased model that generated a higher return without drastically discriminating against one gender (e.g., 55% returned to one gender, and 65% returned to the other). On average, only 25.70% of the participants chose the fair model in these scenarios.

We observed a similar effect of gender as we did in Experiment 1 ($p < 0.001$). Men were less likely to choose the fair model overall (36.52%, $SE = 0.058$), compared to women (47.14%, $SE = 0.058$). We also saw an interaction between fairness conditions (e.g., 55%/65%, 35%/85%) and gender ($F(3,9504) = 117694.219$, $p < 0.001$). Overall, participants preferred choosing the fair model when the biased model discriminated against their own gender, as shown in Figure 4, where the

men and women data flipped between columns. Women seemed less willing to choose the biased model than men both in the case where bias was advantageous to their gender and when it was not.

The few participants who identified as non-binary or other chose differently for different conditions. They behaved similarly across different bar chart styles, where 31.2% chose to be fair on average. However, we observed that they tended to trust the fair model in the conditions where men get more especially if the difference in return between men and women was large (i.e., men return: 85%, women: 35%) where 66.7% chose fair, and for the 65/85 condition (men return: 65%, women: 55%) 29.2% chose to be fair. Whereas in the conditions where women get more (e.g., men return: 35%, women: 85%), they tended to trust the biased model more often (14.6% chose fair).

### 4.4 Discussion and Summary

The difference between bar and text reactions seems to generalize across small variations. Whether overall performance or fairness is preferred depended on the difference in return. In regards to gender, men, and women behaved differently: women seemed to be less willing to choose the biased model even if it favors women. But men seemed to be more willing to choose the biased model when it favors their gender. However, interestingly, men seem to also be more willing than women to choose the biased model that is biased against their own gender, which might be due to the social desirability bias [61, 64] as they refrained from expressing potential prejudices against women.

## 5 Experiment 3: RQ4: Scatterplots and Tables

We investigated the effect of scatterplots and tables on the perceived bias to examine whether other visual designs have similar affordances as bar charts or textual descriptions. We manipulated the number of data points shown to explore the effect of the amount of information.

### 5.1 Design and Procedure

We chose to test our conditions with different styles of scatterplots and tables as they were found, along with bar charts, as the top most used visualizations in order to enhance trust in ML models and effectively reduce potential visualization bias [19]. We adopted a 2x2 design for this experiment and showed participants two variations of tables and scatterplots (see Figure 4), either with 40 data points (20 for men, 20 for women) or 10 data points (5 for men, 5 for women). To make the results comparable to those from the bar and text conditions in Experiment 2, we kept the same four conditions of the biased model.

### 5.2 Participants

We recruited 207 participants for this experiment and after following similar exclusion protocol to our previous experiments, we ended up with a total of 205 participants. Among them, 96 identified as women ($M_{age}$ = 38.76, $SD_{age}$ = 14.62), 103 identified as men ($M_{age}$ = 34.83, $SD_{age}$ = 11.71), 4 identified as non-binary ($M_{age}$ = 29.25, $SD_{age}$ = 7.23), and 2 identified as other or prefer not to disclose ($M_{age}$ = 23.00, $SD_{age}$ = 4.24). In terms of Effort, 73.7% of the participants reported to have put in a lot of effort in the survey, 25.9% reported to put some effort, and 0.49% put very little effort.

### 5.3 Quantitative Analysis

We performed bootstrap re-sampling following a similar protocol to Experiment 2. For each bootstrap sample, we calculated the percentage of participants in the sample who chose the fair robot across 2 scatterplots and 2 table representations. We performed an ANOVA test on the bootstrapped samples, comparing the percentage of people that chose the fair model across visual representation types (scatterplot or table) and gender. We found a main effect of visual representation ($F(1,3168) = 20.333$, $p < 0.001$), such that participants more often chose the fair model after reading the table (44.43%) than the scatterplot (43.81%), although the effect size is small ($\eta^2_{par} < 0.01$). We also found a main effect of the amount of data shown in a visualization ($F(2,3168) = 1371.248$, $p < 0.001$). Participants chose the fair model more often when they read scatterplots and tables with more information (40 data points, 47.5% chose fair) compared to less information (10 data points, 40.7% chose fair).

