# Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems

Meric Altug Gemalmaz
Purdue University
West Lafayette, Indiana, USA
mgemalma@purdue.edu

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

## ABSTRACT

The wide application of AI-based decision systems in many high-stake domains has raised concerns regarding fairness of these systems. As these systems will lead to real-life consequences to people who are subject to their decisions, understanding what these *decision subjects* perceive as a fair or unfair system is of vital importance. In this paper, we extend prior work in this direction by taking a perspective of *repeated interactions*—We ask that when decision subjects interact with an AI-based decision system repeatedly and can strategically respond to the system by determining whether to stay in the system, what factors will affect the decision subjects' fairness perceptions and retention in the system and how. To answer these questions, we conducted two randomized human-subject experiments in the context of an AI-based loan lending system. Our results suggest that in repeated interactions with the AI-based decision system, overall, decision subjects' fairness perceptions and retention in the system are significantly affected by whether the system is in favor of the group that subjects themselves belong to, rather than whether the system treats different groups in an unbiased way. However, decision subjects with different qualification levels have different reactions to the AI system's biased treatment across groups or the AI system's tendency to favor/disfavor their own group. Finally, we also find that while subjects' retention in the AI-based decision system is largely driven by their *own* prospects of receiving the favorable decision from the system, their fairness perceptions of the system is influenced by the system's treatment to people in *other* groups in a complex way.

## CCS CONCEPTS

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

## KEYWORDS

AI-based decision systems, fairness, fairness perceptions, retention, human-subject experiments, human-AI interaction

**ACM Reference Format:**
Meric Altug Gemalmaz and Ming Yin. 2022. Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534201



*AIES'22, August 1–3, 2022, Oxford, United Kingdom*

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534201

## 1 INTRODUCTION

With the rapid advance of the Artificial Intelligence (AI) technologies, a growing number of AI-based decision systems have been developed to automate the decision making in a wide range of high-stake domains such as loan lending [31], hiring [4], and immigration policing [18]. Unfortunately, it is found that many of these AI-based decision systems inherit the pre-existing biases in the dataset on which the systems get trained, and treat people coming from different socio-demographic groups in an unfair manner. For example, an AI model that reviews the credit card applications was found to exhibit gender bias as it gave a male applicant 20 times more credit limit compared to a female applicant with the same qualification [43]. In another case, an AI model widely used in US hospitals to allocate health care to patients was found to systematically discriminate against African Americans [23, 33].

The possibility of AI-based decision systems to behave unfairly has sparked great interests among researchers to investigate various methods for ensuring fairness in such systems. While earlier works tackle this challenge mostly by adjusting the training data, processes, and outputs of the AI systems [1, 20, 48, 49], more recently, an increasing number of studies start to take a more human-centered view by probing deeper into what does a "fair" AI-based decision system mean to *people*. For example, user interfaces have been designed to elicit diverse subjective fairness notions from different stakeholders [6]. Experimental studies have been conducted to understand people's preferences over multiple fairness definitions that potentially compete with one another [17, 37, 40]. Frameworks have also been proposed to learn context-aware fairness notions from humans' situated fairness judgements [46].

As AI-based decision systems bring about real-world consequences to people's lives, another important line of research regarding fairness of these systems is to examine what factors affect the fairness perceptions of *decision subjects* of an AI-based decision system (i.e., those people about whom the decisions are made by the system), and how. To this end, Wang et al. [44] show that in the one-shot interaction with an AI-based decision system, decision subjects perceive the system to be fairer both when the system is not biased against any specific group (i.e., the system is "fair" across groups), and when the system's decisions on them are in their favor. However, in practice, decision subjects often can repeatedly interact with an AI-based decision system and strategically respond to the system by, for example, actively deciding whether they want to

stay in the system or depart from it. In these scenarios of repeated interactions, how will a decision subject's *fairness perceptions* in an AI-based decision system be affected by various factors regarding the system's decision outcomes, such as the system's fairness level across groups and its tendency to favor the subject's group? And will these same factors also impact the subject's *retention* in the system? To complicate things further, different decision subjects have different characteristics, such as their qualification levels (i.e., they "deserve" a favorable decision to different degree) and sensitivity to fairness (i.e., they "value" fairness to different degree). What roles these individual characteristics play, both on their own and as potential moderating factors, in changing the subject's fairness perceptions and retention in the AI-based decision system?

In this paper, we made an initial attempt to answer these questions by conducting randomized human-subject experiments. In our experiments, we recruited human subjects to play a game in which they would play as a small business owner with randomly assigned group identity and qualification levels (i.e., credit score levels), and they needed to apply for loans from a bank to support their business for a period of at most 10 rounds. Subjects were told that the bank utilized an AI system to make its lending decisions. If the subject applied for a loan from the bank in one round, the bank's lending decisions on her as well as the summary information of the bank's decisions on all other applicants during the same period would be revealed to her. Importantly, after interacting with the AI-based banking system for at least once, the subject could decide to not apply loans from the bank anymore at any time and therefore depart from the system in any round as she wished.

Our first study involved two experimental treatments by varying decision outcomes of the bank's AI system across loan applicants of different groups. Through this study, we found that when a decision subject interacts with an AI-based decision system repeatedly, both her fairness perceptions and her retention in the system is significantly affected by *whether the system is in favor of her own group*, rather than whether the system treats people of different groups in an unbiased way. Decision subjects with higher qualification levels also had significantly higher retention in the system, but their fairness perceptions of the system did not change. More interestingly, we noted that the decision subject's qualification level moderates the impacts on her fairness perceptions and retention in the AI system that are brought up by the system's decision outcomes. As for the decision subject's sensitivity to fairness, we observed that subjects who value fairness more tended to perceive AI-based decision systems as more unfair and be less willing to participate in these systems, but we did not find any significant moderating effects associated with the subject's fairness sensitivity.

The findings of our first study led to a natural follow-up question—When decision subjects increase/decrease their fairness perceptions and retention in an AI system as it favors/disfavors the subject's group, are the changes driven by the subject's *own* prospects of receiving the favorable decision, or the subject's *relative* advantage/disadvantage over people in other groups in receiving the favorable decision? We conducted a second study to explore the answers to this question. Our results suggest that decision subjects' retention in the AI system is primarily driven by their own prospects of receiving the favorable decision. In contrast, the system's treatment to people in *other* groups did significantly contribute to subjects' fairness perceptions of the system, both via establishing a baseline for subjects to see the relative advantage/disadvantage of their own group, and perhaps surprisingly, via giving subjects a sense of the system's overall tendency to grant favorable decisions.

We conclude by discussing the implications of our study on understanding humans' repeated interactions with AI-based decision systems, and address the limitations of our work.

## 2 RELATED WORK

There is a growing body of work on understanding how humans adopt, interact with, and trust the AI-based decision-making systems [7, 9, 13, 26, 29, 34–36, 45, 47, 52]. Among many other factors, whether the AI system is "fair" is deemed as a critical factor that will affect people's perceptions of and reactions to the system. However, while there are numerous fairness definitions being proposed in the computer science literature [11, 12, 16, 22, 28, 30, 42], there is no clear agreement over a particular definition [37], and fairness requirements could be highly context-dependent [27]. Therefore, many empirical studies have been designed to solicit human preferences over different fairness definitions for a variety of decision-making contexts [17, 37, 40]. For example, for loan lending decisions, Saxena et al. [37] compared three fairness definitions and found that the calibrated fairness definition tends to be preferred by laypeople.

