The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features

NAVITA GOYAL*, University of Maryland, USA
CONNOR BAUMLER*, University of Maryland, USA
TIN NGUYEN, University of Maryland, USA
HAL DAUMÉ III, University of Maryland & Microsoft Research, USA

AI systems have been known to amplify biases in real world data. Explanations may help human-AI teams address these biases for fairer decision-making. Typically, explanations focus on salient input features. If a model is biased against some protected group, explanations may include features that demonstrate this bias, but when biases are realized through proxy features, the relationship between this proxy feature and the protected one may be less clear to a human. In this work, we study the effect of the presence of protected and proxy features on participants' perception of model fairness and their ability to improve demographic parity over an AI alone. Further, we examine how different treatments—explanations, model bias disclosure and proxy correlation disclosure—affect fairness perception and parity. We find that explanations help people detect direct biases but not indirect biases. Additionally, regardless of bias type, explanations tend to increase agreement with model biases. Disclosures can help mitigate this effect for indirect biases, improving both unfairness recognition and the decision-making fairness. We hope that our findings can help guide further research into advancing explanations in support of fair human-AI decision-making.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: indirect biases, fairness, human-AI decision-making, explanations

1 INTRODUCTION

Improving the fairness and trustworthiness of AI systems is often cited as a goal of explainable AI (XAI) [e.g., 4, 17, 19, 20, 39, 42, 62]. Research in XAI aims to improve fairness in human-AI decision-making by providing insights into model predictions, and thereby allowing humans to understand and correct for model biases. On the other hand, in the context of human-AI decision-making, previous work has noted that humans often over-rely on AI predictions, and explanations can exacerbate this concern [9]. This is especially troubling if the underlying model contains systematic biases, which may go unnoticed even when teamed with a human. In order for the human-AI team to be successful, the human needs to be able to determine when to rely on or override potentially biased AI predictions. Previous work has shown that explanations can help human-AI teams alleviate model biases when those biases depend directly on protected attributes [18, 54], but little is known in the very common case that protected attributes are not explicitly included, and rather the features used for prediction contain proxies thereof (e.g., zip code for race, length of credit for age, and university attended for gender). In particular, it may be difficult for humans to identify and resolve biased model predictions based on the proxy features present in real-world data, even when explanations are provided.

In this work, we study whether explanations can help people to identify model biases and to calibrate their reliance on an AI model based on these biases. We extend this line of investigation beyond direct biases that are revealed through the use of protected (i.e., sensitive) features by considering the effect of explanations when indirect bias is revealed

*Both co-first authors contributed equally to this manuscript, and each has the right to list their name first on their CV.

Authors' addresses: Navita Goyal, navita@umd.edu, University of Maryland, College Park, MD, USA; Connor Baumler, baumler@umd.edu, University of Maryland, USA; Tin Nguyen, University of Maryland, USA, tintn@umd.edu; Hal Daumé III, University of Maryland & Microsoft Research, USA, me@hal3.name.

through proxy features which may be less obvious to a human. Further, we examine whether explicitly disclosing model biases and correlations between the proxy and protected features can help humans calibrate their trust in a biased model. Our study aims to evaluate whether explanations can directly help notice model biases, even when the biases are obfuscated by the presence of proxy features and whether explanations can help users correct model biases when they are known to be present, through the use of bias disclosure and correlation disclosure. We study the effect of these treatments (explanations, model bias disclosure and proxy correlation disclosure) on the fairness, including fairness perception and fairness in decision-making (measured by group-wise parity), as well as the accuracy of the decisions made by human-AI teams.

We conduct our study in the context of micro-lending outcome prediction—a setting that entails judging whether a loan applicant will fulfill their loan request based on profile information of the applicant (e.g., size of the loan, borrower occupation, etc). For our experiments, we use semi-synthetic data where the majority of the features in an applicant profile as well as the final loan repayment status comes from the website Prosper.[1] To incorporate fairness considerations, we add to the applicant profiles (binary) gender, which is a protected feature, and university which, when considering women's vs co-ed colleges, can be a proxy for gender. Because we seek to test whether people can correct for model bias, we intentionally train a biased predictor with outcomes skewed against applicants with gender assigned as female or university assigned as a women's college.

We find that explanations alone can help people notice unfairness in the case of direct bias (through protected features, e.g., gender), but not in the case of indirect bias (through proxy features, e.g., university). Surprisingly, regardless of whether people notice the unfairness in the AI decisions, explanations lead people to accept model's biased decisions leading to less fair decisions. In the case of direct bias, as participants often recognize clear-cut gender bias before an explicit disclosure, disclosing model biases does not further affect participants' fairness perception. However, in the case of indirect bias, disclosing both model bias and the correlation between protected and proxy features or disclosing partial information with the addition of explanations significantly improves participants' awareness of the unfairness. However, contrary to explanations alone, this change is *not* paired with a worsening of decision-making fairness. Instead, with these disclosures, people increase their rate of positive predictions for the disadvantaged group, improving decision-making fairness. Our work aims to highlight methods to assist users in effectively leveraging explanations, especially in scenarios where bias may be indirect and not apparent through explanations alone.

2 BACKGROUND AND RELATED WORK

Biases in Models and Humans. Both models and humans can be biased. Humans have been known to exhibit many implicit and unconscious biases [31]. For instance, Bertrand and Mullainathan [6] find that an applicant with a "White-sounding name" on a resume that is otherwise identical to a resume with an "African-American-sounding name" is more likely to receive an interview callback.

Models, in turn, can inherit human-like biases (e.g., through biased data [3, i.a.]), even if this is not intended by the developers. For instance, Angwin et al. [1] show that training on data collected from a racist justice system can lead to a model that predicts that white defendants are less likely to recidivate than their black peers.

XAI and Decision-Making. The potentially complementary strengths and weaknesses of humans and machines raises a question of whether human-AI teams can overcome the biases that exist in each individually (e.g., in the case of recidivism prediction [e.g., 14, 21, 61]). Existing work in explainable AI (XAI) has focused on providing explanations of

[1] https://www.kaggle.com/datasets/yousuf28/prosper-loan

the model decisions to help improve the outcomes of human-AI decision-making [2, 8–12, 26, 29, 35, 37, 38, 43, 49, 55, 58, 60, 61, 63, 65]. However, these studies find varying utility of explanations. Much work has found that explanations can help humans collaborate more effectively with AI [11, 26, 29, 30, 37, 38, 58, 63], for instance helping them answer trivia questions more accurately [26] or understanding how the AI system works [13]. Other work has found that explanations can worsen human-AI performance [2, 8–10, 12, 35, 49, 55, 60, 61] even below the performance of the human or AI alone. Further, the utility of explanation can also vary based on the participant's level of expertise in the task [e.g., 61], the participant's math and logic skills [57], how easy the explanations are to understand [63], etc.

Beyond explanations, other work has considered how further transparency can or cannot be beneficial to a human-AI team such as tutorials [38], disclosing model confidence [51], disclosing model accuracy [23], and disclosing whether test examples fall into the scope of model training data [15].

XAI and Fairness. Improving model fairness is often cited as a potential benefit of XAI systems [4, 17, 19, 20, 39, 42, 62]. XAI is hoped to help "diagnose the reasons that lead to algorithmic discrimination" [20], to "highlight an incompleteness" in problem formalization that leads to unfairness [19], or to show compliance with fairness requirements [62].

Previous work has examined how explanations affect humans' perceptions of AI systems' fairness [7, 18, 40, 50, 54, 64]. Rader et al. [50] find that participants that are told that an AI system is being used in decision-making rate the system as significantly less fair even without any specific system information. Lee et al. [40] find that explanations of an AI system's general decision-making process do not increase perceived fairness while input-output level explanations of individual outcomes have mixed effects on fairness perceptions. Binns et al. [7] consider how four different styles of explanations affect justice perception, finding no clear winner between the approaches. Dodge et al. [18] further study the explanations styles in [7] and find that local explanations (such as presenting outcomes for similar examples) help surface fairness discrepancies between different cases while global explanations (such as describing how each feature influenced the decision for a given example) increase user confidence in their understanding of the model and enhance users' fairness perceptions.

As self-reported perceptions do not always align with observed behaviors in human-AI decision-making [8, 47], recent work has begun to expand out of fairness perceptions and into observed fairness in decision-making [54, 59]. Schoeffer et al. [54] study how explanations can help users appropriately rely on potentially unfair AI predictions.They find that explanations that highlight protected features negatively affect fairness perceptions and that decreases in fairness perception are associated with an increase in overrides of AI predictions, even on examples where this override is detrimental to the fairness of the human-AI team. Wang et al. [59] consider the effects of the level of model bias and the presence of explanation on the fairness of human decisions. They find that explanations lead participants to make more unfair decisions, even when participants were no longer given access to model predictions or explanations.

Existing work has primarily studied fairness when the model decision is directly based on a protected feature, like gender or race. However, models can produce biased outcomes, even without access to protected features, by relying on proxy features [34, 48]. For instance, a model that has direct access to a "race" feature and one with access to features like zip code, name, or language spoken at home could produce similarly biased predictions. In contrast to existing work considering the relationship between explanations and fairness perceptions or decision-making fairness, we consider not only direct bias through a protected feature but also indirect bias through a proxy feature.

3 RESEARCH QUESTIONS

We study the effect of explanations in improving the fairness of decisions made by human-AI teams when bias stems from different kinds of features and when participants are given different kinds of information about the model and its training data. In our study, model biases can be direct: stemming from the protected feature (gender), or indirect: stemming from a proxy feature (university) that is correlated with the protected feature. Participants may receive disclosures about the level of model bias and the strength of correlation between the proxy and protected feature. Our study addresses the following research questions:

RQ1a: Are explanations beneficial to the fairness of a human-AI team?

RQ2a: Without explanations, does disclosing only model bias or disclosing model bias and proxy correlation benefit human-AI fairness?

RQ3a: With explanations, does disclosing only model bias or disclosing model bias and proxy correlation benefit human-AI fairness?

RQ4a: Does the joint intervention of adding explanations and disclosures benefit human-AI fairness?

RQ1-4b: Do the answers to RQ1-4a change when models exhibit direct (e.g., gender) vs indirect (e.g., university) bias?

We consider the utility of explanations and disclosures under three lenses: the accuracy of **human's perception of fairness** and the improvement in **demographic parity in human-AI decision-making** over AI-only parity, and the **decision-making quality** (namely, accuracy, false negative rate (FNR), and false positive rate (FPR)) **of human-AI decisions** compared to the AI alone.

Beyond these primary research questions, we also consider:

RQ5: How does dispositional trust ("an individual's enduring tendency to trust automation" [32, 44]) affect decision-making and fairness perceptions when working with models exhibiting direct or indirect bias?

RQ6: How do explanations and disclosures affect self-reported learned trust (based on "past experience or the current interaction" [32, 44]) in models exhibiting direct or indirect bias?

4 STUDY DESIGN

To answer the research questions posed in §3, we study decisions made by human-AI teams.[2] In this study, the AI teammate is a classification model trained on partially-synthetic data in the context of loan prediction. We choose the task of loan prediction from a micro-lending platform as it is a decision-making task performed by laypeople which means that crowd-workers are more likely to have intuitions about the task and the features used in predictions. In our study, participants are shown either the protected feature of binary gender[3] or the proxy feature of university.

4.1 Conditions

In our study, we vary conditions based on the directness of bias, whether explanations are shown, and the kind of disclosure the participant receives.

- Protected or Proxy: Whether the participant is shown a model and features including the protected (gender) or proxy (university) feature.

[2]This study design is IRB approved. (#1941548-2),

[3]We only consider binary gender in this study. Since each participant sees only a handful of examples per task phase, it would be difficult to both show the participants a statistically realistic number of non-binary applicants and get a good sense of how participants handle anti-trans model bias. We leave this for future work.

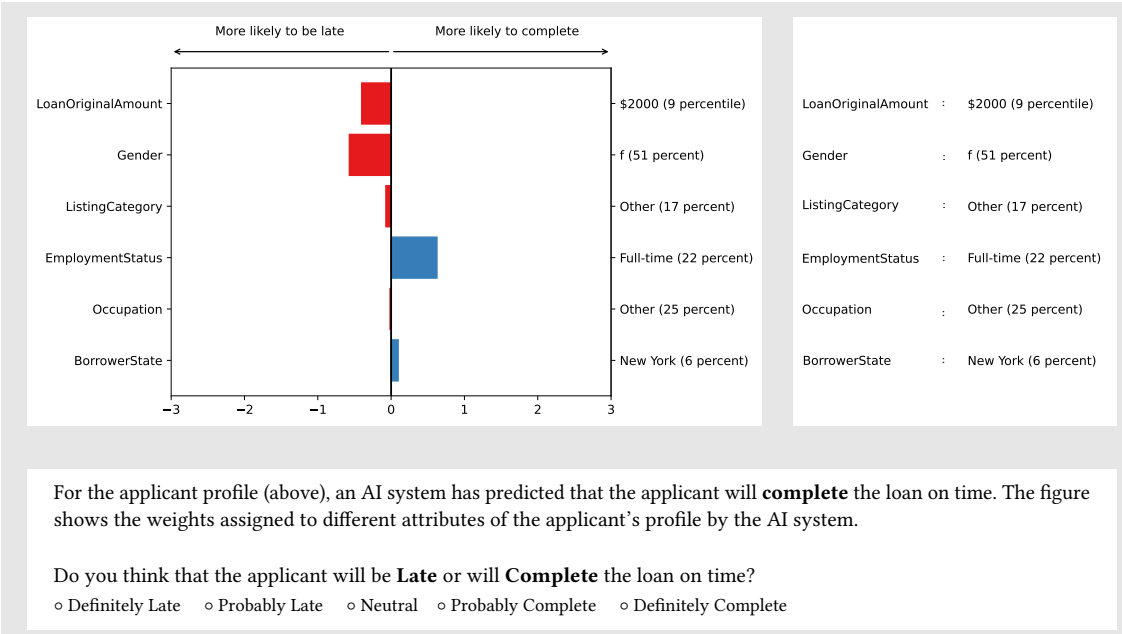

Fig. 1. Example profile with explanation from the "protected" model (left) without an explanation (right) and question to the user (below). The predicted outcome is completing the loan on time. The labels on the left show the name of each feature. The labels on the right show the value of each feature for the current applicant and the percent/percentile of this value in the training data. For the explanation, on the x-axis positive blue values correspond to "Complete" predictions and negative red to "Late". See Figure 8 in the Appendix for an example profile as shown in the study interface.