We observed similar results on the trade-off between prioritizing performance and fairness. Participants were more tolerant of the biased model which had a small discrepancy between returns, but that tolerance decreased significantly once the biased model drastically discriminated against one gender (difference in return increased) despite generating a higher return. We observed overall the same effect of gender as we did in Experiment 2. However, we additionally uncovered a novel insight here with regard to the interaction between gender and the amount of information for the men: 65%, women: 55% condition (leftmost column of Figure 4). It seems that when more data is provided, while both the likelihood of men and women picking the fair model increases, the increase is larger for women where they were much more likely to choose the fair model coming across visualizations with more data points (62.6% women and 21.6% men chose fair) than visualizations with fewer (37.1% women and 12.6% men chose fair).

We found that the few participants who identified as *Non-binary* or *Prefer to Self-Describe*, tended to choose the fair model when coming across scatterplots (50.02% chose fair) as compared to tables (33.45% chose fair). They were more likely to choose the fair model in conditions when the biased model favored men (66.7% chose fair when the discrepancy between men and women return is large, and 54.2% when the discrepancy is small), and less likely to choose the fair robot when the biased model favored women (25.0% when the discrepancy is large, and 20.8% when it is small).

### 5.4 Discussion and Summary

In summary, participants tended to choose the fair model when coming across scatterplots and tables with more data. They found tables to be slightly more fair than scatterplots. Overall, when more data is provided in a visualization, while both the likelihood of men and women picking the fair model increases, the increase is larger for women.

## 6 Experiment 4: RQ6: Explicit Bias Warning

We explore how textual annotation warnings on models affect fairness perception and model selection given the rich literature illustrating the profound effects of textual annotations [12, 42, 53, 65]. For example, explicit textual warnings can mitigate bias in reasoning and decision making [53]. However, user interpretation of visualizations can be subconsciously affected by the presence of slanted titles [42]. More generally, the inclusion of graphic titles and text improves visualization recall [12]. Additional work has also revealed that the type of textual annotation and its positioning can alter what information viewers recall after viewing visualizations [65], and that users prefer graphics with greater textual annotations [65]. Based on the literature, we hypothesize that explicit warnings will substantially affect people's perception of model bias as evidenced by changes to their model selections.

### 6.1 Design and Procedure

This experiment follows the same procedure and design protocols as Experiment 2 (Section 4.1). Participants read bar charts and select either the fair or the biased model to invest in, across the same four combinations of discrimination values for the biased model as used in Experiment 2 (e.g., men get 65%, women get 55%), either for themselves, or on behalf of a gender-unknown client. The difference is that participants came across bar chart identical to those used in Experiment 1 (control), or a bar chart with an annotated textual warning above the biased model, or a bar chart with the annotated warning above the fair model - in a randomized order. Half of the participants invested for themselves while the other half invested on behalf of a client. The annotated warning read, "This robot is unfair to a specific gender". We compared participants' perception of fairness when using the default bar chart to make a selection, to that when using bar charts with explicit warnings. This set-up also allows us to account for the presence of warning overall, and compare the effect of warning alignment to generate insights with regard to how people react when the warning is misaligned with the actual model fairness.


Fig. 5. The results showing how people perceive bias when coming across pairs of models, represented using regular bar charts, and bar charts with or without explicit warning (saying it is biased towards a specific gender). The dots in orange show the percentage women choosing the fair model, and the dots in blue represent men choosing the fair model.

### 6.2 Participants

We recruited a total of 212 participants. After excluding the participants following the same criteria as previous experiments, we ended up with 209 participants. Among them, 101 identified as men ($M_{age}$ = 35.85, $SD_{age}$ = 13.41), 99 identified as women ($M_{age}$ = 39.43, $SD_{age}$ = 13.60), and 9 identified as Non-binary or preferred not to disclose ($M_{age}$ = 27.11, $SD_{age}$ = 5.80).