More recently, a more human-centered perspective has been taken to understand the fairness of AI-based decision systems. That is, instead of searching for a single, objective definition of fairness, fairness is increasingly being treated as a subjective concept, and various studies have been carried out to examine the range of factors that may affect people's fairness perceptions of an AI system. For instance, Hannan et al. [15] found that in resource allocation scenarios, people's fairness perceptions are influenced by what resource is being allocated, who allocates the resources, and sometimes even how the questions regarding fairness perceptions are asked. Other key influencing factors include whether and how the system's decisions are explained [3, 10, 38], the ways that the AI system's decisions are presented and visualized [41], and people's personal experience related to the algorithmic decision making scenario [14]. Researchers have also explored fairness perceptions of an AI-based decision systems from different stakeholders' points of view. For example, by *independently* controlling the AI system's decisions on individuals (i.e., favorable vs. unfavorable) and the system's treatment across groups (i.e., biased vs. unbiased), Wang et al. [44] showed that decision subjects' fairness perceptions of an AI system are predominately affected by whether the system makes a decision that is in their favor, although holding all else equal, decision subjects also perceive a system that exhibits unbiased treatment across different groups as fairer. On the other hand, from system developers' perspectives, factors used and processes involved in algorithmic decision making of an AI system are essential for them to judge the system's fairness level [21].

Compared to the prior work, in this paper, we aim to re-examine decision subjects' fairness perceptions and retention in AI-based decision systems as they interact with these systems *repeatedly*. This perspective of repeated/long-term interactions has been taken

in more theoretical examinations of fairness in AI, in which researchers often argue that an AI model that makes one-shot fair decisions by enforcing static fairness constraints may not lead to long-term well-being of those groups it aims to protect. This is because the decisions that an AI model makes on people, in the long run, can also reshape people, including changing the qualification distributions of different groups [24, 51], changing the group representation over time [50], affecting people's willingness to invest in their qualification [25], and causing different levels of precarity to people [32]. Our work complements this theoretical line of work from two perspectives: First, we extend the discussions from decision subjects' well-being in repeated interactions to their fairness perceptions in repeated interactions, and we suspect these perceptions may also be different from those perceptions in one-shot interactions. For example, the impact of an AI model being unfair across groups on decision subjects' fairness perceptions of the model may either be strengthened in the repeated interactions due to people's repeated exposure to the model's biased treatment, or be weakened as some people see the possibility to "exploit" such biased treatment to optimize their own utility. Second, as many theoretical studies use simulated models to capture the long-term user dynamics [8, 50], our investigation into decision subjects' retention in repeated interactions with an AI model could provide empirical evidence for characterizing the user dynamics more realistically.

## 3 STUDY 1

In our first study, to understand decision subjects' fairness perceptions and retention in an AI-based decision system as they repeatedly interact with the system, we conducted a randomized human-subject experiment. In particular, we ask:

- **RQ1**: How are decision subjects' fairness perceptions and their willingness to participate in the AI-based decision system affected by properties of the decision outcomes, such as the AI system's *fairness level across groups* (i.e., whether the AI system treats decision subjects of different groups equally), and the AI system's *tendency to favor the subject's own group*?
- **RQ2**: What role does a decision subject's qualification level play in influencing her fairness perceptions and retention in the AI system, both on its own and as a potential moderator of the impacts of the AI system's decision outcomes on the subject?
- **RQ3**: What role does a decision subject's sensitivity level to fairness play in influencing her fairness perceptions and retention in the AI system, both on its own and as a potential moderator of the impacts of the AI system's decision outcomes on the subject?

### 3.1 Experimental Design

*3.1.1 Experimental Tasks.* We recruited human subjects to play a game in our experiment, in which each subject was asked to play as a small business owner, and would interact with a bank repeatedly by applying loans from it to support her business. Subjects were told that this bank uses an AI model to make lending decisions, and they could decide whether to keep applying for loans from this bank of their own volition. The main interface of this game is shown in Figure 1. More specifically, upon arrival at the game, each subject was assigned with a loan applicant profile that represents her *throughout* this game (Figure 1A), which included 5 features:

- **Group**: the applicant's group identity, with two possible values—red or blue.
- **Credit score range**: a 30-point range of the applicant's credit score, which can be one of the 12 possible ranges in the set {480–510, ⋯, 630–660, 660–690, ⋯, 810–840}. Subjects were told that their precise credit score varies over time, but it typically falls into the range on their profile. They were also told that a credit score *above* 660 is generally considered to be "high," and the higher their credit scores are, the more they could hope to get their loans approved.
- **Credit history**: the number of years that the applicant has a credit history, which takes a value between 10 and 20.
- **Home ownership**: the ownership of the applicant's home, with two possible values—rent or own.
- **Small business industry**: The type of industry the applicant's small business belongs to, which can be one of the five values—software and IT services, advertising and marketing, food and accommodation service, healthcare service, and construction.

The subject's loan applicant profile was created by *uniformly* randomly sampling a value from the set of possible values for each of the 5 features. Note that for the applicant's group identity, we chose to not bind it with a particular definition of socio-demographic groups (e.g., gender, race) to avoid the possible noisy data resulted from a mismatch between a subject's group identity in the real world and in the game. In addition, the credit score range of a subject was used to reflect the subject's "*qualification level*," i.e., to what extent the subject deserves a favorable decision (i.e., getting the loan)—the higher the credit score of a subject, the more "qualified" she was for receiving a loan. Finally, the last 3 features were added into the subject's profile to make the profile more realistic.

Beyond the loan applicant profile, the subject was also given an "account" with an initial balance of 600 "coins." The subject then needed to interact with the bank for *at most* 10 rounds. In each round, the subject was asked to decide whether she'd like to continue to apply for a loan from the bank (Figure 1B). If yes, 50 coins would be deducted from her account as the application fee. The bank's lending decision, which was decided by the AI model, would then be revealed to her (Figure 1C)—if the AI model approved her loan application, the subject would receive a reward of 100 coins; otherwise the subject would receive nothing. The subject would also be able to view the summary information of the AI model's decisions on *all* applicants in this round—broken down by the applicant's group—before moving on to the next round (Figure 1D; see Section 3.1.2 for details). However, in one round, if the subject decided not to continue to apply for a loan from the bank, she would immediately leave the game and be re-directed to the end of the experiment.