- Explanations: Whether the participant is shown an input-influence explanation of how features contributed to the AI's prediction (Figure 1, top).
- Disclosure: In the case of direct bias, we show the participant a *bias* disclosure, which reveals the demographic parity (described in §5.1) of the system. In the case of indirect bias, the participant may be shown only the bias disclosure or a full *bias and correlation* disclosure which additionally explains the relationship between university and gender features (Figure 3).

In this study, we consider six conditions. In the first three conditions, we do not show participants explanations. Here, we consider one **Protected** condition with bias disclosure, and two Proxy conditions: one **Proxy with correlation disclosure** and one **Proxy without correlation disclosure**. We similarly consider three conditions with explanations allocating the biased feature and disclosure-types in the same fashion.

We assess the effect of explanations (RQ1a) comparing conditions with and without explanations (before any disclosures) in a between-subjects analysis. Similarly, we assess the effect of disclosures without explanations (RQ2a) and the effect of disclosures with explanations (RQ3a) in a within-subject analysis comparing fairness perceptions and human-AI decision-making pre- and post-disclosures. This allows us to study how disclosures may help participants in identifying model biases over what is apparent before any disclosures. These first three effects are summarized in Figure 4. Lastly, we assess the effect of explanations and disclosures jointly (RQ4a) by comparing conditions in which participants are *not* shown explanations pre-disclosures with conditions in which participants are shown explanations

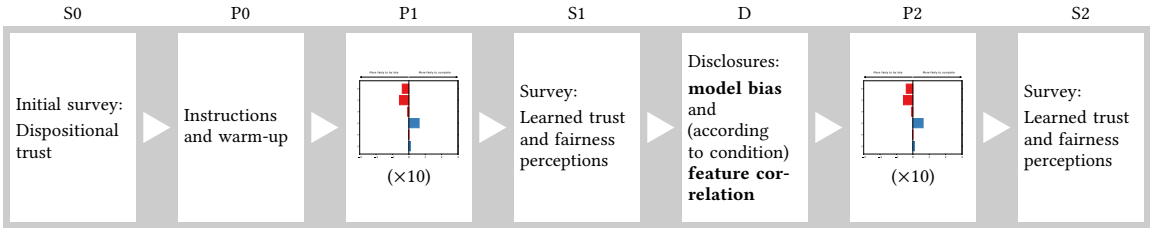

Fig. 2. Order of study phases.

post-disclosures. These experiments are repeated both for protected and proxy conditions to assess the differences in interventions therein (RQ1-4b).

4.2 Procedure

Our study procedure consists of three surveys (S0, S1, S2), one tutorial and warm-up phase (P0), two task phases (P1, P2), and a disclosure interlude (D) ordered as shown in Figure 2.

Task Phases. In each task phase (P1, P2), the participant is shown 10 profiles of loan applicants: their features and the overall AI prediction. Depending on the condition, they may or may not be shown an explanation of the AI prediction (Figure 1 left and right, respectively). This profile will, according to the condition, include either a "gender" or a "university" feature but not both. Participants are asked to mark on a five-point-scale whether they think the applicant will complete their loan on time or be late in repaying their loan (Figure 1, below). Their response to this question serves as the decision made by the human-AI team.

In each phase, we control the distribution of gender and AI predictions. The participant sees applications from 2 women who are predicted as "Complete" and 3 women who are predicted as "Late" and vice versa for men. (This is true in the underlying data even if the participant and the model do not directly see each applicant's gender.) We hold this ratio constant to avoid any effect due to the gender distribution or the rejection rate observed by different participants.

To discourage participants from making decisions without any consideration of the prediction and (when applicable) explanation, we ask participants for a free-text justification of why they agreed or disagreed with the model prediction (or was neutral) after they have chosen their prediction on selected profiles. We randomly select one application in each gender + prediction combination for collecting these free-text justifications. These justifications also help us qualitatively assess the reasoning behind participants' decisions. Further, to help filter out low-quality responses, participants are shown an attention check question, asking them to recall the previous AI prediction (Figure 9 in the Appendix) after seeing the first applicant in P1.

Disclosures. Before proceeding to P2, participants may be shown general explanatory materials or specific disclosures on model bias and feature correlations. In the model bias disclosure (Figure 3a), participants are told that the model they saw in P1 had a low demographic parity (below 80%) (see §5.1 for details about demographic parity). In the correlation disclosure Figure 3b, participants are told the correlation between each university and gender in the model's training data. The bias disclosure is shown across conditions, whereas correlation disclosure is only shown in the proxy conditions *with correlation disclosure*. In the proxy conditions *without correlation disclosure*, participants are only told

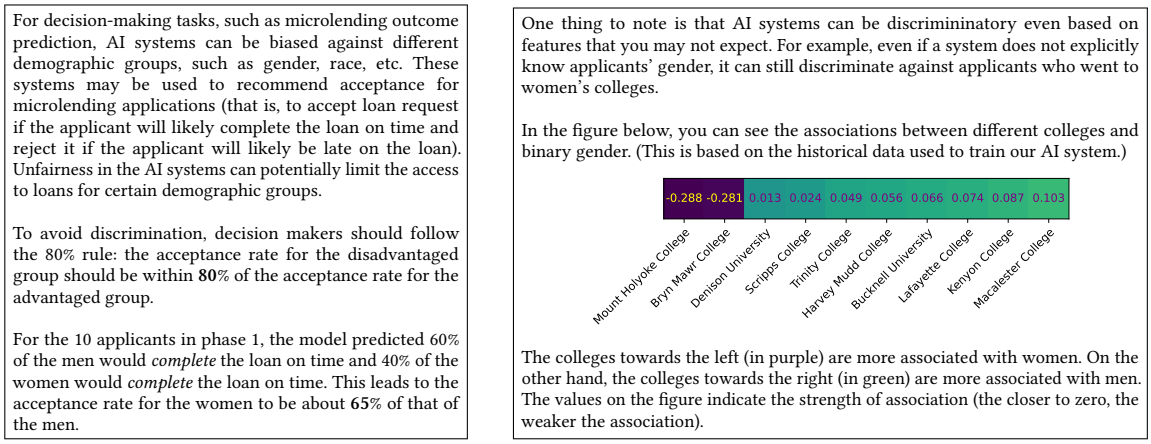

Fig. 3. a) Bias disclosure. b) Full correlation disclosure. Proxy "no correlation disclosure" conditions include the top paragraph but with the example of a hiring system relying on the relationship between zip code and race. See Figure 10 and Figure 11 in the Appendix for how these disclosures are shown in the study interface.

that models can rely on proxy features to make biased predictions, without specifying the correlation between gender and university. This is done to make participants aware of potential biases without explicitly disclosing the correlations.

Based on the disclosures seen, participants are asked up to two comprehension questions (Figure 13 in the Appendix). All are asked whether the model's demographic parity was above 80%. Those who received correlation disclosure are asked to select one university that is highly associated with women.

Note that before or during the first phase of the task, participants are never encouraged or primed to think about fairness explicitly. We only refer to fairness directly after phase 1. This allows us to measure how well participants can notice, or account for, unfairness when they aren't explicitly told to look out for it in phase 1. Following this, in phase 2, we can then measure how participants perceive and account for unfairness when they know it is a salient concern.

Surveys. The three surveys (S0, S1, S2) aim to capture participants' trust and fairness perceptions. All surveys include questions asking participants to rate their level of agreement with statements relating to trust (on a scale of 1-5) [33]. In S0, participants are asked about their trust in AI systems generally, assessing their dispositional trust (Figure 14 in the Appendix). In S1 and S2, participants are asked about their trust in the system presented in the task phases, assessing their learned trust in the AI system that they interact with in the study (Figure 15 in the Appendix).

In the post-task surveys (S1 and S2), alongside trust-related questions, participants are also asked about their perception of the fairness of the system they have been interacting with (whether "*the AI system was fair across different genders*"). Additionally, participants are asked the reason(s) that led to the their disagreements with AI such as the explanations including irrelevant features or the decisions being unfair towards applicants of different genders.

Tutorial and Warm-up. In P0, participants are acclimatized to the task with a full tutorial example. They are shown one tutorial example with a walk-through of the task, the AI decisions and explanations (when applicable). Then they are shown warm-up examples. In the conditions without explanation, they are shown two example with no AI prediction or explanation. In the conditions with explanations, they are first shown a version of this example with no

AI prediction or explanation. This is designed to encourage participants to properly engage with the features present. Second, they are shown the same example with the AI feature explanation (still without any prediction) as this setting has been shown to benefit decision quality and support learning by encouraging participants to cognitively engage with explanations [27].

4.3 Participants

We recruit 369 participants for our study through the crowdsourcing platform Prolific.[4] Each participant is restricted to taking the study only once. Participation is restricted to US participants, fluent in English. We compensate all participants at an average rate of US$15 per hour. We discard responses that fail more than one attention check, leaving a total of 350 participants, with 51, 48, 45 participants in the *protected condition*, the *proxy condition with correlation disclosure*, and the *proxy condition without correlation disclosure* without model explanation and 68, 69, 69 participants in the three respective conditions with model explanations. 42% of participants self-identified as women, 52% as men, 3% as non-binary/non-conforming, 3% as transgender, and 1% as a different gender identity, with 1% of participants opting not to respond.[5] 19% of participants were between the ages of 18-25, 46% between 25-40, 27% between 40-60, and 6% over the age of 60, with 2% of participants opting not to respond.

5 SYSTEM OVERVIEW

We conduct our study using model predictions and explanations from logistic regression models trained on partially synthetic micro-lending data. Since the participant's perceptions of how the model is interacting with the profile features is key to answering our research questions, we want to avoid any potential confounding effects from using artificial or Wizard-of-Oz model explanations, or entirely synthetic data.

The scenario of predicting whether an applicant will complete microloan repayment on time or will be late is one that our participants will likely be sufficiently familiar with to have reasonable prior intuitions about what features are relevant. A challenge is that under US law, protected features like gender cannot be considered when making loan allocation decisions [52] and therefore is not in the dataset that we consider. For this reason, we augment our data with a synthetic "gender" feature which we correlate with outcome to induce model bias. We also generate a proxy feature, university, which allows us to finely control the level of correlation between the proxy and gender.

Data. Our loan prediction data comes from a modified set of microloans from the website Prosper.[1] The original dataset contains 79 features of microloans including their status (completed, past due, etc). We group the loan statuses into "Complete" (including "Final Payment in Progress"), "Late" (including "Defaulted" and "Charged-Off"), or "Other" (including "Current" or "Canceld"). We keep the ~14000 profiles with "Complete" or "Late" statuses (with a 7:3 train-test split). This grouped loan status is the feature that the participants and the model will predict. As showing all 79 features to the participant may be overwhelming, [49] we select 5 features (the original amount of the loan, the category of the listing, the applicant's occupation and employment status, and their state of residence) that are both important to loan prediction and are likely interpretable by a layperson.

As described above, we generate values for our protected characteristic (binary gender) synthetically. The existing applicants are assigned a gender such that women "Complete" vs are "Late" in repayment with a 2:3 ratio and vice

[4]https://www.prolific.com/
[5]These do not add up to 100 as participants may have selected multiple options.

versa for men. This simulates historically biased data which will cause the model to associate femaleness with being late on loans and maleness with completing them.

Using the generated "gender" feature, we further generate the proxy feature (university). We include co-ed and women's colleges, setting the joint distribution of gender and university such that most co-ed universities have relatively balanced gender ratios (See Figure 3b). For women's colleges, the distributions reflect real-life statistics. We choose exclusively liberal arts colleges with similar US News rankings[6] to avoid confounding due to the effect of perceptions of liberal arts vs non-liberal arts schools and perceptions of school rankings.

Since, in our biased dataset, gender is correlated with outcome and, of course, the existing features are correlated with outcome, all features may be weakly correlated with gender. To confirm that university is the only strong proxy in our data, we compare the correlation of each categorical and continuous feature with gender. For continuous features (and one-hot features of each university), we use Pearson's r coefficient. We find that the women's colleges have at least an absolute correlation of 0.273 across Proxy conditions, whereas the maximum absolute correlation for other continuous features is 0.014, which is much lower. Similarly, for categorical features, we use Cramer's V, finding that the university feature has at least an absolute correlation of 0.417 while the maximum absolute correlation for the remaining categorical features is 0.082, which is also lower. Overall, we see that university (especially women's colleges) has a much stronger correlation with gender than any other feature shown to the participants.

Models. For our AI predictions, we use logistic regression models as explanations on simple models may be more useful to humans [37]. We train the models on 14 pre-selected features from the Prosper dataset (of which participants will only see 5) and, when applicable, the gender or university feature. These models have an average accuracy of about 65% when compared to the original ground-truth values before adding synthetic features. Since we are using logistic regression, we can create a simple input-influence explanation of the AIs' predictions using feature weights. For continuous features like `LoanOriginalAmount`, we multiply the normalized feature value by the corresponding feature weight. For categorical features like `EmploymentStatus`, we take only the feature weight corresponding to the feature value (e.g., the weight of the `EmploymentStatus` = `Full-Time` feature). These values are graphed as in Figure 1 (left).