### 6.3 Quantitative Analysis

We found a significant main effect of the explicit warning ($F(2,4752) = 9550.6343, p < 0.001$). When there was no explicit warning of bias, 37.6% of the participants chose the fair model. Participants became more likely to choose the fair investment model (54.4%) when the biased model was explicitly labeled to be unfair. Interestingly, they seem to be significantly impacted by the annotation warning that, when the annotation was paired with the actual fair model, participants were less likely to trust the fair model (25.4%), see Figure 5.

We also replicated findings from Experiment 1 with regard to scenario and gender. Participants were slightly more likely to choose the fair investment model in the scenario when they were choosing on behalf of a gender-unknown client (39.1%) compared to on behalf of themselves (39.0%), although the effect size is small ($\eta^2_{part} = 1.14 \times 10^{-4}$). They were more likely to choose the fair investment model when the difference in return was larger between men and women, despite the model favoring their own gender. $F(3,4752) = 7896.4289, p < 0.001$. The effect of gender also persisted, such that women were more likely to choose the fair investment model. Participants who identified as Non-binary choose the fair robot in the conditions when the difference in return was larger (i.e., 85%/35%, 35%/85%), 49.94% chose fair and 37.04% chose fair in the conditions when the difference was small.

### 6.4 Discussion and Summary

Participants tended to be more fair when the explicit warning was on the biased model and less so when the explicit warning was on the fair model. This shows that people tend to make decisions based on the textual description and not what the model shows.

## 7 RQ7: Reasoning Strategies

To understand how model trust decisions are made, we asked our 1,519 participants from Experiments 1 and 3 to explain their reasoning after making their choice to select a model. We used an inductive thematic analysis [13, 14] for coding the collected data. Two authors independently constructed a set of codes from a subset of the data after going through each response. They then conducted a converging exercise to integrate their independent codes into a standard set of codes, with definitions and prototypical examples. They then used these codes to categorize the participant responses independently. The first author helped resolve disagreements.

### 7.1 Strategies

We identified seven strategies the participants used:

**Average**: The participants computed the average return for each model and compared the two model's average returns.

**Delta**: The participants computed the model's bias (difference in return between the two genders) and compared the two model's biases.

**Indifferent**: The participants were indifferent about their choice.

**Misaligned**: The participants' response did not align with the choice that they made, such as selecting the more biased model, having explained as selecting it for being less biased.

**Personal Beliefs**: Participants used information that was not provided, such as by making up assumptions for why a model favors one gender.

**Reliability/Consistency**: The participants reported that they selected the "safe," "reliable," or "consistent across trials" model.

**Self Profit**: The participant chose the model that historically had higher returns for the participant's gender.

### 7.2 Results

Figure 6 shows that women tended to use the delta strategy more and were more likely to choose the fairer model. It is plausible that the delta strategy, which involves computing model bias, relies on people noticing the discrepancies between returns in the biased model, which drives drive them to choose the fair model. In Experiment 1, participants more readily chose the fair model when shown textual model descriptions than bar charts. This suggests that textual descriptions with numerical values afford the computation of bias, while bar charts do not, and this affordance might be what drives participants to select the fair model when interacting with text.

This inference aligns with existing work on how people reason with data: explicitly showing numbers facilitates difference computations, while visualizing values using bar charts, instead, draws people's attention to salient large values [73]. This bottom-up attraction to the salient large bars, along with the top-down effects of paying attention to self-relevant data, potentially explains why participants more often used the self-profit strategy with bar charts. Depending on whether that self-relevant bar happened to be relatively large or small, participants ended up choosing the biased or fair model. We see supporting evidence of this, as the self-profit strategy is less predictive of selecting the fair model than the delta strategy. This effect of letting the salient large bar and self-interest drive attention and decision became attenuated when the participants made decisions on behalf of a client, as the participant was left without having a specific gender bar to focus on.