Overall, this game was designed to closely reflect the real-world scenario that decision subjects can freely decide whether they are willing to "stay in the system" to take part in AI-based decision making (i.e., whether they want to be subject to a particular AI system's decisions) as they repeatedly interact with the system. These participation decisions are often made as the decision subjects—with some knowledge of their own qualification levels—observe the AI system's decisions on themselves and on others over time. Moreover, while the decision to participate is often costly (i.e., the

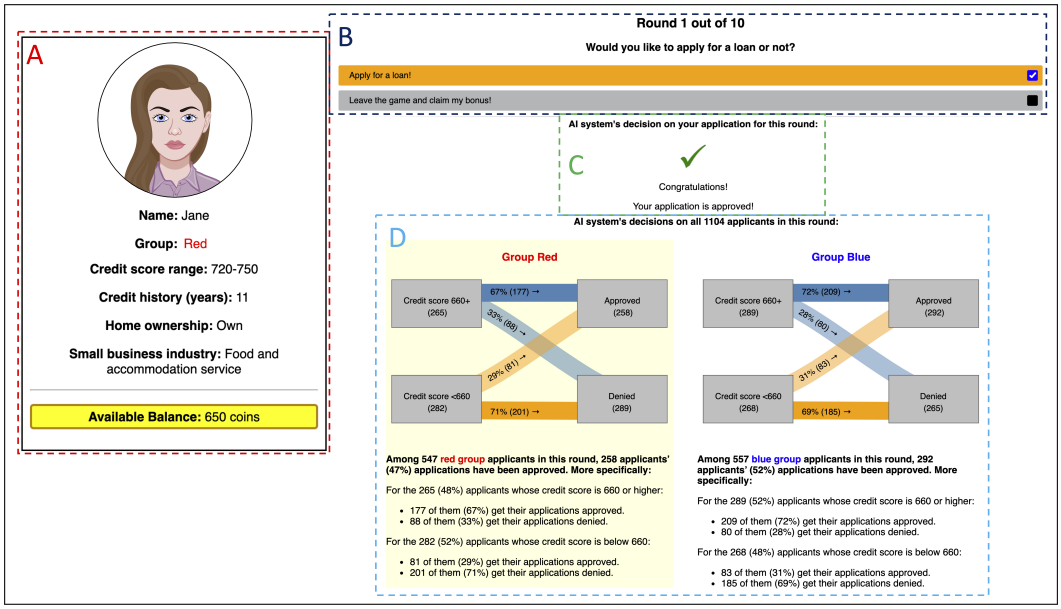



**Figure 1: The main interface of the game. The loan applicant profile assigned to the subject is presented on the interface (Part A). In each round, the subject needed to decide whether to continue to apply loans from the bank (Part B). If the subject decided to apply for a loan in one round, the AI model's lending decision on the subject would be revealed to her (Part C), and the subject could also get a summary of the AI model's decisions on all applicants in this round (Part D).**

participate decision in the game is associated with a "fee" of 50 coins), decision subjects could benefit from such participation when the AI system grants a favorable decision to them (i.e., a favorable decision is associated with a "reward" of 100 coins in the game). Note that in this game, we assume the AI model is not updated over time, and the model's decisions do not result in changes in the decision subjects' qualification levels that are significant enough to affect the model's future decisions on them. Understanding decision subjects' fairness perceptions and retention in repeated interactions with the AI-based decision systems after relaxing these assumptions will be an interesting future work.

*3.1.2 Experimental Treatments.* We created two treatments by varying properties of the bank's AI model:

- **"Fair" model**: In this treatment, the bank's AI model treats applicants from different groups *equally*. Specifically, this AI model makes a stochastic lending decision based on a "decision matrix," as shown in Table 1a. According to this matrix, *regardless of the group identity of the applicant*, the chance for this AI model to approve the loan for an applicant with a high credit score (i.e., score $\geq 660$) is 70%, while the chance to approve the loan for an applicant with a low credit score (i.e., score $< 660$) is 30%.
- **"Unfair" model**: In this treatment, the bank's AI model is unfair as it is systematically *in favor of applicants from the red group*. Specifically, for applicants of the red group, the AI model makes its stochastic lending decision based on the decision matrix as shown in Table 1b—the probabilities for the AI model to approve the loan for a red group applicant with a high or low credit score are 90% or 40%, respectively. In contrast, for applicants of the blue group, the AI model makes its stochastic lending decision based on the matrix as shown in Table 1c—the probabilities for the AI model to approve the loan for a blue group applicant with a high or low credit score are 50% or 20%, respectively.

Specifically, in one round of the game, if the subject decided to apply for a loan from the bank, the bank's lending decision on her would be made by the AI model of the subject's *assigned treatment*—Given the decision matrices of the AI model, the subject's group identity and credit score range would be used to determine the probability of loan approval, and then the AI model would randomly realize its lending decision on the subject according to this probability. Moreover, to allow the subject to get a sense of the AI model's overall decisions on applicants of different groups, we told the subject that many other people had also applied loans from the bank in the same time period, and we showed the summary information of the AI model's decisions on *all* these applicants to the subject[1]. In particular, in each round, we simulated another $N$ loan applicants where $N$ is an integer uniformly randomly drawn from the interval of $[1000, 1200]$. Again, for each of these $N$ applicants, we randomly generated her profile (i.e., each feature value was uniformly randomly sampled from the set of possible values) and determined the lending decision for her using the AI model of the subject's *assigned treatment*. Finally, we displayed the AI model's decisions on all these $N$ applicants to the subject through flowcharts (Figure 1D)[2], with decisions on red group applicants and blue group applicants shown in separate flowcharts. We also provided textual explanations along with the flowcharts to help subjects better interpret information in the flowcharts.

[1] In reality, decision subjects may get access to such summary information of the AI model's decisions on decision subjects of different groups due to media coverage or scientific investigation of the AI model, such as [2, 5].
[2] We chose to use flowcharts since previous study suggested that flowcharts could best support laypeople's understanding of the performance of algorithmic models [39].

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 70% | 30% |
| < 660 | 30% | 70% |

**(a) Fair model: Red/Blue group**

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 90% | 10% |
| < 660 | 40% | 60% |

**(b) Unfair model: Red group**

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 50% | 50% |
| < 660 | 20% | 80% |

**(c) Unfair model: Blue group**

**Table 1: The decision matrices used by the AI model in different treatments of Study 1 on loan applicants of different groups. Number in each cell represents the probability for the AI model to approve/deny an applicant when the applicant's credit score falls into the range as specified in the corresponding row.**

We note that if we consider each loan applicant's qualification level (i.e., their credit score range) as the "ground truth" for the lending decision, the decision matrices in Table 1 are effectively the AI models' *confusion matrices*. Since each loan applicant's profile was generated uniformly randomly, when considering the AI model's *overall* performance regardless of the group identity of the decision subjects, the fair AI model and the unfair AI model had exactly the *same* expected performance with respect to a range of metrics such as accuracy, positive prediction rate (PPR), false positive rate (FPR), and false negative rate (FNR)[3]. However, while the fair AI model treats decision subjects of different groups equally, the unfair model is in favor of decision subjects from the red group according to all these metrics—it had a higher accuracy, a higher PPR, a higher FPR, and a lower FNR, on red group applicants.

*3.1.3 Experimental Procedure.* Our experiment was posted as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (MTurk)[4]. This HIT was open to U.S. workers only, and each worker was allowed to take the HIT only once.

Upon arrival at the HIT, the subject was asked to create a nickname and select an avatar to represent herself in the game. She would then be presented with the instruction of the game. In particular, we used an interactive tutorial to explain to her the meaning of all the information shown in her assigned, randomly generated, loan applicant profile, the interface of the game (e.g., how to read the flowcharts), as well as the rules of the game. At the end of the instruction, we prepared 4 questions to test the subject's understanding of the game. The subject was only qualified to proceed to the actual game after correctly answering all these questions.