5.1 Metrics

We evaluate study outcomes based on participants' perceptions and decisions. We consider two fairness perception metrics based on questions in the post-phase surveys, and we consider the decisions made by the human-AI teams based on one fairness measure—demographic parity—and three decision quality metrics: accuracy, false negative rate, and false positive rate.

We measure all metrics in the two task phases across conditions. We count both "Likely Complete" and "Definitely Complete" as "Complete" and similarly for "Late". We count "Neutral" as agreement with the system prediction.

5.1.1 Decision-Making Fairness Measure. We employ demographic parity [25, i.a.] as a measure of fairness in decision-making, which captures the independence between protected characteristics and prediction. There are other measures of fairness [46], however, not all definitions can be simultaneously satisfied [16, 36]. Demographic parity has been found to be more understandable to laypeople and better capture their perception of fairness than competing metrics [53, 56].

[6] https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges

We can calculate the demographic parity for human-AI teams in task phases 1 and 2 across conditions as follows.

$$\text{Parity} = \frac{\mathbb{E}[\hat{Y}_{i,j} = 1 \mid \text{Gender} = \texttt{female}]}{\mathbb{E}[\hat{Y}_{i,j} = 1 \mid \text{Gender} = \texttt{male}]},$$

where $\hat{Y}_{i,j}$ is the predicted decision for the applicant by the participant j. We obtain one demographic parity score in this way for each participant's decisions in each phase.

A parity close to 1 means an equal acceptance rate. As the acceptance rate for the advantaged group increases over the disadvantaged group, parity becomes closer to 0. If the acceptance rate of the disadvantaged group increases above the advantaged group, then the parity can increase above 1. A parity of less than $\frac{4}{5}$ is considered "evidence of adverse impact" under US Anti-Discrimination law [24]. In our model bias disclosure, we tell the participants about this 80% rule and that the model failed this test in phase 1, that is, the demographic parity of the model is below 80% (Figure 3a).

5.1.2 Fairness Perception Measures. Based on our post-task surveys (described in §4.2), we calculate two measures of how participants perceive the degree of model unfairness. First, we consider how much participants agree with the statement "The AI model was fair across different genders". Here, the participant's *fairness rating* is higher when they believe the model is more fair. We also consider whether participants mark "unfairness" as a reason that they disagreed with model decisions. Here, the participant's *fairness saliency* is higher when they have a greater belief that they disagreed with the model due to unfairness.

5.1.3 Decision Quality Measures. Measures such as accuracy require Y_i's: a ground-truth "Complete" or "Late" value to compute. We have access to the ground-truth loan completion status for the original applicants. However, as we discuss in §5, our study uses an edited set of applicants with synthetic gender and university features which are made to be correlated with the outcome. We estimate the loan completion status for the edited profile from the ground-truth completion status of the original applicants and our defined sampling rates of the synthetic features using Bayes' rule. In turn, we compute an expected accuracy, expected FPR, and expected FNR using the estimated loan completion status as our decision-quality measures. See Appendix A for more details.

5.2 Statistical Analyses

To answer our research questions (§3), we perform separate multi-way ANOVA tests for different treatments (explanations, disclosures without explanations, and disclosures with explanations) for both protected and proxy conditions. For each statistical test, we construct a linear model with a fixed effect term for each independent treatment variable and one fixed effect term representing the participant's dispositional trust, which is calculated by averaging the scores from the pre-study trust survey. Additionally, in the within study comparisons (that is, moving from phase 1 to phase 2), we include the participant ID as a random effect.

The independent treatment variables are determined by the factors that vary between the effect of interest. For instance, to estimate the effect of explanation (that is, the vertical arrow in Figure 4), the treatment variable is the *presence of explanations*. For the effect of disclosure (that is, the horizontal arrows in Figure 4), the treatment variables are: (1) whether only *bias disclosure* has been shown (i.e., is this a phase 2 measurement with no correlation disclosure), and (2) whether full *bias and correlation disclosure* has been shown. For the effect of adding both explanations and disclosures (that is, the diagonal in Figure 4 going from without explanation and disclosures in phase 1 to with explanation and disclosures in phase 2), we include all three treatment variables.

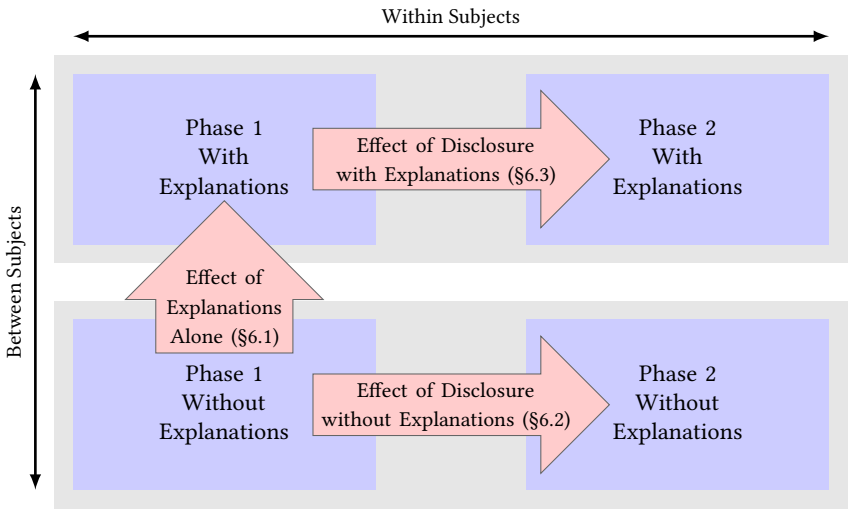

Fig. 4. Summary of primary effects considered in our study. Participants are assigned to either with or without explanations conditions and then complete the study moving horizontally from phase 1 to phase 2. We then compare the results of different combinations of phases and explanation conditions to investigate the effects of explanations alone, disclosures without explanations, and disclosures with explanations.

In each ANOVA test, we consider the data from the relevant sections. For instance, to estimate the effect of explanations alone in the case of direct bias through a protected feature, we only consider the data in phase 1 of the "protected" conditions (left vertical section in Figure 4), and similarly for "proxy" conditions.

We fit a separate model for each fairness perception and decision-making metric as the dependent variable for each of the above effects. Although the ratio of "Complete" and "Late" model decisions shown to each participant is kept the same in each phase across conditions (leading to a constant AI-only parity of ~0.67), the model accuracy, FPR, and FNR varies across conditions. To account for this variation, we subtract the model score from the score of the human-AI team. Similarly, when considering learned trust measures, we also adjust for dispositional trust in AI by subtracting the participant's answer to the pre-study survey for the corresponding question.

In addition to our key metrics detailed in §5.1, we also study the effect of the treatments on learned trust. We follow the same procedure as before, but with the learned trust measures as the dependent variable in the ANOVA test. In this case, however, we adjust participants' post-phase 1 or post-phase 2 survey responses based on their baseline responses, and we do not include the overall dispositional trust term. Lastly, we perform additional ANOVA tests considering the difference between dispositional and learned trust. For this, we consider the participants' trust ratings in the pre-study survey and surveys after phase 1 or phase 2 as the dependent variable. We fit two linear models, one for phase 1 and 2 each, testing whether the phase has a fixed effect on the participants' trust ratings in different conditions (with the participant added as a random effect).

We perform Benjamini-Hochberg correction to avoid multiple testing effect with a false discovery threshold of 0.05 [5]. This leads to a significance threshold of 0.0175 for the reported results.

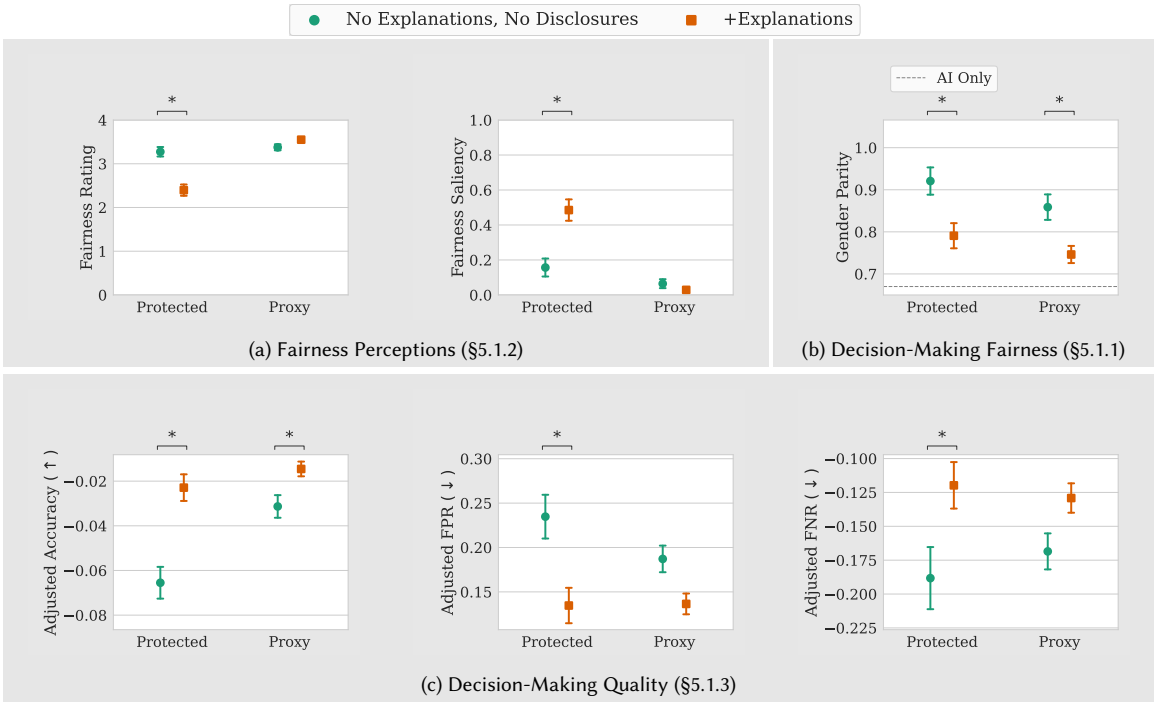

Fig. 5. Effect of explanations alone on various metrics when bias stems from usage of a protected vs proxy feature. The marks show the average and standard error of the given metric across participants in the given condition.

6 QUANTITATIVE RESULTS

In this section, we report our findings on the effects of different interventions on the decision-making and fairness perception metrics. We first discuss the primary effects detailed in Figure 4: the effect of explanations (§6.1), the effect of disclosures without explanations (§6.2) and the effect of disclosures with explanations (§6.3). Next, we consider the effect of the joint intervention of adding both explanations and disclosures to assess their aggregate benefits, if any (§6.4). Lastly, we discuss the effect of dispositional trust on the decision-making and fairness perception, the effect of different interventions on participants' trust, and the differences in participants' dispositional trust vs trust in our system in §6.5 (full results in Appendix B).

6.1 Effect of Explanations Alone

First, we consider the effects of explanations alone by comparing between the first phase of with and without explanations conditions with either type of bias.

In the case of direct bias through a protected feature, we find that explanations alone have a significant effect on all metrics; however, the direction of the effect is not consistent. Explanations alone significantly improve participants' ability to recognize unfairness (Figure 5a). Surprisingly, despite participants being more able to recognize that the model is unfair when shown explanations, when considering decisions instead of perceptions, we see that explanations significantly decrease gender parity (Figure 5b). Looking closer, we find that explanations lead to a significantly lower acceptance rate for female applicants, whereas the acceptance rate for male applicants does not change significantly

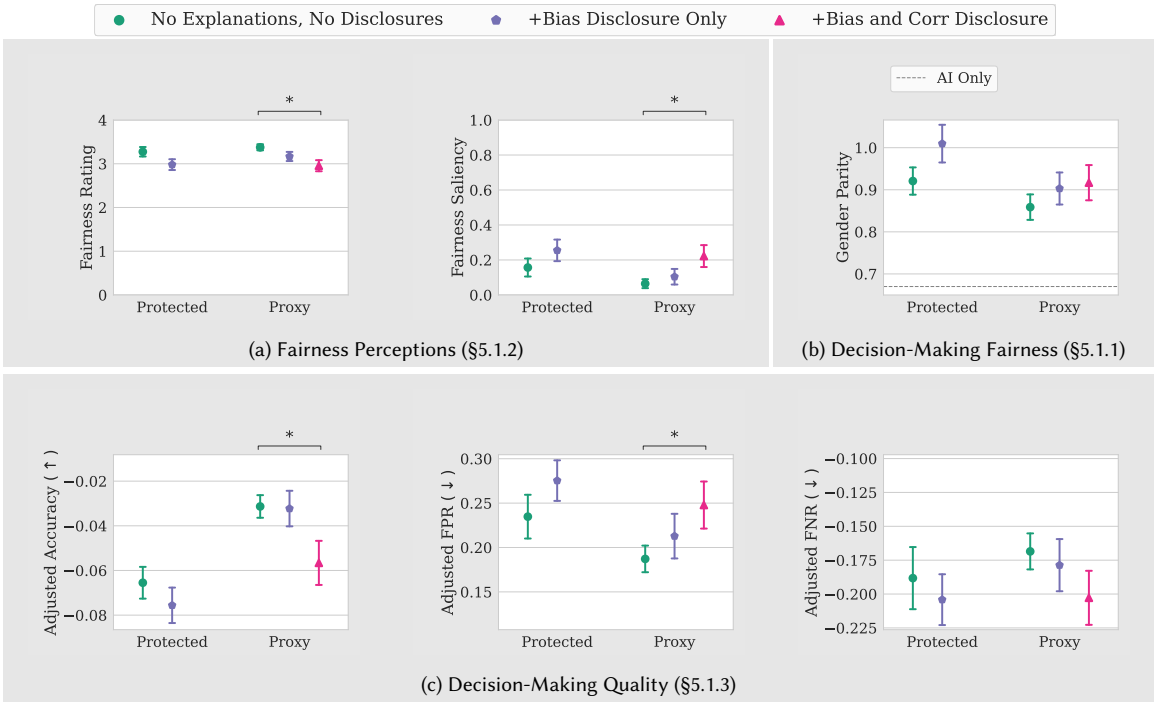

Fig. 6. Effect of disclosures *without* explanations on various metrics when bias stems from usage of a protected vs proxy feature. The marks show the average and standard error of the given metric across participants in the given condition.