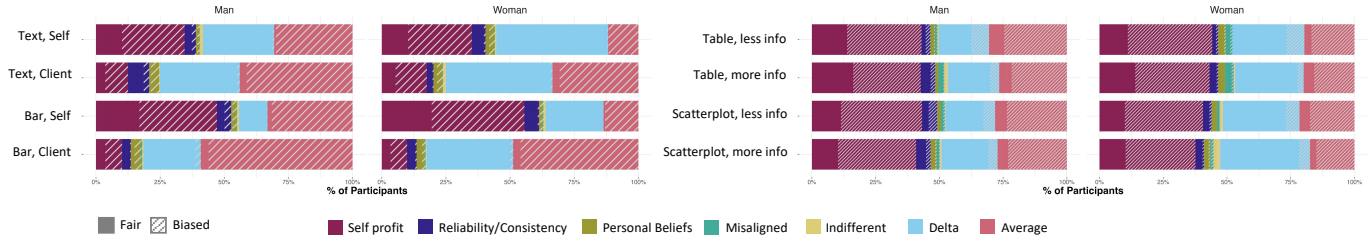


Fig. 6. The fraction of participants using each strategy who chose the fair or biased model in Experiment 1 (left) and Experiment 3 (right).

The strategy distribution for the scatterplot and table conditions fell between that of the bar chart and text: participants were less likely to use the self-profit strategy in these conditions compared to the bar chart condition and less likely to use the delta strategy compared to the text condition. One plausible explanation could be that, since both scatterplots and tables display multiple data points for the participant to consider, computing the average difference between returns becomes less intuitive compared to the text condition, and participants became less likely to use the delta strategy. Furthermore, unlike for the bar chart conditions, because there no longer exists one salient large value to focus on in the scatterplot and table conditions, participants became less likely to use the self-profit strategy as well.

Overall, fairness perception appears to be driven by the perceptual salience of self-relevant data values (which corroborates existing findings that suggest data is personal [52]), and how much the representation affords difference computation. Ultimately, fairness perception in visualizations seems closely related to the ease of perceiving differences between groups, potentially indicating that designing visualizations to highlight between-group differences and minimizing the salience of one large value related to self-interest might sway people from choosing the biased models.

## 8 Design Implications and Future Work

Our results consistently demonstrated that visualization design has profound consequences on how people perceive fairness and which models they trust. This section makes recommendations for designing model fairness visualizations and systems.

**Design for specific stakeholder-model relationships.** Our work provides preliminary evidence that investing on one's behalf vs. on behalf of a client affects their trust decisions (RQ2). But with some notable exceptions designed to support MLOps engineers and ML practitioners tackling validation issues [48, 81], prior work on visualization for fairness assessments [68, 71, 72] has generally not considered this relationship. We envision this relationship playing a greater role in the design of such systems, explicitly considering which stakeholders will use the tool, and validating the tool with the corresponding group.

**Embrace the diversity of user perspectives.** How individuals trust models is affected significantly by their demographics (RQ1, RQ5). For example, women are more likely to trust fair models than men are. Moreover, people follow a broad set of strategies in making trust decisions (RQ7). Visualization design must move beyond the monolithic "user" and embrace individual differences [33, 39].

**Use explicit bias warnings with caution.** Explicitly telling people a model may be biased can overpower the effect of a model's biased history (RQ6). While these warnings can enhance communication with the user, erroneous or malicious labeling can have a strong detrimental effect. The effects of explicit warnings can be more substantial than presentation modality (compare Figures 5 and 4). Visualization designers should only use explicit warnings following extensive consultation with stakeholders regarding when and how to deliver the warnings.

**Account for designer and user biases.** The majority of men and women trust and select models biased in their favor (Figure 3). This result aligns with the neoclassical economists' view that people seek to maximize their profits without considering effect on others [32]. This has two implications for relevant visualization design. First, designers must consider and account for how the users' perception of personal advantage will affect their decision-making. Second, the designers' motivations, potential personal gains, and subconscious biases may affect the design and should be explicitly considered.

**Consider moving beyond aggregate and average metrics.** Our results on scatterplots and tables suggest that the amount and granularity of data affects behavior: seeing more historical data increased trust in fair models (Figure 4). But consuming large, raw data can be overwhelming. Designers should consider showing information beyond aggregate metrics, but perhaps limit the use of raw data to small but consequential bias properties, as our results suggest aggregate metrics were sufficient for scenarios with large bias.