Once qualified, the subject would be *randomly* assigned to one of the two experimental treatments and start to play the game. As explained in Section 3.1.1, in each round of the game, the subject decided whether to continue to apply for a loan from the bank. If yes, the bank's lending decision on her, as well as on all applicants in this round, would then be revealed to the subject[5]. The subject's account balance would also be updated based on the lending decision she received. To make sure that the subject at least had some interaction with the AI system, we required each subject to apply for a loan in the *first* round. After that, the subject could continue to apply for loans for a maximum of 9 more rounds, but she could also decide not to apply for loans anymore in any round, which would immediately redirect the subject to the end of the game.

At the end of the game, the subject was asked to answer a few exit-survey questions. In particular, the subject first reported some demographic information (e.g., gender, age). Then, the subject was asked to indicate how fair she perceived the bank's AI system was—We adapted a set of six fairness perception statements from those used in [44] (e.g., "The bank's AI system is fair to manage loan applications."), and the subject evaluated how much she agreed with each statement on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The subject could also comment on her perceived fairness of the AI system via free text. To quantify the subject's sensitivity to fairness (i.e., to what extent the subject values fairness), we created another set of 4 statements as below:

- It is very important to me that an AI system making decisions about people is fair (i.e., it treats everyone fairly and does not discriminate).
- I would only use an AI system if it is fair to everyone.
- I would stop using an AI system if it is unfair, even if it tends to be in favor of me.
- When I decide whether to use an AI system or not, I seldom think about whether the system is fair. (Negative)[6]

Finally, we conjectured that a decision subject's fairness perceptions and retention in an AI-based decision system might be influenced by the subject's *risk attitude* (i.e., how much the subject is willing to take risks). We thus included another set of statements created in previous studies [19] to measure the subject's risk attitude (e.g., "I prefer friends who are exciting and unpredictable."). Again, for statements related to both fairness sensitivity and risk attitude, the subject rated how much she agreed with each statement on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

After completing the exit-survey, we would reveal to the subject the amount of bonus payment she received in this game—We converted the amount left in the subject's account to her bonus payment using a rate of 500 coins to $1.5. Thus, together with the base payment of $1.5 of this HIT, the subject could earn a maximum of $4.8 from this game[7].

## 3.2 Analysis Methods

We adopted two main dependent variables in our analysis: (1) the decision subject's *perceived fairness level* of the AI system, which was the sum of the subject's ratings on those statements in the exit-survey regarding her fairness perceptions of the AI system—the higher the total rating, the fairer the subject found the AI system to be; (2) the decision subject's *retention* in the AI system, which was quantified through the number of rounds that the subject decided

[3] We considered the decision of approving the loan as the positive decision. On expectation, both the fair model and the unfair model had an accuracy of 70%, a positive prediction rate of 50%, a FPR of 30%, and a FNR of 30%, across all decision subjects.
[4] All of our experiments were approved by the IRB of the authors' institution.
[5] On the interface, we used a light yellow background to highlight the AI model's decisions on applicants coming from the subject's *own* group to allow subjects better contrast the model's performance on different groups.
[6] We reversed the rating for negative statements.
[7] The median value of time that subjects spent on our HIT was 20 minutes, and the median payment to subjects was $3.3, leading to an effective hourly wage of $9.9.

to apply for a loan from the bank—the larger the number, the more the decision subject was willing to stay in the AI system.

We then fit our experimental data into regression models to answer our research questions. More specifically, for **RQ1**, we first defined a binary variable "*biased treatment*" to indicate whether the AI system treats decision subjects of different groups in a biased way (i.e., the fair model: 0; the unfair model: 1). Using this variable as the independent variable, we constructed linear regression models to analyze how the AI system's fairness level across groups affects decision subjects' fairness perceptions and retention in the AI system, while the subject's risk attitude was included in the regression models as a covariate[8].

Similarly, to see how decision subjects' fairness perceptions and retention in the AI system are affected by the system's tendency to favor the group that the subject belongs to, we created two other binary variables—"*advantaged*" and "*disadvantaged*," which reflects whether the AI system placed the subject's group at an advantaged or disadvantaged position, respectively, *compared to the other group*, with respect to receiving the favorable decisions (i.e., fair model: advantaged=disadvantaged=0; unfair model, red group: advantaged=1, disadvantaged=0; unfair model, blue group: advantaged=0, disadvantaged=1). Again, regression models were built using these two variables as the independent variables while controlling for the subject's risk attitude.

Next, to examine the role that a decision subject's qualification level plays in influencing the subject's fairness perceptions and retention in the AI-based decision system (**RQ2**), we mapped each subject's credit score range into a value between 0 and 11 (higher credit score ranges were mapped into larger values). We then incorporated this credit score level into the set of regression models that we previously had for **RQ1**—For each regression model, we first included only the subject's credit score level as a covariate. Then, we further included the interaction term(s) between the subject's credit score level and the independent variable(s), which allowed us to understand whether the subject's qualification level moderates the impacts of the AI system's decision outcomes on the subject.

Finally, for **RQ3**, we computed each subject's fairness sensitivity score based on her responses on the relevant statements in the exit-survey—the higher the score, the more the subject values fairness. Again, we constructed a new set of regression models on the basis of what we previously had for **RQ1** by adding the fairness sensitivity score as well as its interaction(s) with the independent variable(s) into them subsequently. However, since subjects' fairness sensitivity scores were found to be highly correlated with their risk attitude, we removed the subject's risk attitude from this set of regression models to avoid the multicollinearity problems.

## 3.3 Results

In total, 809 subjects participated in our experiment. In the following, we analyzed the data that we collected from these subjects to answer our research questions.

*3.3.1 RQ1: The impacts of the AI system's decision outcomes.* We start by examining whether the AI system's fairness level across groups influences decision subjects' fairness perceptions and retention. Results are shown in Table 2 (Models 1 and 3). We find that whether the AI system treats decision subjects of different groups equally does *not* significantly affect either decision subjects' perceived levels of fairness of the AI system (Model 1) or their willingness to participate in the system (Model 3). Indeed, as shown in Figure 2a, regardless of whether the AI system treats decision subjects of different groups in a biased way or not, overall, subjects in the two treatments departed from the AI system at a similar rate.

| | **Perceived Fairness** | | **Retention** | |
|---|---|---|---|---|
| | *Model 1* | *Model 2* | *Model 3* | *Model 4* |
| Biased treatment | -0.12<br>(0.28) | | -0.12<br>(0.26) | |
| Advantaged | | 0.57<br>(0.35) | | $0.58^{\dagger}$<br>(0.32) |
| Disadvantaged | | $-0.68^{*}$<br>(0.33) | | $-0.69^{*}$<br>(0.30) |
| Risk attitude | $0.33^{***}$<br>(0.03) | $0.34^{***}$<br>(0.03) | $0.18^{***}$<br>(0.03) | $0.18^{***}$<br>(0.03) |
| Constant | $9.15^{***}$<br>(0.50) | $9.09^{***}$<br>(0.51) | $4.03^{***}$<br>(0.46) | $3.97^{***}$<br>(0.46) |

**Table 2: Regression models predicting decision subjects' perceived levels of fairness and retention in the AI system, based on properties of the AI system's decision outcomes. Coefficients and standard errors are reported. $^{\dagger}$, $^{*}$, and $^{***}$ represent significance levels of 0.1, 0.05, and 0.001, respectively.**

In contrast, whether the AI system is in favor of the group that the decision subject belongs to affects both the subject's fairness perception and retention (Models 2 and 4). When a decision subject's group is favored by the AI system in receiving the preferable decision, the decision subject seems to perceive the model as fairer (Model 2, though not significant), and has a marginally higher level of willingness to stay in the AI-based decision system (Model 4, $p = 0.075$)—As shown in Figure 2b, subjects who were assigned to the treatment with the unfair AI model and the red group (i.e., the group being favored by the unfair model) tended to apply for loans from the bank for more rounds. However, when a decision subject's group is disfavored by the AI system, the decision subject significantly decreases her perceived fairness level of the system (Model 2, $p = 0.039$), as well as her retention in the system (Model 4, $p = 0.023$; also see the blue solid line in Figure 2b). Notably, we also find that decision subjects' risk attitude is significantly correlated with their fairness perceptions and retention in the AI system—the more risk-seeking the decision subject is, the fairer she perceives the AI system to be and the more she is willing to stay in the system.