(Table 5 in the Appendix). This decrease in acceptance rates also leads to a significant increase in the FNR and a significant decrease in FPR, with an overall higher accuracy (Figure 5c).

In the case of indirect bias through a proxy feature, we find that explanations alone significantly reduce gender parity, similar to the case of direct bias. (Figure 5b). Analogous to the direct bias, this occurs due to a significant decrease in acceptance of female applicants (Table 5 in the Appendix). Further, explanations also lead to a significant increase in accuracy in the case of indirect bias as well. However, unlike in the case of direct bias, explanations have no significant effect on fairness perceptions in the case of indirect bias (Figure 5a).

Overall, we find that explanations can help people recognize unfairness in the case of direct bias but not indirect. However, regardless of fairness perceptions, in line with Wang et al. [59], we find that explanations lead people to accept model biases leading to less fair decisions.

6.2 Effect of Disclosures without Explanations

We consider the effects of disclosures without explanations by comparing between phase 1 and phase 2 in without explanations conditions with either type of bias.

In the case of both direct and indirect bias, we find that disclosing model bias alone does not have a significant effect on any of the outcome metrics (gender parity, fairness perception, accuracy, FPR, and FNR).

However, in the case of indirect bias, when we disclose both the model bias and the relationship between the protected and proxy feature (i.e., that some universities in the study are women's colleges), participants were significantly more

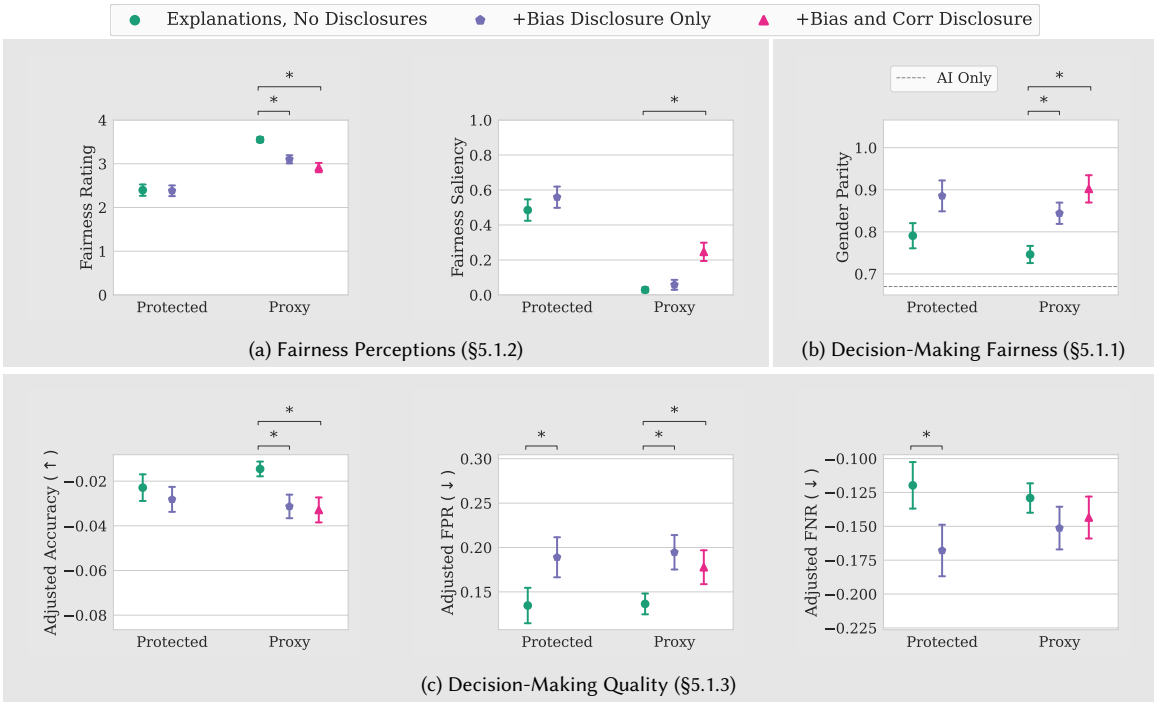

Fig. 7. Effect of disclosures *with* explanations on various metrics when bias stems from usage of a protected vs proxy feature. The marks show the average and standard error of the given metric across participants in the given condition.

likely to report that the model is unfair or that this unfairness caused them to disagree with the model's decisions (Figure 6a). Interestingly, this still does not translate to fairer decisions—as seen in Figure 6b, the gender parity does not change significantly on disclosing both model bias and correlations in the case of indirect bias.

In sum, we find that, interestingly, being explicitly told that the model is biased does not affect participants' fairness perception of the model decisions (in both direct and indirect bias conditions). But, disclosing both the model bias and the correlation between protected and proxy feature does lead to participants perceiving the model as less fair in the case of indirect bias. However, this is not sufficient to improve the decision-making fairness.

6.3 Effect of Disclosures with Explanations

We consider the effects of disclosures with explanations by comparing between phase 1 and phase 2 in explanations conditions with either type of bias.

In the case of direct bias through a protected feature, we find that bias disclosure with explanations has no significant effect on fairness perceptions (Figure 7a). Different from bias disclosure without explanations (§6.2), we find that bias disclosure with explanations significantly increases the acceptance rate for female applicants (participants flip models' "Late" prediction for female applicants at a much higher rate), with the acceptance rate for the male applicants unchanged (Table 5 in the Appendix). Bias disclosure with explanations also results in a significant increase in FPR and a significant decrease in FNR, leading to an overall insignificant change in accuracy (Figure 7c). Even despite a higher

rate of acceptance for female applicants, the increase in gender parity is not significant (Figure 7b). This is likely due to the normalizing effect of the acceptance rate of male applicants, which also increases, but not significantly.

In the case of indirect bias through a proxy feature, we find that disclosures with explanations have a positive impact on fairness both with respect to perceptions and decision-making. Similarly to what we saw in §6.2, disclosing both model bias and the association between gender and university while including explanations significantly decreases perceived model fairness (Figure 7a). In the case of fairness rating, this effect is significant even without the correlation portion of the disclosure. Further, as seen in Figure 7b, disclosing model bias alone, as well as disclosing model bias along with the correlation between protected and proxy feature with explanations leads to a significant increase in gender parity. This stems from a significantly higher acceptance rate for female applicants (Table 5 in the Appendix). The acceptance rate for male applicants remains unchanged. This also results in a higher FPR (significant) and a lower FNR (not significant), with an overall drop in accuracy (significant).

In sum, we find that in the case of direct bias, even though bias disclosure with explanations does not improve recognition of model unfairness or improve the decision-making quality or fairness significantly, it does reduce agreement with model's biased decisions, leading to a significantly higher acceptance rate for female applicants. This is in stark contrast with the effect of explanations alone (§6.1), which improved recognition of fairness, but led to more biased decisions overall.

Further, in the case of indirect bias, disclosing the model bias and correlations between protected and proxy feature with explanations significantly increases both recognition of fairness and gender parity in decision-making. We also observe an approximately 1% drop in accuracy in the case of indirect bias after disclosing model bias and correlations with explanations (which might be acceptable in certain cases). Overall, we conclude that neither explanations nor bias or correlation disclosures alone is sufficient. We observe better decision-making fairness outcomes when participants are not only shown the model explanations, but also made aware of the biases underlying them.

6.4 Effect of Joint Intervention

We have seen that explanations alone decrease decision-making fairness, while disclosures with explanations can, in the case of indirect bias, have the opposite effect. Here, we consider the effect of adding explanations and giving disclosures over including neither (i.e., comparing phase 1 without explanations to phase 2 with explanations).

We find that, for decision-making metrics, the effect of joint intervention is never significant (Table 6 in the Appendix). For fairness perception metrics, since the effect of explanations alone and disclosures with explanations pointed in the same direction, as expected, adding both explanations and disclosures also significantly improves recognition of unfairness. In sum, the joint intervention of including both explanations and disclosures (over including neither) helps participants in recognizing model biases but not in correcting them.

6.5 Effects of Additional Variables

Beyond the primary effects considered in our study, we also investigate the effect of participants' dispositional trust levels on decisions and perceptions (RQ5) and the effect of our interventions on learned trust (RQ6). We include a detailed discussion of these results in Appendix B, along with a discussion of the differences between dispositional and learned trust and the relationship between a participant's gender and their decisions and perceptions.

Does dispositional trust affect decision-making and fairness perception measures? As discussed in §5.2, we include a measurement of a participant's dispositional trust in AI as a fixed effect in our linear models. The effects and their

significance were generally not consistent across models. Overall, we find that dispositional trust does not affect fairness perception in the case of direct bias, but it indeed leads to a significantly higher perceived fairness in the case of indirect bias, that is, participants with higher levels of dispositional trust were also less able to recognize indirect bias. Additionally, we find higher dispositional trust in AI was associated with making less fair decisions by relying more on the biased model. This is in line with previous findings that a person's dispositional trust significantly affects their reliance on a machine [45]. However, we find that this effect is alleviated after including disclosures, both in the case of direct and indirect model bias.

Do explanations and disclosures affect self-reported learned trust? In addition to the fairness perception, decision-making fairness and quality measures discussed above, we additionally consider the effect of the explanation and disclosure interventions on learned trust when compared to dispositional trust in AI generally. We find that our interventions generally have no effect on learned trust ratings in models exhibiting direct bias, except for explanations leading to significantly lowered feelings that the AI system works well. When model biases are indirect, full disclosures with explanations (or sometimes full disclosures without explanations) lead to a drop in learned trust. Lastly, explanations alone and full disclosures alone also lead to an increase in the predictability of the underlying model in the case of indirect bias but not in the case of direct bias.

7 QUALITATIVE RESULTS

As described in §4.2, for a selected set of applicants in each phase, we also ask participants to write a free text response after they have made their prediction, with a justification for why they agreed or disagreed with AI (or marked it as "*neutral*"). In addition to encouraging careful thinking, this also helps us gauge the kinds of reasoning participants employ in their decision-making.

As the main goal of our study is to understand how humans interact with AI decisions when the AI is biased, we primarily focus our qualitative analysis on rationales concerning biases. To analyze how participants perceive and use (or discard) the biased feature (gender or university), we consider justifications that directly reference the protected (gender) or proxy feature (university) by using a set of keywords for both. For keywords, we started with an initial set (e.g., "gender", "female", "university") and, based on reading a subset of the justifications, expanded to include spelling variations (e.g., "skool" and "collage") and other topically relevant words (e.g., abbreviated names of schools). We discuss our qualitative findings on justifications involving the protected feature in §7.1 and justifications involving the proxy feature in §7.2. Lastly, we discuss additional observations indicating over-reliance based on a random sample of justifications in §7.3.

7.1 Justification Involving Protected Feature

Here, we analyze justifications that explicitly mention "gender", "male", "female", etc. in the direct bias conditions. First, we consider such justifications before we disclose model bias to participants. In this phase, when we do not provide participants with explanations, we almost never observe participants directly discussing "gender" as a salient part of their justification. When they are given explanations, participants mention trying to "ignore gender" when making their decision. We observe that participants who mentioned gender in a prediction about a female applicant tended to be making "Complete" or "Neutral" predictions. Thus, even though explanations significantly decreased gender parity (§6.1) overall, they do seem to help certain participants correct model biases in some cases.

After bias disclosure, justifications mentioning gender still predominantly appear in the condition with explanations. However, we do find some participants who, even without explanations, mention gender bias and flip "Late" predictions for female applicants. For example, one participant explained they overrode such a prediction because they had seen an example of a male applicant with the same occupation who was predicted as "Complete". In a condition with explanations, many participants mention that "*gender should not be a deciding factor*" and that they try to ignore this feature when making their prediction. Some participants even explain when they *are* agreeing with a "Late" prediction for a female applicant that their decision was based on other features ("*The large amount of negatives aside from gender still point towards being late.*"). However, even in the case of direct bias and with both explanations and bias disclosure, we still observe participants who agree with model biases. For example, one participant agreed with a "Complete" prediction of a male applicant "*[b]ecause according to AI, male gender is more likely to complete loan...*"

7.2 Justifications Involving Proxy Feature

Here, we analyze justifications that explicitly mention "university", "college", "school", etc. in the indirect bias conditions. First, we consider how participants used the "university" feature in their justifications before any disclosures of model bias or correlation. Overall, some participants mention that attending college in the first place generally increases the likelihood of repayment, regardless of the specific school (for example, mentioning that an applicant is "college educated" and predicting "Complete").

However, other participants do consider the specific college as a reason to accept or reject an applicant. In the conditions without explanations, we see evidence of participants relying on their own judgment of school quality. In the underlying model, going to a co-ed school is counted as a positive. We instead see participants mentioning that they have "*Never heard of Kenyon College*," or that the co-ed school in the application "*...is not a particularly prestigious university.*" In contrast, when participants are provided explanations, we instead see examples of participant's evaluation of a school aligning with the model's biases. For example, on applicants from women's colleges, participants claimed "*The applicant didn't go to a good college*," or that "*...College history was a major contributing factor to being late on loan*," while on applications from co-ed schools, participants claimed that the applicant "*...attended a good university*," or mention that "*the university is listed as a good one.*" This supports our quantitative finding that explanations alone lead participants to be more influenced by model biases in their decision-making (§6.1).