**Account for diverse users.** Gender is not binary, but our study specifically studied the effects of bias against men and women. Some participants wondered how the model would perform for someone who does not identify as a man or a woman. Excluding explicit mentions of potential bias against non-binary individuals led more participants who identify as non-binary to employ the indifferent strategy. This was an unintended flaw in our study design, and our ongoing work is tackling a broader exploration of gender-based biases. Designers (and researchers) should consider the implications of presenting biased data for specific groups when the visualizations will be consumed by members of other groups (which has also been done with race [34]). As a community, we must carefully monitor and assess our work for such inadvertent, implicit assertions and address their impact.

## 9 Limitations

All our participants were U.S.-based and all participants received the same compensation, without incurring consequences for their choices. Future work should study other cultures and trust games with financial incentives to increase the external validity of our findings.

Our study focused on bias against men and women, but gender is not binary. This choice may have adversely affected engagement from certain participants and future work is needed to understand both the effects on non-binary participants and how to include a more inclusive definition of gender when modeling how bias affects behavior.

Uncertainty can play an important role in visualization and decision making, and our experiments using scatterplots and tables explored its role. These experiments form a baseline for future work to more explicitly model and study uncertainty and its effect on trust.

## 10 Conclusion

Our study is the first exploration of how visual design choices, model performance and fairness, and user characteristics affect trust in ML models. We find that visual design plays a significant role in which models people trust and that women prioritize fairness more often than men do. People tolerate more bias when making decisions on their own behalf than on the behalf of others, and explicit warnings of bias have a bigger effect on trust than a biased history. Finally, we identify seven strategies people use when reasoning about trust, which suggests that affordances of difference computation and self-relevancy-driven salience can impact fairness perception. Overall, our findings support the importance of studying how visual design affects trust and the perception of fairness and identifies new research directions in this important field.

## REFERENCES


[1] Supplementary materials for My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning. https://osf.io/er5a3/?view_only=ce6801454c35476780d8591056f7c450, 2023.

[2] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, vol. PMLR 80, pp. 60–69. Stockholm, Sweden, July 2018.

[3] J. Ahmad, E. Huynh, and F. Chevalier. When red means good, bad, or canada: Exploring people's reasoning for choosing color palettes. In *2021 IEEE Visualization Conference (VIS)*, pp. 56–60, 2021. doi: 10.1109/VIS49827.2021.9623314

[4] Y. Ahn and Y.-R. Lin. Fairsight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1086–1095, 2020. doi: 10.1109/TVCG.2019.2934262

[5] V. Anderhub, D. Engelmann, and W. Güth. An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization*, 48(2):197–216, 2002. doi: 10.1016/S0167-2681(01)00216-5

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] K. Arun, G. Ishan, and K. Sanmeet. Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering*, 18(3):18–21, 2016.

[8] A. Bell, L. Bynum, N. Drushchak, T. Herasymova, L. Rosenblatt, and J. Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice, 2023.

[9] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, 1810.01943, 2018.

[10] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142, 1995. doi: 10.1006/game.1995.1027

[11] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[12] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics*, 22(1):519–528, 2015.

[13] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa

[14] V. Braun and V. Clarke. *Thematic analysis.*, pp. 57–71. 01 2012.

[15] M. Brülhart and J.-C. Usunier. Does the trust game measure trust? *Economics Letters*, 115(1):20–23, 2012. doi: 10.1016/j.econlet.2011.11.039

[16] J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, pp. 77–91. PMLR, New York, NY, USA, Feb. 2018.

[17] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 46–56, 2019. doi: 10.1109/VAST47406.2019.8986948

[18] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020. doi: 10.1177/1473871620904671

[19] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, vol. 39, pp. 713–756. Wiley Online Library, 2020.

[20] J. Coleman and A. C. of Learned Societies. *Foundations of Social Theory*. ACLS Humanities E-Book. Belknap Press of Harvard University Press, 1990.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226. Cambridge, MA, USA, August 2012.

[22] H. Elhamdadi, A. Gaba, Y.-S. Kim, and C. Xiong. How do we measure trust in visual data communication? In *2022 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, 2022. doi: 10.1109/BELIV57783.2022.00014

[23] H. Elhamdadi, L. Padilla, and C. Xiong. Processing fluency improves trust in scatterplot visualizations. *IEEE VIS 2022 Posters*, submitted.