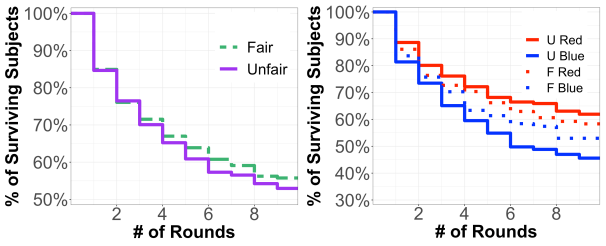
*3.3.2 RQ2: The role of decision subjects' qualification levels.* To first see how a decision subject's qualification level, *by itself*, correlates with her fairness perceptions and retention in an AI-based decision system, we simply add the qualification level as a covariate into each of the four regression models that we have constructed for **RQ1**. Results are reported as Models 1, 3, 5, 7 in Table 3, which consistently indicate that a decision subject with a higher qualification level is significantly *more* likely to participate in the AI-based decision system ($p < 0.001$ for both Models 5 and 7), but her perceived fairness level of the AI system is not significantly different.

Next, we explore whether a decision subject's qualification level *moderates* the impacts of the AI system's decision outcomes on the

[8]The subject's risk attitude was computed by summing up her ratings on the relevant statements in the exit-survey; higher total ratings imply more risk-seeking subjects.

| | Perceived Fairness | | | | Retention | | | |
|---|---|---|---|---|---|---|---|---|
| | *Model 1* | *Model 2* | *Model 3* | *Model 4* | *Model 5* | *Model 6* | *Model 7* | *Model 8* |
| Biased treatment | -0.12 (0.28) | 0.86$^{\dagger}$ (0.52) | | | -0.14 (0.25) | 0.37 (0.48) | | |
| Advantaged | | | 0.57 (0.36) | 1.98$^{**}$ (0.67) | | | 0.58$^{\dagger}$ (0.32) | 0.64 (0.60) |
| Disadvantaged | | | -0.68$^{*}$ (0.33) | -0.08 (0.62) | | | -0.73$^{*}$ (0.30) | 0.05 (0.56) |
| Qualification | 0.00 (0.04) | 0.09 (0.06) | 0.01 (0.04) | 0.09 (0.06) | 0.19$^{***}$ (0.04) | 0.23$^{***}$ (0.05) | 0.19$^{***}$ (0.04) | 0.23$^{***}$ (0.05) |
| Qualification × Biased treatment | | -0.18$^{*}$ (0.08) | | | | -0.09 (0.07) | | |
| Qualification × Advantaged | | | | -0.26$^{*}$ (0.11) | | | | -0.01 (0.10) |
| Qualification × Disadvantaged | | | | -0.11 (0.09) | | | | -0.14$^{\dagger}$ (0.09) |
| Risk attitude | 0.34$^{***}$ (0.03) | 0.34$^{***}$ (0.03) | 0.32$^{***}$ (0.03) | 0.35$^{***}$ (0.03) | 0.18$^{***}$ (0.03) | 0.18$^{***}$ (0.03) | 0.18$^{***}$ (0.03) | 0.18$^{***}$ (0.03) |
| Constant | 9.13$^{***}$ (0.55) | 8.64$^{***}$ (0.59) | 9.06$^{***}$ (0.54) | 8.54$^{***}$ (0.59) | 3.04$^{***}$ (0.49) | 2.78$^{***}$ (0.54) | 2.96$^{***}$ (0.49) | 2.76$^{***}$ (0.53) |

**Table 3: Regression models predicting decision subjects' perceived levels of fairness and retention in the AI system, based on properties of the AI system's decision outcomes and subjects' qualification levels. Coefficients and standard errors are reported. $^{\dagger}$, $^{*}$, $^{**}$, and $^{***}$ represent significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.**




**(a) Aggregate across groups** **(b) Broken down by groups**

**Figure 2: Survival curves showing the fraction of subjects who continued to apply for a loan from the bank after the X-th round in the two experimental treatments. In Figure 2b, "U" ("F") represents the treatment with unfair (fair) model.**

subject's fairness perceptions and retention. We do so by including the interaction term(s) between the subject's qualification level and the independent variable(s) representing properties of the AI system's decision outcomes into the regression models. Results are reported as Models 2, 4, 6, 8 in Table 3. For example, consider the impacts of the AI system's fairness level across groups on subjects—In Model 2, we find that while decision subjects with lower qualification levels perceive the unfair AI system to be marginally fairer than the fair AI system ($p = 0.099$ for "Biased treatment"), decision subjects with higher qualification levels significantly decrease their perceived fairness level of the AI system when the AI system is unfair across groups (Model 2, $p = 0.027$ for the interaction). However, such decrease does not result in any significant different change in highly-qualified subjects' retention in the AI system as compared to low-qualified subjects (Model 6). When it comes to the impacts of the AI system's tendency to favor a subject's own group on the subject, we detect a significantly negative interaction (Model 4, $p = 0.014$) between the qualification level and the independent variable of "advantaged" when examining subjects' fairness perceptions of the AI system. This means that highly-qualified subjects increase their perceived fairness level of the AI system to a *smaller* degree compared to low-qualified subjects when the AI system favors the group that they belong to. Finally, we detect a marginally negative interaction between the qualification level and independent variable of "disadvantaged" in influencing subjects' retention in the AI system (Model 8, $p = 0.099$), suggesting that highly-qualified subjects decrease their retention in the system to a slightly *larger* extent compared to low-qualified subjects when their own group is placed at the disadvantaged position by the system.

| | Perceived Fairness | | Retention | |
|---|---|---|---|---|
| | *Model 1* | *Model 2* | *Model 3* | *Model 4* |
| Biased treatment | -0.16 (0.30) | | -0.14 (0.26) | |
| Advantaged | | 0.43 (0.38) | | 0.51 (0.33) |
| Disadvantaged | | -0.65$^{\dagger}$ (0.35) | | -0.66$^{*}$ (0.31) |
| Fairness sensitivity | -0.18$^{*}$ (0.07) | -0.18$^{*}$ (0.07) | -0.13$^{*}$ (0.06) | -0.13$^{*}$ (0.06) |
| Constant | 16.50$^{***}$ (0.75) | 16.52$^{***}$ (0.75) | 8.25$^{***}$ (0.66) | 8.28$^{***}$ (0.65) |