Next, we consider how participants talk about the "university" feature after disclosures. With bias disclosure alone (and especially without explanations), mentions of university were quite sparse. Some participants mention that university is given too much weight ("*I just find it hilarious that the borrower's state and university is such a huge factor.*") but may not make the connection that this over-use is due to indirect bias. One participant knew outside of the study that Bryn Mawr is a women's college, but they were still generally not confident about how to find biased predictions without direct access to protected features: "*After learning more about possible discriminatory predictions on the AI's part... I'm specifically concerned about gender and race… but don't quite know how to discern that from these charts... This applicant profile gave me pause because I *think* Bryn Mawr College is an all-women's college.*"

When participants were instead given the full bias and correlation disclosure, the "university" feature appears frequently in decision justifications. Here, participants continue to point out that university is given too much weight in explanations generally but also point to undue negative weight towards women's colleges. (e.g., "*While the system said late, I thought this was unfair because it placed a strong negative value on the college, which might be a women's college.*"). While some participants bring up this university bias as a justification for flipping model predictions, many acknowledge the bias and make a neutral prediction or agree with predictions of women being late in repayment. We

		Explanations Only		Disclosures Without Explanations			Disclosures With Explanations			Joint Intervention		
		Prot	Prox	Prot BD	Prox BD	Prox BD+CD	Prot BD	Prox BD	Prox BD+CD	Prot BD	Prox BD	Prox BD+CD
Fairness Perception	Fairness Rating	↓	·	·	·	↓	·	↓	↓	↓	↓	↓
	Fairness Saliency	↑	·	·	·	↑	·	·	↑	↑	·	↑
DM Fairness	Gender Parity	↓	↓	·	·	·	·	↑	↑	·	·	·
DM Quality	Accuracy	↑	↑	·	·	↓	·	↓	↓	·	·	·
	FNR	↑	·	·	·	·	↓	·	·	·	·	·
	FPR	↓	↑	·	·	↑	↑	↑	↑	·	·	·

Table 1. Summary of our main results. Arrows represent significant effects and point in the direction of the change. "BD" and "CD" represent bias and correlation disclosures, respectively.

also find that some participants struggled to recall which universities were co-ed, and others misinterpreted the co-ed schools as being predominantly male which may have limited their ability to intervene to correct model biases.

7.3 Justifications Indicating Over-reliance

In addition to the positive examples of explanations and disclosures helping participants notice and correct model biases, in many instances, as gleaned through their justifications, we also find instances of participants making predictions entirely based on the AI prediction or the corresponding explanations, regardless of disclosures. For instance, even after bias disclosure, one participant in a direct bias condition agreed with a prediction of a female applicant being "Late" saying that "*They seem to have more negatives than positives.*" Similarly, even after bias and correlation disclosures, a participant in an indirect bias condition changed the prediction for an applicant from a women's college from "Complete" to "Late", making a similar justification. For the applicants in question, however, if the biased features (gender and university, respectively) were to be ignored, the positives in fact would outweigh the negatives. This indicates that some participants continue to focus their decision-making process on explanations that they know contain biases.

We also find instances of over-trust in AI even after being told that AI is biased such as "*I have no reason to disagree with the AI, if the AI is discriminating it probably has a good reason to,*" or "*An AI is usually better than a professional let alone an amateur like me.*" This indicates that even despite explanations and disclosures, there is much room for improvement in educating and training humans to avoid unwarranted trust in AI systems and promote fair decision-making.

8 DISCUSSION AND LIMITATIONS

In this work, we studied the effect of explanations and disclosures on fairness perceptions and decision-making when humans are provided predictions from models exhibiting direct or indirect bias. Our findings are summarized in Table 1. Regardless of intervention, we consistently observed that human-AI teams made fairer decisions than the AI alone. We found that explanations alone significantly improved participants' ability to notice unfairness in the case of direct bias only. However, explanations led participants to be more influenced by model biases, whether they noticed these biases or not. Disclosures were an effective tool for helping users recognize unfairness in the case of indirect bias, especially with the help of explanations. And we saw that this increased recognition of bias was paired with fairer human-AI decisions, showing that disclosures helped participants understand when and how to intervene on model decisions to produce fairer outcomes.

However, we found that the joint intervention of including both explanations and disclosures (over including neither) was only effective in helping participants recognize model bias, not correct it. If the main objective is to help users notice model unfairness in the case of direct bias, we recommend including explanations, and if it is to help users notice model unfairness in the case of indirect bias, we recommend including explanations and disclosing both model bias and the correlations between protected and proxy features. But if the main objective is to help the human-AI team produce fairer outcomes, we did not find including explanations with disclosures to be an effective intervention. However, if explanations are to be used, then disclosures may help contextualize explanations and the potential biases, especially when these biases are indirect. While in a perfect world, such known biases could be addressed in the model itself instead of relying on human intervention, this may not always be possible. In many cases, we may have limited access to the underlying model (e.g., only having API access) or may not be able to non-superficially "debias" it [28]. Disclosures may help uncover these biases to humans, possibly leading to fairer human-AI decisions.

A key limitation of work is that since we show our participants partially-synthetic loan data, we cannot directly rely on the existing ground truth. Instead, we calculate the expectation of ground-truth based metrics (accuracy, FNR, and FPR) which means that there are applicants for which neither choice is very likely to be "correct" (i.e., both $P(Y_i = 1)$ and $P(Y_i = 0)$ are close to 0.5). We handle this in part by adjusting for the baseline AI-only scores; however, using a fully non-synthetic dataset and original ground truth values may lead to cleaner results. This lack of a true ground-truth, in part, led us to use demographic parity which has been argued to be insufficient as a notion of fairness [22].

Another limitation is that our study design forces participants to make decisions one at a time, without seeing the entire pool of applicants. It is our hope that the percent/percentile information given for each feature helped give participants a better sense of how each applicant's profile compared to the general pool, even without seeing many profiles. However, we recognize that it may be difficult for participants to conceptualize what a "strong" or "weak" candidate looks like under this design. This may make it more difficult for participants who, for example, wish to increase the acceptance rate of women in phase 2 to decide which female applicants are "most deserving" of having their prediction flipped to "Complete".

Despite these limitations, our work provides insights into the effect of explanations on fairness in human-AI decision-making, especially when the biases are indirect (through a proxy features). We conclude that neither explanations nor disclosures alone improve the fairness of decisions made by a human-AI team. Our findings serve to caution the wider community from treating explanations as a foolproof solution to human-AI collaborative decision-making: explanations may not always make model biases clear and may make people more prone to align with model biases, leading to less fair decisions. When people are repeatedly exposed to explanations that justify or rationalize biased predictions, they may begin to accept these biases as valid or even desirable, rather than critically questioning and challenging them. We highlight that explanations and disclosures in conjunction may be helpful to some extent. However, more work is needed to further examine how best to aid humans not only in identifying indirect model biases, but also in systematically correcting these biases.

ACKNOWLEDGMENTS

We sincerely thank the CHI TRAIT reviewers, Md Naimul Hoque, and the members of the UMD CLIP and HCIL labs for their valuable feedback.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. http://www.jstor.org/stable/24758720

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[6] 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 (September 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[7] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[10] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 95–106. https://doi.org/10.1609/icwsm.v14i1.7282

[11] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 251–263. https://doi.org/10.1145/3581641.3584080

[12] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. arXiv:2301.07255 [cs.HC]

[13] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[14] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 348, 18 pages. https://doi.org/10.1145/3544548.3581015

[15] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *Proceedings of the 13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 120–129. https://doi.org/10.1145/3447535.3462487

[16] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047 PMID: 28632438.

[17] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *CoRR* abs/2006.11371 (2020). arXiv:2006.11371 https://arxiv.org/abs/2006.11371

[18] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[19] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/ARXIV.1702.08608

[20] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2021. Fairness in Deep Learning: A Computational Perspective. *IEEE Intelligent Systems* 36, 4 (2021), 25–34. https://doi.org/10.1109/MIS.2020.3000681

[21] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2022. The Influences of Task Design on Crowdsourced Judgement: A Case Study of Recidivism Risk Evaluation. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1685–1696. https://doi.org/10.1145/3485447.3512239

[22] 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* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[23] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7 Trust and Technology.

[24] Equal Employment Opportunity Commission. 1978. Uniform Guidelines on Employee Selection Procedures. Code of Federal Regulations, Title 29, § 1607.4. https://www.law.cornell.edu/cfr/text/29/1607.4 43 FR 38295, 38312, Aug. 25, 1978, as amended at 46 FR 63268, Dec. 31, 1981.

[25] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[26] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[27] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[28] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

[29] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1103–1116. https://doi.org/10.18653/v1/2021.findings-acl.95

[30] Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare R. Voss, Marine Carpuat, and Hal Daumé III. 2023. What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on AI Systems. arXiv:2305.14331 [cs.CL]

[31] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1 (1995), 4. https://doi.org/10.1037/0033-295X.102.1.4

[32] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 PMID: 25875432.

[33] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[34] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf

[35] Joon Sik Kim, Valerie Chen, Danish Pruthi, Nihar B. Shah, and Ameet Talwalkar. 2023. Assisting Human Decisions in Document Matching. arXiv:2302.08450 [cs.LG]

[36] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[37] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[38] 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* (Atlanta, GA, USA) *(FAccT '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[39] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[40] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (nov 2019), 26 pages. https://doi.org/10.1145/3359284

[41] Joseph Lev. 1949. The Point Biserial Coefficient of Correlation. *The Annals of Mathematical Statistics* 20, 1 (1949), 125–126. http://www.jstor.org/stable/2236816

[42] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[43] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. https://doi.org/10.1145/3479552

[44] Stephen Marsh and Mark R. Dibben. 2003. The role of trust in information science and technology. *Annual Review of Information Science and Technology* 37, 1 (2003), 465–498. https://doi.org/10.1002/aris.1440370111

[45] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors* 50, 2 (2008), 194–210. https://doi.org/10.1518/001872008X288574 PMID: 18516832.

[46] Arvind Narayanan. 2018. Translation Tutorial: 21 Fairness Definitions and Their Politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, Vol. 1170. New York, USA, 3.

[47] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (mar 2022), 33 pages. https://doi.org/10.1145/3495013

[48] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[49] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[50] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[51] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. https://doi.org/10.1145/3491102.3501967

[52] Federal Reserve. 2006. *Consumer Compliance Handbook*. Chapter Fair Lending Regulations and Statutes: Overview. https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_over.pdf

[53] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 8377–8387. https://proceedings.mlr.press/v119/saha20c.html

[54] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2209.11812* (2022). https://arxiv.org/abs/2209.11812

[55] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[56] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[57] 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* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 315–324. https://doi.org/10.1145/3394231.3397922

[58] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[59] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3076–3084. https://doi.org/10.24963/ijcai.2023/343 Main Track.

[60] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[61] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 27 (nov 2022), 36 pages. https://doi.org/10.1145/3519266

[62] Richard Warner and Robert H. Sloan. 2021. Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables. *Criminal Justice Ethics* 40, 1 (2021), 23–39. https://doi.org/10.1080/0731129x.2021.1893932

[63] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[64] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161

[65] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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* (Barcelona, Spain) *(FAccT '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

A DERIVING METRICS USING PROBABILISTIC GROUND TRUTH

Here, we consider in more detail how to calculate the probability of ground-truth completion and the expected value of the decision quality metrics.

A.1 Probability of Loan Completion

In our study, each applicant i is only evaluated by a single participant j in a given condition, and each participant j evaluates 20 applicants across the two phases (§4.2). We represent the set of observed decisions as $S = \{(i, j) \mid \text{participant } j \text{ sees applicant } i\}$. For the i^{th} applicant, we want to know the probability that the true outcome should be complete, that is, $P(Y_i = 1)$. Let $\boldsymbol{x}_i$ be the set of original features of the i^{th} applicant's and x_i^* be the assigned (synthetic) gender or university. Note, we drop the subscript i when referring to a general applicant. We can write the probability of the true outcome for applicant with features $(\boldsymbol{x}, x^*)$ as

$$P(Y = 1 \mid \boldsymbol{x}, x^*) = \frac{P(x^* \mid Y = 1, \boldsymbol{x})P(Y = 1 \mid \boldsymbol{x})}{P(x^* \mid \boldsymbol{x})}$$

Since we assign the values of the protected or proxy feature x^* based solely on the ground-truth outcome, we can assume that x^* and $\boldsymbol{x}$ are independent given y. So we say that $P(x^* \mid Y = 1, \boldsymbol{x}) = P(x^* \mid Y = 1)$.

$$\begin{aligned} P(Y = 1 \mid \boldsymbol{x}, x^*) &= \frac{P(x^* \mid Y = 1)P(Y = 1 \mid \boldsymbol{x})}{P(x^* \mid \boldsymbol{x})} \\ &= \frac{P(Y = 1 \mid x^*)P(x^*)}{P(Y = 1)} \frac{P(Y = 1 \mid \boldsymbol{x})}{P(x^* \mid \boldsymbol{x})} \end{aligned}$$

Then, removing any terms not containing Y, which can be normalized away, we are left with

$$P(Y = 1 \mid \boldsymbol{x}, x^*) = \frac{P(Y = 1 \mid x^*)P(Y = 1 \mid \boldsymbol{x})}{P(Y = 1)}$$

In the protected case, we know the probability of acceptance given the synthetic feature (that is, $P(Y = 1 \mid x^*)$) based on our selected 60/40 male/female acceptance ratio. In the proxy case, we estimate the probability of acceptance given the university using the joint distribution of gender and university and the probability of acceptance given gender. For each applicant, we estimate the probability of the applicant completing their loan $P(Y = 1 \mid \boldsymbol{x})$ based on only the non-synthetic features $(\boldsymbol{x})$ using a linear regression model that does not have access to the synthetic information. Finally, we can calculate $P(Y = 1)$ based on the rate of ground-truth acceptances in the original data.