[24] Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf, May 2016.

[25] S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, December 2005. doi: 10.1257/089533005775196732

[26] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[27] A. Gaba, V. Setlur, A. Srinivasan, J. Hoffswell, and C. Xiong. Comparison conundrum and the chamber of visualizations: An exploration of how language influences visual design. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1211–1221, 2022.

[28] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: Testing software for discrimination. In *Proceedings of the 11th Joint Meeting of the European Software Engineering Conference and ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pp. 498–510. Paderborn, Germany, September 2017. doi: 10.1145/3106237.3106277

[29] B. Ghai and K. Mueller. D-bias: A causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):473–482, 2023. doi: 10.1109/TVCG.2022.3209484

[30] S. Giguere, B. Metevier, Y. Brun, B. C. da Silva, P. S. Thomas, and S. Niekum. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, April 2022.

[31] E. L. Glaeser, D. I. Laibson, J. A. Scheinkman, and C. L. Soutter. Measuring trust. *The quarterly journal of economics*, 115(3):811–846, 2000.

[32] R. Goodland and G. Ledec. Neoclassical economics and principles of sustainable development. *Ecological Modelling*, 38(1):19–46, 1987. Ecological Economics. doi: 10.1016/0304-3800(87)90043-3

[33] K. W. Hall, A. Kouroupis, A. Bezerianos, D. A. Szafir, and C. Collins. Professional differences: A comparative study of visualization task performance and spatial ability across disciplines. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):654–664, 2022. doi: 10.1109/TVCG.2021.3114805

[34] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 392–402. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3351095.3372831

[35] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693, 2019. doi: 10.1109/TVCG.2018.2843369

[36] E. Holder and C. Xiong. Dispersion vs disparity: Hiding variability can encourage stereotyping when visualizing social outcomes. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):624–634, 2022.

[37] IBM. AI Fairness 360 Open Source Toolkit. https://aif360.mybluemix.net, 2019.

[38] B. Johnson, J. Bartola, R. Angell, K. Keith, S. Witty, S. J. Giguere, and Y. Brun. Fairkit, fairkit, on the wall, who's the fairest of them all? Supporting data scientists in training fair models. *EURO Journal on Decision Processes*, 2023. arXiv: abs/2204.10370.

[39] F. A. Khan and J. Stoyanovich. The unbearable weight of massive privilege: Revisiting bias-variance trade-offs in the context of fair prediction. *arXiv preprint arXiv:2302.08704*, 2023.

[40] S. S. Komorita. *Social dilemmas*. Routledge, 2019.

[41] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

[42] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12.

Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174012

[43] D. Kreps. Corporate culture and economic theory. in alt, j., shepsle, k.(eds.). perspectives on positive political economy, 1990.

[44] B. C. Kwon, U. Kartoun, S. Khurshid, M. Yurochkin, S. Maity, D. G. Brockman, A. V. Khera, P. T. Ellinor, S. A. Lubitz, and K. Ng. Rmexplorer: A visual analytics approach to explore the performance and the fairness of disease risk models on population subgroups. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 50–54, 2022. doi: 10.1109/VIS54862.2022.00019

[45] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, vol. 80, pp. PMLR 3150–3158, 2018.

[46] E. Mayr, N. Hynek, S. Salisu, and F. Windhager. Trust in information visualization. In *TrustVis at EuroVis*, pp. 25–29, 2019.

[47] B. Metevier, S. Giguere, S. Brockman, A. Kobren, Y. Brun, E. Brunskill, and P. S. Thomas. Offline contextual bandits with high probability fairness guarantees. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), Advances in Neural Information Processing Systems 32*, pp. 14893–14904. Vancouver, BC, Canada, December 2019.

[48] D. Munechika, Z. J. Wang, J. Reidy, J. Rubin, K. Gade, K. Kenthapadi, and D. H. Chau. Visual auditor: Interactive visualization for detection and summarization of model biases. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 45–49, 2022. doi: 10.1109/VIS54862.2022.00018

[49] S. Nadella. The partnership of the future. *Slate*, June 28, 2016. `http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html`.