**Table 4: Regression models predicting decision subjects' perceived levels of fairness and retention in the AI system, based on properties of the AI system's decision outcomes and decision subjects' sensitivity to fairness. Coefficients and standard errors are reported. $^{\dagger}$, $^{*}$, and $^{***}$ represent significance levels of 0.1, 0.05, and 0.001, respectively.**

*3.3.3 RQ3: The role of decision subjects' sensitivity to fairness.* Finally, we examine the role that a decision subject's sensitivity to fairness plays in influencing her fairness perceptions and retention in the AI-based decision system. Similar as our analyses in Section 3.3.2, we first include the subject's fairness sensitivity score as a covariate into our regression models, and results are shown in Table 4. In all these models, we consistently find that the decision subject's sensitivity to fairness is significantly negatively correlated with the subject's fairness perceptions and retention in the AI system ($p < 0.05$). In other words, the more the decision subject values fairness, the more unfair she perceives the AI system to be, and the less she is willing to participate in the system. We next add the interaction term(s) between the subject's fairness sensitivity score and the independent variable(s) into each of the regression models, but we do not detect any significant interactions in all these models, suggesting that subjects' sensitivity to fairness do not seem to moderate the impacts of the AI system's decision outcomes on subjects' fairness perceptions and retention.

## 4 STUDY 2

In Study 1, we found that, overall, as decision subjects interact with an AI-based decision system repeatedly, their fairness perceptions and retention in the system are mainly influenced by the system's tendency to favor the subject's own group, rather than the system's fairness level across groups. However, it is still unclear what the cause underlying such behavior is:

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 70% | 30% |
| < 660 | 30% | 70% |

(a) Red group (all) & Blue group ("Unbiased")

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 50% | 50% |
| < 660 | 20% | 80% |

(b) Blue group ("Red advantaged")

| Credit/Decision | Approve | Deny |
|---|---|---|
| ≥ 660 | 90% | 10% |
| < 660 | 40% | 60% |

(c) Blue group ("Red disadvantaged")

**Table 5: The decision matrices used by the AI model in different treatments of Study 2 on loan applicants of different groups. Number in each cell represents the probability for the AI model to approve/deny an applicant when the applicant's credit score falls into the range as specified in the corresponding row.**

- **RQ4**: Are the changes in subjects' fairness perceptions and retention in the AI system when the system favors (disfavors) their own group caused by subjects' *own* prospects of receiving the favorable decision, or the *relative* advantage (disadvantage) that the AI system grants to the subject's group over other groups?

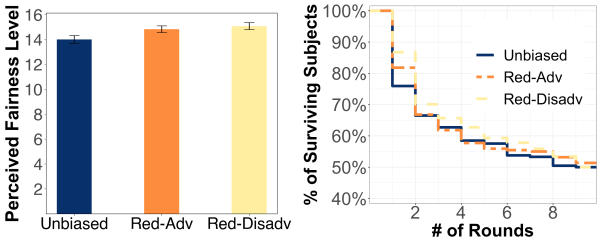
Study 1 does not provide a direct answer to this question, since in Study 1, whenever the AI system is in favor of a subject's group, the AI system places the subject's group at the advantaged position *by* providing a higher prospect of the favorable decision to the subject' group. Therefore, to answer **RQ4**, we conducted a second randomized human subject experiment.

## 4.1 Experimental Design

We again recruited human subjects to play the same game of loan application as that in Study 1, with only one key difference—In this experiment, *all subjects were assigned to the red group*. We then created three experimental treatments by varying how the AI system treats the red group applicants in relative to blue group applicants, while *controlling the prospects of the favorable decision for red group applicants*. Specifically, in all three treatments, the AI system makes its stochastic lending decisions to applicants of the red group based on the same decision matrix as shown in Table 5a (i.e., approve the loan for a red group applicant with a high or low credit score with a probability of 70% or 30%, respectively). However, the AI system makes lending decisions to the blue group applicants in different treatments based on different decision matrices:

- **Unbiased**: In this treatment, the bank's AI system makes lending decisions on blue group applicants based on the same decision matrix as that for the red group (i.e., Table 5a). Thus, the bank places applicants in neither group at the advantaged position.
- **Red advantaged**: In this treatment, the bank's AI system uses the decision matrix as shown in Table 5b to make its lending decisions on blue group applicants (i.e., approve the loan for a blue group applicant with a high or low credit score with a probability of 50% or 20%, respectively). Thus, the bank places applicants from the red group at the advantaged position.
- **Red disadvantaged**: In this treatment, the bank's AI system uses the decision matrix as shown in Table 5c to make its lending decisions on blue group applicants (i.e., approve the loan for a blue group applicant with a high or low credit score with a probability of 90% or 40%, respectively). Thus, the bank systematically discriminates against applicants from the red group.

This design may allow us to determine that when the AI system is in favor of a subject's group, whether the subject increases her fairness perceptions and retention in the system simply due to the higher prospect of receiving the favorable decision from the system—If yes, we expect to see minimal differences across the three treatments on subjects' fairness perceptions and retention; otherwise, we expect to see some differences across treatments.




(a) Perceived Fairness (b) Retention

**Figure 3: The fairness perceptions and retention for subjects in Study 2. (3a): Subject's average perceived levels of fairness of the AI system; error bars represent the standard errors of the mean. (3b): Survival curves showing the fraction of subjects who continued to apply for a loan from the bank after the X-th round in the three treatments of Study 2.**

Again, we posted this experiment as a HIT on MTurk to U.S. workers only, where it had an identical procedure as the experiment in Study 1 (see Section 3.1.3) except for the following differences: (1) Workers who had participated in the experiment in Study 1 were *not* allowed to participate in this experiment; (2) workers were randomly assigned to one of the three treatments as defined above.

## 4.2 Experimental Results

In total, we collected data from 636 subjects for Study 2. We adopted the same dependent variables as those used in Study 1 in the analyses, while the main independent variables we used were "advantaged" and "disadvantaged," indicating whether the subject's group was placed at the advantaged or disadvantaged position in receiving the favorable decision as compared to the other group (i.e., "Unbiased": advantaged=disadvantaged=0; "Red advantaged": advantaged=1, disadvantaged=0; "Red disadvantaged": advantaged=0, disadvantaged=1). To answer **RQ4**, we visualize our experimental data, and then fit them into regression models to see whether the AI system's tendency to favor/disfavor a subject's group still has any impact on the subject's fairness perceptions and retention, after fixing the subject's prospect of receiving the favorable decision.