A.2 Expected Accuracy, FNR, and FPR

Using our calculated probability of the ground-truth acceptance of a given applicant (with original features $\boldsymbol{x}$ and synthetic feature x^*), we can calculate an expected accuracy, FNR, and FPR for a given human-AI team.

For example, for expected FNR, we consider the expected number of false negatives over the expected number of ground truth positives.

$$\text{Expected FNR} = \frac{\mathbb{E}[\text{\# False Negatives}]}{\mathbb{E}[\text{\# Positives}]}$$

Let $\hat{y}_{i,j}$ be the human-AI decision for the i^{th} applicant by the j^{th} participant, such that $\hat{y}_{i,j}$ is 1 if the human-AI decision is "Complete" and 0 if it is "Late". We can write the expected number of false negatives $\mathbb{E}[\text{\# False Negatives}]$ as $\sum_{(i,j)\in S}(1-\hat{y}_{i,j})P(Y_i=1|\boldsymbol{x}_i, x_i^*)$, that is, the probability of the ground truth label for the i^{th} applicant being 1 but the human-AI decision for the same applicant being 0. Similarly, we can write the expected value of positives $\mathbb{E}[\text{\# Positives}]$ as $\sum_i P(Y_i=1|\boldsymbol{x}_i, x_i^*)$. Thus, we can write

$$\text{Expected FNR} = \frac{\sum_{(i,j)\in S} P(Y_i=1|\boldsymbol{x}_i, x_i^*) \times (1-\hat{y}_{i,j})}{\sum_i P(Y_i=1|\boldsymbol{x}_i, x_i^*)}.$$

Using a similar process, we can calculate the expected FPR and Accuracy.

$$\begin{aligned}\text{Expected FPR} &= \frac{\mathbb{E}[\text{\# False Positives}]}{\mathbb{E}[\text{\# Negatives}]} \\ &= \frac{\sum_{(i,j)\in S} P(Y_i=0|\boldsymbol{x}_i, x_i^*) \times \hat{y}_{i,j}}{\sum_i P(Y_i=0|\boldsymbol{x}_i, x_i^*)}\end{aligned}$$

$$\begin{aligned}\text{Expected Accuracy} &= \frac{\mathbb{E}[\text{\# True Positives + \# True Negatives}]}{\mathbb{E}[\text{\# Samples}]} \\ &= \frac{1}{|S|}\sum_{(i,j)\in S}\mathbb{E}[TP_{i,j}+TN_{i,j}] \\ &= \frac{1}{|S|}\sum_{(i,j)\in S}\left(P(Y_i=1|\boldsymbol{x}_i, x_i^*)\times\hat{y}_{i,j}\right) + \left(P(Y_i=0|\boldsymbol{x}_i, x_i^*)\times(1-\hat{y}_{i,j})\right)\end{aligned}$$

B EXTENDED RESULTS

In this section, we report additional results regarding dispositional trust, learned trust, and participant gender. We additionally include a full table detailing the primary effects considered in the study (Table 6) as well as the effects of our interventions on reliance split by applicant gender (Table 5).

B.1 Does dispositional trust affect decision-making and fairness perception measures?

As discussed in section 5.2, we include a measurement of a participant's dispositional trust in AI as a fixed effect in our linear models. The effects and their significance was generally not consistent across models (See Table 2).

Metric	Bias Type	Model	F	p		Coef	Std Error
Fairness Rating	protected	Expl	$F(1, 115) = 4.07$	0.0459		1.31	0.592
		Disclosure	$F(1, 49) = 5.86$	0.0192		1.70	0.700
		Disclosure with Explanation	$F(1, 66) = 1.6$	0.2104		0.89	0.702
	proxy	Expl	$F(1, 227) = 32.7$	0.0000	*	1.54	0.280
		Disclosure	$F(1, 91) = 25.8$	0.0000	*	1.69	0.332
		Disclosure with Explanation	$F(1, 136) = 10.11$	0.0018	*	1.13	0.355
Fairness Saliency	protected	Expl	$F(1, 115) = 0.87$	0.3540		−0.31	0.284
		Disclosure	$F(1, 49) = 0.18$	0.6737		−0.15	0.354
		Disclosure with Explanation	$F(1, 66) = 0.85$	0.3598		−0.30	0.331
	proxy	Expl	$F(1, 227) = 2.52$	0.1135		−0.12	0.087
		Disclosure	$F(1, 91) = 9.11$	0.0033	*	−0.45	0.150
		Disclosure with Explanation	$F(1, 136) = 0.01$	0.9361		−0.01	0.125
Parity	protected	Expl	$F(1, 115) = 6.25$	0.0138	*	−0.35	0.149
		Disclosure	$F(1, 99) = 0.53$	0.4681		−0.15	0.208
		Disclosure with Explanation	$F(1, 133) = 3.27$	0.0730		−0.28	0.153
	proxy	Expl	$F(1, 227) = 5.73$	0.0175	*	−0.22	0.112
		Disclosure	$F(1, 182) = 0.31$	0.5778		0.07	0.128
		Disclosure with Explanation	$F(1, 135) = 3.55$	0.0618		−0.19	0.100
Accuracy	protected	Expl	$F(1, 115) = 2.09$	0.1511		0.04	0.031
		Disclosure	$F(1, 49) = 4.71$	0.0349		0.10	0.044
		Disclosure with Explanation	$F(1, 66) = 0.23$	0.6366		0.02	0.033
	proxy	Expl	$F(1, 227) = 2.97$	0.0863		0.02	0.018
		Disclosure	$F(1, 91) = 1.73$	0.1917		0.03	0.026
		Disclosure with Explanation	$F(1, 136) = 2.22$	0.1382		0.03	0.019
FNR	protected	Expl	$F(1, 115) = 0.06$	0.8126		0.01	0.096
		Disclosure	$F(1, 49) = 0.01$	0.9113		−0.01	0.125
		Disclosure with Explanation	$F(1, 66) = 0.53$	0.4674		0.08	0.104
	proxy	Expl	$F(1, 227) = 8.8$	0.0033	*	0.14	0.054
		Disclosure	$F(1, 91) = 0.6$	0.4400		0.06	0.072
		Disclosure with Explanation	$F(1, 136) = 9.69$	0.0023	*	0.20	0.063
FPR	protected	Expl	$F(1, 115) = 0.7$	0.4051		−0.08	0.107
		Disclosure	$F(1, 49) = 0.73$	0.3983		−0.12	0.144
		Disclosure with Explanation	$F(1, 66) = 0.59$	0.4458		−0.10	0.125
	proxy	Expl	$F(1, 227) = 9.44$	0.0024	*	−0.16	0.059
		Disclosure	$F(1, 91) = 2.07$	0.1533		−0.12	0.085
		Disclosure with Explanation	$F(1, 136) = 8.79$	0.0036	*	−0.22	0.073

Table 2. Effects of dispositional trust in AI on different outcome metrics (Perception, Parity, Accuracy, FPR, and FNR).

We consistently find that dispositional trust has no significant impact on fairness ratings in the protected conditions, while it significantly increases fairness ratings in the proxy conditions. In other words, when biases are direct, people are equally able to notice model biases even when they tend to trust AI in general; however, when biases are indirect, people with higher dispositional trust in AI are less likely to believe that the model is unfair. For participants' fairness saliency, we see that there is never a significant effect in the case of direct bias, but higher dispositional trust significantly decreased the rate of disagreement only when considering disclosures without explanations.

We also find that, under models that consider the effect of explanations and disclosures with explanations, increases dispositional trust in AI significantly increased FPR and decreased FNR in the proxy conditions only. This is likely due

				better		confidence		predictable		safe		wary		works well	
Feature	Phase	Expl?	Disclosure	sig	coef	sig	coef	sig	coef	sig	coef	sig	coef	sig	coef
Protected	1	-	-	*	−0.61	*	−0.43	*	−0.51	*	−0.29		0.04	*	−0.43
	1	✓	-	*	−0.93	*	−0.66	*	−0.37		−0.26		0.10	*	−0.81
	2		Bias	*	−0.33	*	−0.23	*	−0.25	*	−0.20		0.02	*	−0.28
	2	✓	Bias	*	−0.51	*	−0.38	*	−0.19	*	−0.22		0.12	*	−0.48
Proxy	1	-	-	*	−0.55	*	−0.34	*	−0.68	*	−0.22		−0.15	*	−0.32
	1	✓	-	*	−0.54		−0.19		0.01		0.03		−0.14	*	−0.29
	2	-	Bias	*	−0.39	*	−0.21	*	−0.29		−0.05		0.02	*	−0.25
	2	-	Full	*	−0.31	*	−0.28	*	−0.23		−0.13		0.01	*	−0.37
	2	✓	Bias	*	−0.30	*	−0.29		0.02		−0.09		0.06	*	−0.30
	2	✓	Full	*	−0.36	*	−0.20		−0.04		−0.08		0.00	*	−0.26

Table 3. Comparison of trust in AI generally vs trust in our models in varied conditions and phases.

to participants with higher trust in AI being more influenced by subtle indirect biases leading to lower acceptance rates for female applicants. This is also supported by the models measuring the effect of explanations on gender parity. Here, we see that an increased dispositional trust in AI significantly decreased parity under both types of bias.

Overall, people with a greater dispositional trust in AI tended to make more unfair decisions (when working with a biased model) and were less likely to notice indirect bias.

B.2 Do explanations and disclosures affect learned trust?

In this section, we discuss the effect of our interventions—explanations, disclosures without explanations, and disclosures with explanations (Figure 4)—on learned trust over dispositional trust in AI generally. We perform statistical tests similar to the ones described in §5.2. We consider the different trust measures as the dependent variable and the treatment as the fixed effect term. We also control for the dispositional trust level as a fixed effect.

As seen in Table 6, we find that our treatments generally have no effect on trust ratings in models exhibiting direct bias, except for explanations alone leading to significantly lowered feelings that the AI system works well. In the case of indirect bias, we often see that full disclosure with explanations (and sometimes also full disclosure without explanations or bias disclosure with explanations) has a significant effect on learned trust. These effects demonstrate lowered feelings that the AI system works well as well as decreased feelings that the AI system can perform as well as an untrained human, decreased confidence in the system, decreased feelings of safety when relying on the system, and increased wariness of the AI system.

We also find that explanations alone significantly increase participant's perception of model predictability in the proxy conditions but not the protected conditions. Without explanations, full bias and correlation disclosure also significantly increased predictability. Likely due to explanations increasing predictability on its own, no disclosure significantly affected predictability when paired with explanations. This is to say that our models are already seen as relatively predictable when biases are direct, but when biases are indirect, explanations or disclosure of model bias and the model's usage of the university feature help make the model more predictable.

	Metric	Coefficient	p
Fairness Perception	Fairness Rating	0.045	0.403
	Fairness Saliency	−0.026	0.624
DM Fairness	Gender Parity	−0.058	0.280
DM Quality	Accuracy	0.094	0.079
	FNR	0.083	0.120
	FPR	−0.107	0.045
Acceptance Rate	Female Applicants	−0.102	0.057
	Male Applicants	−0.068	0.204

Table 4. Correlation between participants self-describing as male and various performance metrics.

B.3 Does dispositional trust differ from learned trust?

In the previous section, we discussed how interventions affected learned trust when controlling for baseline dispositional trust. Here, we study whether there is a significant difference between dispositional trust and learned trust in the biased models across the questions described in §4.2. These results are shown in Table 3.

We find that participants usually thought our model worked worse than AI does generally, that it inspired less confidence, and was less predictable. Participants also regarded our AI system as less safe than AI in general, but this is primarily true only in the case of direct bias. Surprisingly, participants did not consider our AI systems to be less safe than general in phase 1 when they were given explanations (which would have directly indicated that the system used gender as a feature to determine loan outcomes). Participants' wariness of the biased models was not significantly different from their baseline wariness in AI.

B.4 Does participant gender correlate with decision-making and fairness perception measures?

Because our models exhibits gender bias, it stands to reason that participants of varied gender may react differently to the models. Namely, non-male participants may be more sensitive to bias against women. Using point-biserial correlation tests [41], we consider whether gender[7] correlates with our fairness perception, decision-making fairness, and decision-making quality metrics as well as the rate of "Complete" predictions for female and male applicants directly.

We find no significant correlations between gender and behavior or perceptions in our task (See Table 4). However, we do find marginally significant correlations with acceptance rate for female applicants, FPR, and accuracy. This shows there may be a weak trend in male participants accepting fewer female candidates ($p = 0.05$), leading to a lower FPR and higher accuracy.

[7]Here, we use a binary indicator variable of whether a participant's self-reported gender included the "male" checkbox. We did not have enough non-binary or gender non-conforming participants to analyze separately.

		Female		Male	
Feature	Intervention	sig	coef	sig	coef
Protected	+Expl	*	−0.13		−0.04
	+Bias Disclosure		0.06		0.00
	+Bias Disclosure with Explanation	*	0.08		0.02
Proxy	+Expl	*	−0.07		0.00
	+Bias Disclosure		0.03		0.01
	+Bias and Corr Disclosure		0.08		0.01
	+Bias Disclosure with Explanation	*	0.07		0.00
	+Bias and Corr Disclosure with Explanation	*	0.10		−0.03

Table 5. Effect of interventions on acceptance rate for female and male applicants across conditions and phases.