[50] S. Otto and S. Weinzierl. Comparative simulations of adaptive psychometric procedures. *Jahrestagung der Deutschen Gesellschaft für Akustik*, pp. 1276–1279, 2009.

[51] S. Palan and C. Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[52] E. M. Peck, S. E. Ayuso, and O. El-Etr. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[53] M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–25, 2017.

[54] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445315

[55] Q.ai. How intelligent machines are reshaping investing. *Forbes*, 2022.

[56] I. Qualtrics. Qualtrics. *Provo, UT, USA*, 2013.

[57] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, vol. 29, pp. 1203–1210. Wiley Online Library, 2010.

[58] C. Ross. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *STAT+*, July 2018.

[59] J. B. Rotter. Interpersonal trust, trustworthiness, and gullibility. *American psychologist*, 35(1):1, 1980.

[60] P. K. Roy, S. S. Chowdhary, and R. Bhatia. A machine learning approach for automation of resume recommendation system. *Procedia Computer Science*, 167:2318–2327, 2020.

[61] D. O. Sears and P. J. Henry. Over thirty years later: A contemporary look at symbolic racism. *Advances in experimental social psychology*, 37(1):95–125, 2005.

[62] N. Sheard and A. Schwartz. The movement to ban government use of face recognition. `https://www.eff.org/deeplinks/2022/05/movement-ban-government-use-face-recognition`, May 2022.

[63] N. Singer. Amazon faces investor pressure over facial recognition. *New York Times*, May 2019.

[64] T. H. Stark, F. M. van Maaren, J. A. Krosnick, and G. Sood. The impact of social desirability pressures on whites' endorsement of racial stereotypes: A comparison between oral and acasi reports in a national survey. *Sociological Methods & Research*, p. 0049124119875959, 2019.

[65] C. Stokes, V. Setlur, B. Cogley, A. Satyanarayan, and M. Hearst. Striking a balance: Reader takeaways and preferences when integrating text and charts. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[66] P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019. doi: 10.1126/science.aag3311

[67] N. van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445365

[68] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu. Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1470–1480, 2021. doi: 10.1109/TVCG.2020.3030471

[69] R. Wang, F. M. Harper, and H. Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376813

[70] J. Wexler. The what-if tool: Code-free probing of machine learning models. `https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html`, 2018.

[71] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. doi: 10.1109/TVCG.2019.2934619

[72] T. Xie, Y. Ma, J. Kang, H. Tong, and R. Maciejewski. Fairrankvis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):368–377, 2022. doi: 10.1109/TVCG.2021.3114850

[73] C. Xiong, E. Lee-Robbins, I. Zhang, A. Gaba, and S. Franconeri. Reasoning affordances with tables and bar charts. *IEEE transactions on visualization and computer graphics*, 2022.

[74] C. Xiong, L. Padilla, K. Grayson, and S. Franconeri. Examining the components of trust in map-based visualizations. 2019.

[75] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *IEEE transactions on visualization and computer graphics*, 26(1):853–862, 2019.

[76] A. Yala, C. Lehman, T. Schuster, and T. P. andRegina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 2019. doi: 10.1148/radiol.2019182716

[77] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, 2019.

[78] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):3–36, 2021. doi: 10.1007/s41095-020-0191-7

[79] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Fairness, Accountability, and Transparency in Machine Learning (FAT ML)*. Lille, France, July 2015.

[80] R. Zemel, Y. L. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML), published in JMLR W&CP: 28(3):325–333)*. Atlanta, GA, USA, June 2013.

[81] X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren. Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852, 2023. doi: 10.1109/TVCG.2022.3209465

[82] J. Zheng, E. Veinott, N. Bos, J. S. Olson, and G. M. Olson. Trust without touch: Jumpstarting long-distance trust with initial social activities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, p. 141–146. Association for Computing Machinery, New York, NY, USA, 2002. doi: 10.1145/503376.503402

[83] I. Žliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *International Conference on Data Mining (ICDM)*, pp. 992–1001. Vancouver, BC, Canada, December 2011.

[84] M. Zürn and S. Topolinski. When trust comes easy: Articulatory fluency increases transfers in the trust game. *Journal of Economic Psychology*, 61:74–86, 2017.