Figure 3a compares subjects' perceived level of fairness of the AI system across the three treatments, and Figure 3b shows the subjects' survival curves in the three treatments. In Figure 3b, we observe minimal differences across the three treatments regarding subjects' willingness to stay in the AI system. This seems to suggest that changes in subjects' retention in an AI system is mainly driven by subjects' own prospects of receiving the favorable decision, rather than the relative advantage or disadvantage for subjects' group to receive the favorable decision over the other group.

| | Perceived Fairness | | | Retention | | |
|---|---|---|---|---|---|---|
| | *Model 1* | *Model 2* | *Model 3* | *Model 4* | *Model 5* | *Model 6* |
| Advantaged | 0.92$^{**}$ (0.30) | 0.87$^{**}$ (0.30) | 0.89$^{**}$ (0.29) | 0.04 (0.29) | 0.08 (0.28) | 0.00 (0.28) |
| Disadvantaged | -0.30 (0.28) | -0.26 (0.28) | -0.25 (0.27) | -0.01 (0.27) | -0.04 (0.27) | 0.02 (0.26) |
| Observed favorable decision prob. | 0.81 (0.51) | 0.85$^{\dagger}$ (0.51) | 0.84$^{\dagger}$ (0.50) | 3.74$^{***}$ (0.48) | 3.72$^{***}$ (0.48) | 3.78$^{***}$ (0.48) |
| Observed PPR | 10.49$^{**}$ (3.62) | | | 0.36 (3.45) | | |
| Observed TPR | | 7.06$^{**}$ (2.67) | | | 0.83 (2.55) | |
| Observed FPR | | | 14.72$^{**}$ (4.91) | | | -0.89 (4.69) |
| Risk attitude | 0.37$^{***}$ (0.02) | 0.37$^{***}$ (0.02) | 0.37$^{***}$ (0.02) | 0.16$^{***}$ (0.02) | 0.16$^{***}$ (0.02) | 0.16$^{***}$ (0.02) |
| Study 2 | -0.12 (0.21) | -0.11 (0.21) | -0.12 (0.21) | -0.48$^{*}$ (0.20) | -0.48$^{*}$ (0.20) | -0.48$^{*}$ (0.20) |
| Constant | 2.76 (1.81) | 3.05 (1.87) | 3.57$^{*}$ (1.49) | 2.30 (1.73) | 1.91 (1.79) | 2.74$^{\dagger}$ (1.43) |

**Table 6: Regression models predicting decision subjects' perceived levels of fairness and retention in the AI system, based on properties of the AI model's decision outcomes, after combining data from Studies 1 and 2. Coefficients and standard errors are reported. $^{\dagger}$, $^{*}$, $^{**}$, and $^{***}$ represent significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.**

However, in Figure 3a, we notice a surprising trend in subjects' perceived fairness perceptions of the AI system—According to our regression results, compared to subjects in the "Unbiased" treatment, subjects in both the "Red advantaged" treatment and the "Red disadvantaged" treatment increased their perceived level of fairness of the AI system, either marginally or significantly (e.g., red advantaged vs. unbiased: $p = 0.066$, red disadvantaged vs. unbiased: $p = 0.011$). This clearly indicates that subjects' fairness perceptions of the AI system are *not* solely determined by their own prospects of receiving the favorable decision. Moreover, the fact that subjects reported the highest perceived level of fairness to the AI system which actually places them at a disadvantaged position in receiving the favorable decision also seems to suggest that there are more factors beyond the relative advantage/disadvantage over the other group that substantially influence subjects' fairness perceptions.

To explore what these additional factors could be, we look into the free-text comments that subjects in the "Red disadvantaged" treatment left in the exit-survey explaining why they felt the AI system was fair. Interestingly, we find that some subjects related their perceptions of the AI system's fairness to the system's overall tendency of granting the favorable decision, both among highly-qualified applicants and low-qualified applicants. For example:

- "Majority of those with high credit score were getting their loans approved which is fair."
- "I thought this system was fair since I think it was based on credit scores, and it was still possible for people with low credit scores to be approved."

These comments suggest that subjects' perceived fairness level of an AI system *may* also be influenced by the system's *overall* positive prediction rate (PPR), true positive rate (TPR), and/or false positive rate (FPR), regardless of the applicant's group identity. To see how the three factors—a subject's own prospect of the favorable decision, the relative advantage or disadvantage of the subject's group over the other group in receiving the favorable decision, and the AI system's overall tendency in granting the favorable decision—together, may influence the subject's fairness perceptions and retention in the AI system, we conduct an exploratory analysis by combining the data we obtained from both Study 1 and 2 and fitting them into regression models[9]. Specifically, we continue to use the two independent variables "advantaged" and "disadvantaged" to represent whether a subject's group has relative advantages or disadvantages over the other group in receiving the favorable decision from the AI system. Then, for each subject, we checked the AI system's decision outcome flowcharts for all the rounds in which she decided to apply for a loan, and we defined her "observed favorable decision probability" as the AI system's average probability of approving the loans, across all these rounds, for applicants who had both the same group and the same credit score category (i.e., $\geq$ 660 or <660) as her. We then included it into our regression models to reflect the subject's prospect of receiving the favorable decision. Using a similar approach, we can also compute, for each subject, her observed PPR, TPR, and FPR of the AI (regardless of applicants' group identity), and they are each incorporated in separate regression models to reflect the AI system's tendency of granting the favorable decision. Finally, we include the subject's risk attitude in the regression as a covariate. To account for the possible systematic differences between subjects of the two studies, we also include in the regression models an indicator variable "Study 2" to differentiate subjects of the two studies.

Results of our regression models are reported in Table 6. Regarding subjects' perceived fairness level of the AI system (Models 1–3), we consistently find that an AI system that grants more favorable decisions is perceived as fairer by subjects[10]. Moreover, the relative advantage of a subject's group over the other group in receiving the favorable outcome is also a driver of the subject's increases in her perceived fairness level of an AI system when the system is in favor of her group. In contrast, we find that the subject's prospect in receiving the favorable decision seems to be the sole driver for the changes in her retention in an AI system (Models 4–6).

## 5 CONCLUSIONS AND DISCUSSIONS

In this paper, via two experiments, we examined how decision subjects' fairness perception and retention in an AI-based decision system might be influenced by various factors, as they repeatedly interact with the system. Our results suggest that on average, a subject's fairness perceptions and retention in an AI-based decision system is significantly affected by the system's tendency to

[9]We were not able to conduct this exploratory analysis on the data of either study alone due to multicollinearity problems.

[10]The designs of our experimental treatments in both studies imply that the AI system's PPR, TPR and FPR are correlated, so we can not separate the impacts of these three metrics using our data. We also fit regression models in which the observed AI system's accuracy is included as a covariate rather than its PPR, TPR, or FPR. Despite in our experiment, an AI system's accuracy is correlated with the system's PPR, TPR, and FPR, our regression results suggest that an AI system's accuracy is not significantly correlated with subjects' fairness perceptions of it.

favor/disfavor the subject's group, but not the system's fairness level across groups, although we also detect individual differences between subjects with different qualification levels. Further investigations suggest that while decision subjects' fairness perceptions of an AI system may be influenced by the system's treatment on themselves and on others in a complex way, their retention in the system seems to be mostly driven by their own prospects of receiving the favorable decision from the system.

We now reflect on our findings, provide implications of our study, and discuss the limitations of our work.