Metric	Bias Type	Effect	F	p		Coef	Std Error
Fairness Rating	protected	+Expl	$F(1, 115) = 26.33$	0.0000	*	−0.89	0.174
		+Bias Disclosure	$F(1, 50) = 5.43$	0.0238		−0.29	0.126
		+Bias Disclosure with Explanation	$F(1, 67) = 0.01$	0.9091		−0.01	0.128
		Joint Intervention (BD)	$F(1, 208) = 73.06$	0.0000	*	−1.01	0.118
	proxy	+Expl	$F(1, 227) = 1.51$	0.2207		0.11	0.088
		+Bias Disclosure	$F(1, 122) = 3.54$	0.0623		−0.21	0.114
		+Bias and Corr Disclosure	$F(1, 125) = 12.84$	0.0005	*	−0.42	0.116
		+Bias Disclosure with Explanation	$F(1, 178) = 19.84$	0.0000	*	−0.43	0.096
		+Bias and Corr Disclosure with Explanation	$F(1, 178) = 46.57$	0.0000	*	−0.66	0.096
		Joint Intervention (BD)	$F(1, 277) = 6.14$	0.0138	*	−0.27	0.109
		Joint Intervention (BD+CD)	$F(1, 277) = 13.19$	0.0003	*	−0.46	0.109
Fairness Saliency	protected	+Expl	$F(1, 115) = 15.83$	0.0001	*	0.33	0.083
		+Bias Disclosure	$F(1, 50) = 2.33$	0.1330		0.10	0.064
		+Bias Disclosure with Explanation	$F(1, 67) = 1.19$	0.2785		0.07	0.067
		Joint Intervention (BD)	$F(1, 208) = 72.7$	0.0000	*	0.47	0.055
	proxy	+Expl	$F(1, 227) = 1.21$	0.2733		−0.03	0.028
		+Bias Disclosure	$F(1, 122) = 0.67$	0.4154		0.04	0.050
		+Bias and Corr Disclosure	$F(1, 125) = 9.19$	0.0030	*	0.16	0.052
		+Bias Disclosure with Explanation	$F(1, 182) = 0.62$	0.4311		0.03	0.037
		+Bias and Corr Disclosure with Explanation	$F(1, 182) = 35.06$	0.0000	*	0.22	0.037
		Joint Intervention (BD)	$F(1, 277) = 0.58$	0.4458		−0.04	0.047
		Joint Intervention (BD+CD)	$F(1, 277) = 13.42$	0.0003	*	0.15	0.048

(*table continues*)

Metric	Bias Type	Effect	F	p		Coef	Std Error
Parity	protected	+Expl	$F(1, 115) = 8.32$	0.0047	*	−0.13	0.044
		+Bias Disclosure	$F(1, 99) = 2.58$	0.1117		0.09	0.055
		+Bias Disclosure with Explanation	$F(1, 133) = 4.08$	0.0455		0.09	0.047
		Joint Intervention (BD)	$F(1, 208) = 0.06$	0.8096		0.01	0.042
	proxy	+Expl	$F(1, 227) = 8.58$	0.0038	*	−0.10	0.035
		+Bias Disclosure	$F(1, 182) = 0.77$	0.3815		0.04	0.050
		+Bias and Corr Disclosure	$F(1, 182) = 1.29$	0.2574		0.06	0.051
		+Bias Disclosure with Explanation	$F(1, 188) = 7.75$	0.0059	*	0.10	0.035
		+Bias and Corr Disclosure with Explanation	$F(1, 188) = 20.26$	0.0000	*	0.16	0.035
		Joint Intervention (BD)	$F(1, 277) = 0.78$	0.3785		−0.03	0.038
		Joint Intervention (BD+CD)	$F(1, 277) = 1.01$	0.3149		0.03	0.038
Accuracy	protected	+Expl	$F(1, 115) = 20.89$	0.0000	*	0.04	0.009
		+Bias Disclosure	$F(1, 50) = 1.28$	0.2630		−0.01	0.009
		+Bias Disclosure with Explanation	$F(1, 67) = 0.79$	0.3772		−0.01	0.006
		Joint Intervention (BD)	$F(1, 208) = 3.5$	0.0628		0.01	0.007
	proxy	+Expl	$F(1, 227) = 7.3$	0.0074	*	0.02	0.006
		+Bias Disclosure	$F(1, 123) = 0.02$	0.8846		0.00	0.009
		+Bias and Corr Disclosure	$F(1, 126) = 6.66$	0.0110	*	−0.02	0.010
		+Bias Disclosure with Explanation	$F(1, 182) = 9.18$	0.0028	*	−0.02	0.006
		+Bias and Corr Disclosure with Explanation	$F(1, 182) = 10.26$	0.0016	*	−0.02	0.006
		Joint Intervention (BD)	$F(1, 277) = 2.44$	0.1197		0.01	0.007
		Joint Intervention (BD+CD)	$F(1, 277) = 0.73$	0.3933		0.01	0.007
FNR	protected	+Expl	$F(1, 115) = 5.88$	0.0169	*	0.07	0.028
		+Bias Disclosure	$F(1, 50) = 0.38$	0.5426		−0.02	0.026
		+Bias Disclosure with Explanation	$F(1, 67) = 7.78$	0.0069	*	−0.05	0.017
		Joint Intervention (BD)	$F(1, 208) = 0.05$	0.8269		0.00	0.022
	proxy	+Expl	$F(1, 227) = 3.78$	0.0530		0.03	0.017
		+Bias Disclosure	$F(1, 113) = 0.48$	0.4921		−0.01	0.018
		+Bias and Corr Disclosure	$F(1, 115) = 2.95$	0.0887		−0.03	0.019
		+Bias Disclosure with Explanation	$F(1, 169) = 1.6$	0.2073		−0.02	0.014
		+Bias and Corr Disclosure with Explanation	$F(1, 169) = 1.65$	0.2006		−0.02	0.014
		Joint Intervention (BD)	$F(1, 277) = 1.21$	0.2722		0.02	0.020
		Joint Intervention (BD+CD)	$F(1, 277) = 1.39$	0.2396		0.03	0.020

(*table continues*)

Metric	Bias Type	Effect	F	p		Coef	Std Error
FPR	protected	+Expl	$F(1, 115) = 9.94$	0.0021	*	−0.10	0.031
		+Bias Disclosure	$F(1, 50) = 2.09$	0.1546		0.04	0.028
		+Bias Disclosure with Explanation	$F(1, 67) = 8.35$	0.0052	*	0.05	0.019
		Joint Intervention (BD)	$F(1, 208) = 0.17$	0.6832		−0.01	0.025
	proxy	+Expl	$F(1, 227) = 5.45$	0.0205		−0.04	0.019
		+Bias Disclosure	$F(1, 116) = 1.34$	0.2502		0.03	0.024
		+Bias and Corr Disclosure	$F(1, 118) = 5.95$	0.0162	*	0.06	0.024
		+Bias Disclosure with Explanation	$F(1, 169) = 10.79$	0.0012	*	0.05	0.017
		+Bias and Corr Disclosure with Explanation	$F(1, 169) = 7.44$	0.0071	*	0.05	0.017
		Joint Intervention (BD)	$F(1, 277) = 0.06$	0.8092		−0.01	0.023
		Joint Intervention (BD+CD)	$F(1, 277) = 0.84$	0.3593		−0.02	0.023
better	protected	+Expl	$F(1, 116) = 2.57$	0.1117		−0.32	0.199
		+Bias Disclosure	$F(1, 50) = 0.39$	0.5369		−0.06	0.095
		+Bias Disclosure with Explanation	$F(1, 67) = 1.06$	0.3070		−0.09	0.086
	proxy	+Expl	$F(1, 228) = 0.0$	0.9726		0.00	0.143
		+Bias Disclosure	$F(1, 101) = 5.0$	0.0276		−0.21	0.095
		+Bias and Corr Disclosure	$F(1, 102) = 0.77$	0.3838		−0.09	0.098
		+Bias Disclosure with Explanation	$F(1, 151) = 0.07$	0.7884		0.02	0.085
		+Bias and Corr Disclosure with Explanation	$F(1, 151) = 10.12$	0.0018	*	−0.27	0.085
confidence	protected	+Expl	$F(1, 116) = 1.71$	0.1933		−0.23	0.176
		+Bias Disclosure	$F(1, 50) = 0.04$	0.8496		−0.02	0.103
		+Bias Disclosure with Explanation	$F(1, 67) = 1.2$	0.2765		−0.09	0.080
	proxy	+Expl	$F(1, 228) = 1.41$	0.2356		0.16	0.131
		+Bias Disclosure	$F(1, 109) = 0.5$	0.4798		−0.08	0.106
		+Bias and Corr Disclosure	$F(1, 110) = 3.66$	0.0583		−0.21	0.109
		+Bias Disclosure with Explanation	$F(1, 149) = 13.17$	0.0004	*	−0.28	0.076
		+Bias and Corr Disclosure with Explanation	$F(1, 149) = 19.32$	0.0000	*	−0.33	0.076
predicatable	protected	+Expl	$F(1, 116) = 0.48$	0.4883		0.14	0.204
		+Bias Disclosure	$F(1, 50) = 0.0$	1.0000		0.00	0.125
		+Bias Disclosure with Explanation	$F(1, 67) = 0.01$	0.9038		−0.01	0.121
	proxy	+Expl	$F(1, 228) = 18.55$	0.0000	*	0.69	0.161
		+Bias Disclosure	$F(1, 103) = 0.05$	0.8298		0.02	0.110
		+Bias and Corr Disclosure	$F(1, 103) = 6.39$	0.0130	*	0.29	0.113
		+Bias Disclosure with Explanation	$F(1, 156) = 0.2$	0.6559		0.05	0.110
		+Bias and Corr Disclosure with Explanation	$F(1, 156) = 1.22$	0.2716		−0.12	0.110

(*table continues*)

Metric	Bias Type	Effect	F	p		Coef	Std Error
safe	protected	+Expl	$F(1, 116) = 0.03$	0.8677		0.03	0.176
		+Bias Disclosure	$F(1, 50) = 1.69$	0.1997		−0.10	0.075
		+Bias Disclosure with Explanation	$F(1, 67) = 2.96$	0.0898		−0.18	0.103
	proxy	+Expl	$F(1, 228) = 4.17$	0.0422		0.24	0.119
		+Bias Disclosure	$F(1, 107) = 0.08$	0.7787		0.03	0.099
		+Bias and Corr Disclosure	$F(1, 108) = 0.01$	0.9419		−0.01	0.102
		+Bias Disclosure with Explanation	$F(1, 158) = 3.27$	0.0727		−0.17	0.092
		+Bias and Corr Disclosure with Explanation	$F(1, 158) = 7.76$	0.0060	*	−0.25	0.092
wary	protected	+Expl	$F(1, 116) = 0.12$	0.7283		0.06	0.183
		+Bias Disclosure	$F(1, 50) = 0.0$	1.0000		0.00	0.097
		+Bias Disclosure with Explanation	$F(1, 67) = 1.6$	0.2100		0.13	0.105
	proxy	+Expl	$F(1, 228) = 0.0$	0.9621		0.01	0.118
		+Bias Disclosure	$F(1, 113) = 4.48$	0.0365		0.24	0.114
		+Bias and Corr Disclosure	$F(1, 115) = 1.05$	0.3066		0.12	0.117
		+Bias Disclosure with Explanation	$F(1, 153) = 2.69$	0.1030		0.13	0.080
		+Bias and Corr Disclosure with Explanation	$F(1, 153) = 13.28$	0.0004	*	0.29	0.080
works	protected	+Expl	$F(1, 116) = 6.0$	0.0158	*	−0.38	0.154
		+Bias Disclosure	$F(1, 50) = 1.5$	0.2265		−0.14	0.112
		+Bias Disclosure with Explanation	$F(1, 67) = 2.56$	0.1145		−0.15	0.092
	proxy	+Expl	$F(1, 228) = 0.06$	0.8122		0.03	0.138
		+Bias Disclosure	$F(1, 105) = 4.43$	0.0376		−0.22	0.104
		+Bias and Corr Disclosure	$F(1, 106) = 14.58$	0.0002	*	−0.41	0.107
		+Bias Disclosure with Explanation	$F(1, 153) = 5.43$	0.0211		−0.20	0.088
		+Bias and Corr Disclosure with Explanation	$F(1, 153) = 14.19$	0.0002	*	−0.33	0.088

Table 6. Overall results of tests regarding the primary effects of our study on fairness perceptions, decision-making fairness, decision-making quality, and trust in our model.

C HUMAN STUDY INTERFACE

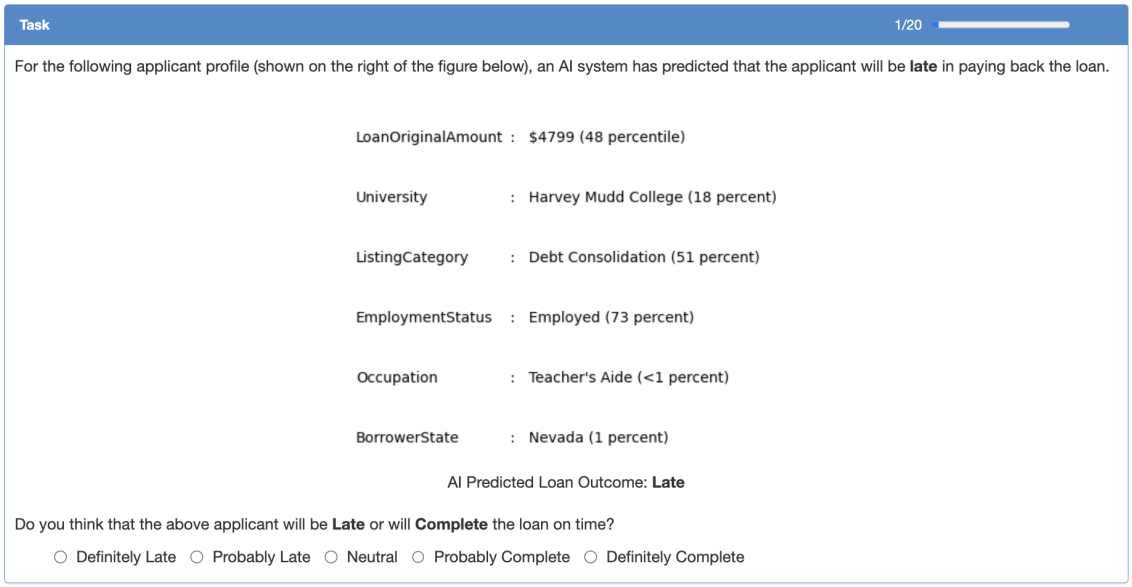

Fig. 8. Example task question without explanations.

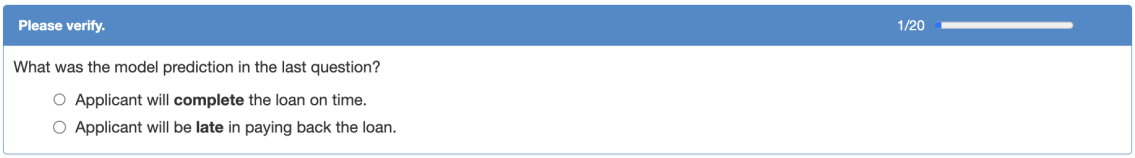

Fig. 9. Example attention check question.

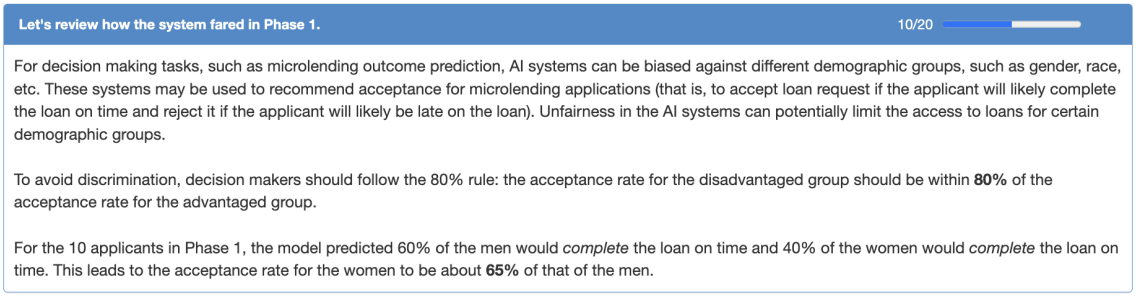

Fig. 10. Bias disclosure showing the demographic parity of the model in phase 1.

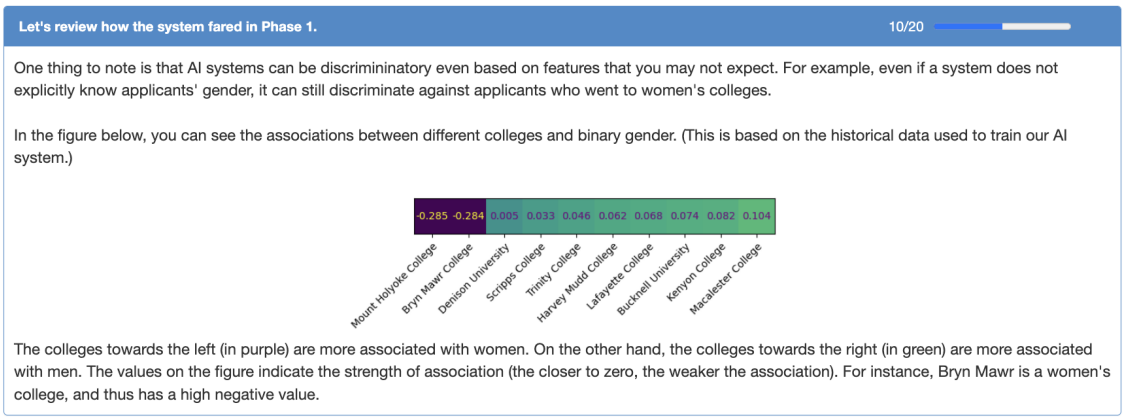

Fig. 11. Correlation disclosure showing the relationship between university and gender in our synthetic data.

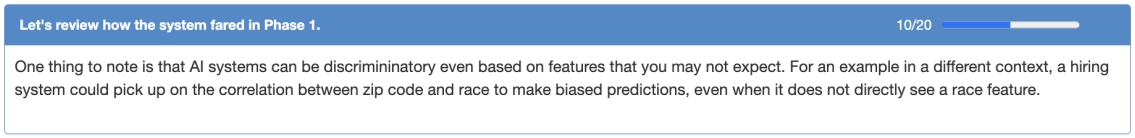

Fig. 12. In proxy conditions where participants are not given correlation disclosure. They are instead given this screen explaining that proxies can general, without mentioning the relationship in our data.

Fig. 13. Comprehension check screen testing both understanding of bias disclosure (Figure 10) and correlation disclosure (Figure 11). The correlation disclosure question is only shown in proxy conditions where the participants are given correlation disclosure.

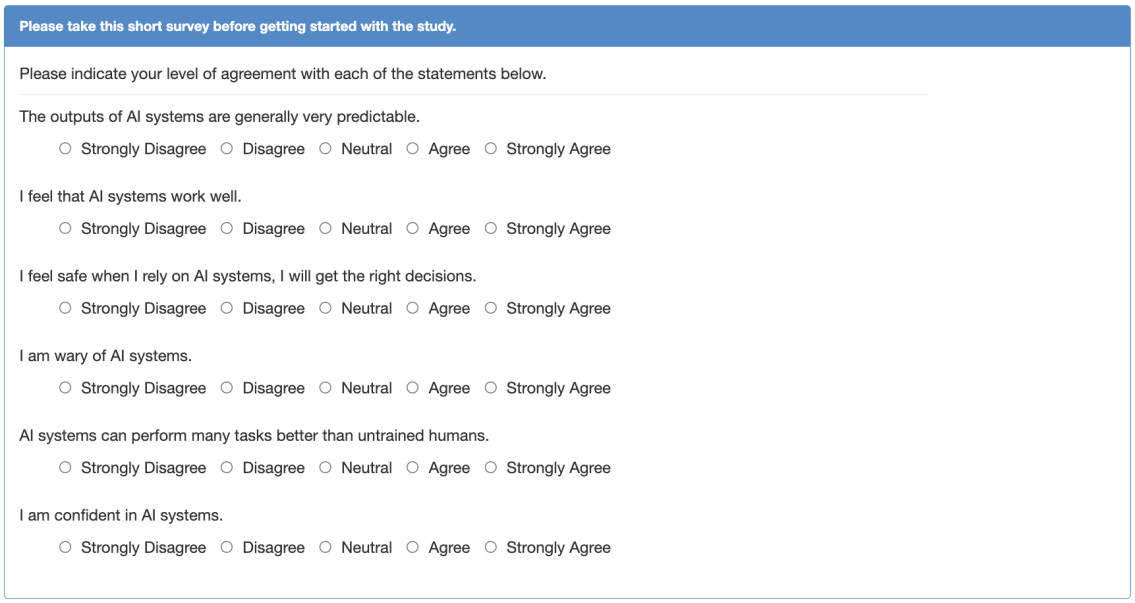

Fig. 14. Initial trust survey given before the task is introduced.

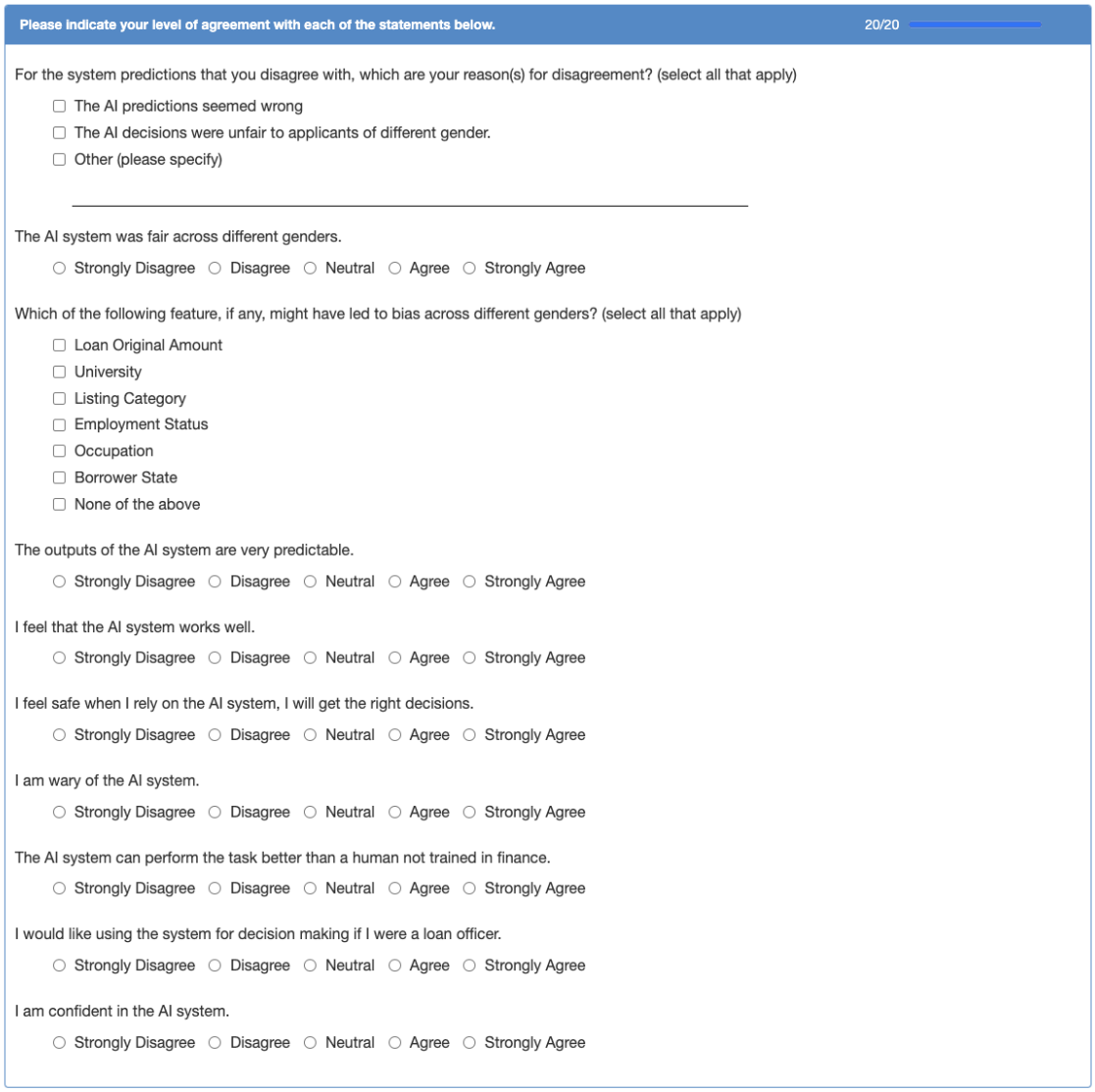

Fig. 15. Example post-task survey. This is the version that is shown after phase 2 of proxy conditions. In protected conditions and after phase 1 of proxy conditions, the question about which feature might have lead to gender bias is omitted.

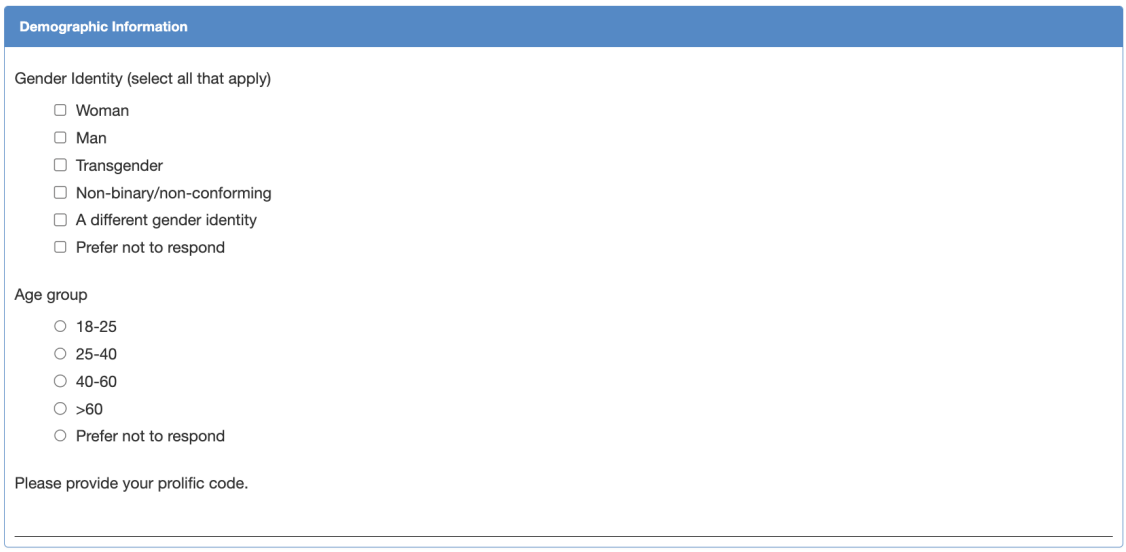

Fig. 16. Participant demographic questions.