**On the impact of the AI system's biased treatment across groups on decision subject's fairness perceptions.** Our finding of that the AI system's biased/unbiased treatment across group does not seem to affect subjects' average level of fairness perceptions of the system in repeated interactions is different from the results reported in Wang et al. [44], when decision subjects can only engage in a one-shot interaction with the AI system. One possible reason for this discrepancy is that in our experiments, we did not associate the group identity of loan applicants with specific socio-demographic features like race and gender, which may come with their unique social and historical contexts that can heighten people's sensitivity to inequality across groups. On the other hand, these results could also reflect the differences in decision subjects' fairness perceptions in repeated vs. one-shot interactions with AI systems. In particular, when decision subjects repeatedly interact with a biased AI system, the ones who are placed at the advantaged position by the system may realize that they could "materialize" the advantages by keeping interacting with the system, while those who are placed at the disadvantaged position can actively choose to "boycott" the system. This possibility for decision subjects to strategically interact with the AI-based decision system in the long run may shift the focus of their fairness perceptions to their own utility, rather than the equality across groups. Our investigation in Section 3.3.2 on the moderating role of decision subjects' qualification levels provides further nuanced results—It turns out that highly-qualified subjects will still significantly decrease their perceived fairness levels when the AI system exhibits biased treatment across groups. However, the low-qualified subjects actually perceive the biased AI system to be slightly fairer, and this is mainly caused by those low-qualified subjects who belong to the group that the AI system is in favor of. In other words, it is those decision subjects who do not deserve a favorable decision yet still be favored by the AI system, who substantially increase their fairness perception of the system, despite it being biased.

**On the complexity of fairness perceptions.** Our analysis in Study 2 suggests that in decision subjects' mind, the perception of "fairness" might be multifaceted. Fairness is partly about "*me*," i.e., how frequently I can get the favorable decision from the AI system. Fairness is also about "*me vs. others*," especially with respect to whether I get an advantage in receiving a favorable decision from the AI system in relative to people in the other groups. While this may appear to be directly contradicting to the classical group fairness definition (i.e., fairness is equality across groups), we suspect that decision subjects may utilize this cross-group contrast to gauge whether the AI system is fair to *me*, rather than whether the AI system is fair across different groups. In other words, decision subjects may not have a fixed standard when evaluating whether the AI system is fair to themselves, and they may need to rely on the comparison with others to make this call. Finally, fairness may also be about "*us*," e.g., how likely the AI system grants favorable decisions to people in general, regardless of their group identity. Our study design does not allow us to identify whether decision subjects' fairness perceptions are affected by the AI system's overall PPR, TPR, FPR, or a subset/all of these. It is also possible that individuals with different characteristics get affected by different factors—for example, highly-qualified subjects may care about overall TPR while low-qualified subjects may care about FPR, and future studies are needed to advance our understandings on this. It might also be useful for future studies to explicitly solicit different dimensions of fairness perceptions of AI, and rigorously examine how they, together, influence people's overall fairness perceptions of AI.

**Group retention in repeated interactions and implications.** While it is often believed that people's fairness perceptions of an AI system will influence their adoption of it, our study results suggest that the relationship between fairness perceptions and usage of the AI system is not linear, at least for decision subjects. For example, as shown in Table 3, while highly-qualified subjects were shown to significantly decrease their fairness perceptions of a biased AI system, they did not significantly decrease their retention in the AI system accordingly. In fact, as shown in Table 6, in our studies, subjects' retention in the AI system seems to be mainly driven by their prospects of receiving the favorable decision. This implies that in the long run, the group of decision subjects who have lower prospects of receiving the favorable decision might become increasingly under-represented over time, which can further influence the AI system's performance on the under-represented group as it continues to update its training data. This is true even if AI system is trained with fairness constraints but the factor equalized across groups by the constraints is not the positive prediction rate [50].

**Limitations and future work.** Our study was conducted in the context of AI-based loan lending systems, and we used a specific set of "parameter" values when designing the game in our experiments (e.g., the cost associated with the participation in the AI system and the reward brought up by a favorable decision). Cautions should be used when generalizing results in this work to different contexts and settings. For example, it would be interesting to see whether our results still hold when the reward/cost ratio is significantly larger or when the decision subjects have more "skin in the game." In addition, our study assumes that decision subjects have full knowledge of the AI system's performance on different groups. In practice, people may only obtain partial knowledge about the AI system's performance on others through, for example, their own social connections, who might be "similar" to themselves on some aspects due to homophily. It's therefore interesting to explore in these cases, how decision subjects' partial knowledge of the AI system's performance affect their fairness perceptions and retention.

## 6 ACKNOWLEDGMENTS

This work is supported in part by the NSF FAI program in collaboration with Amazon under grant IIS-2040800. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

## REFERENCES


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

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica, May 23, 2016.

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[4] Miranda Bogen and Aaron Rieke. 2018. Help wanted: an examination of hiring algorithms, equity, and bias.

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

[6] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (May 2021). https://doi.org/10.1145/3411764.3445308

[7] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[8] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[9] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64 (03 2018), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[10] 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

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. arXiv:1104.3913 [cs.CC]

[12] Pratik Gajane and Mykola Pechenizkiy. 2018. On Formalizing Fairness in Prediction with Machine Learning. arXiv:1710.03184 [cs.LG]

[13] Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115 (2021), 106607. https://doi.org/10.1016/j.chb.2020.106607

[14] Nina Grgić-Hlača, Adrian Weller, and Elissa M Redmiles. 2020. Dimensions of diversity in human perceptions of algorithmic fairness. *arXiv preprint arXiv:2005.00808* (2020).

[15] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 555–565.

[16] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[17] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[18] Anil Kalhan. 2013. Immigration Policing and Federalism Through the Lens of Technology, Surveillance, and Privacy. *Political Institutions: Federalism & Sub-National Politics eJournal* (2013).

[19] Cindy Kam. 2012. Risk Attitudes and Political Participation. *American Journal of Political Science* 56 (10 2012). https://doi.org/10.1111/j.1540-5907.2012.00605.x

[20] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[21] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future Developers' Perception of Fairness in Algorithmic Decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 690–700.

[22] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics* 133, 1 (08 2017), 237–293. https://doi.org/10.1093/qje/qjx032 arXiv:https://academic.oup.com/qje/article-pdf/133/1/237/30636517/qjx032.pdf

[23] Heidi Ledford. 2019. *Millions of black people affected by racial bias in health-care algorithms*. Retrieved Feb 12, 2022 from https://www.nature.com/articles/d41586-019-03228-6

[24] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[25] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 381–391.

[26] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[27] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the applicability of ml fairness notions. *arXiv preprint arXiv:2006.16745* (2020).

[28] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *CoRR* abs/1908.09635 (2019). arXiv:1908.09635 http://arxiv.org/abs/1908.09635

[29] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[30] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (Mar 2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[31] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. 2002. Multi–objective Evolutionary Algorithms for the Risk–return Trade–off in Bank Loan Management. *International Transactions in Operational Research* 9 (2002), 583–597.

[32] Pegah Nokhiz, Aravinda Kanchana Ruwanpathirana, Neal Patwari, and Suresh Venkatasubramanian. 2021. Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 199–208.

[33] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[34] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[35] 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*. 1–52.

[36] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[38] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 2021. A study on fairness and trust perceptions in automated decision making. *arXiv preprint arXiv:2103.04757* (2021).

[39] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I Hong. 2020. Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[40] 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*. 2459–2468.

[41] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[42] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) *(FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[43] Neil Vigdor. 2019. Apple card investigated after gender discrimination complaints. *The New York Times* 10 (2019).

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

[45] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022*. 1697–1708.

[46] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. 2021. A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 1023–1033.

[47] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[49] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[50] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. 2019. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems* 32 (2019).

[51] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* 33 (2020), 18457–18469.

[52] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.