

# Understanding Decision Subjects' Engagement with and Perceived Fairness of AI Models When Opportunities of Qualification Improvement Exist

Meric Altug Gemalmaz<sup>1</sup>, Ming Yin<sup>1</sup>

<sup>1</sup>Purdue University, West Lafayette, Indiana, USA  
mgemalma@purdue.edu, mingyin@purdue.edu

## Abstract

We explore how an AI model's decision fairness affects people's engagement with and perceived fairness of the model if they are subject to its decisions, but could repeatedly and strategically respond to these decisions. Two types of strategic responses are considered—people could determine whether to continue interacting with the model, and whether to invest in themselves to improve their chance of future favorable decisions from the model. Via three human-subject experiments, we found that in decision subjects' strategic, repeated interactions with an AI model, the model's decision fairness does not change their willingness to interact with the model or to improve themselves, even when the model exhibits unfairness on salient protected attributes. However, decision subjects still perceive the AI model to be less fair when it systematically biases against their group, especially if the difficulty of improving one's qualification for the favorable decision is larger for the lowly-qualified people.

## Introduction

The rapid development of Artificial Intelligence (AI) technologies has made it possible to automate decision making in many domains. However, it has been discovered that AI models often acquire pre-existing biases in the dataset used for their training, resulting in the unfair treatment of individuals from different demographic backgrounds (Mesa 2021; Dastin 2018). This increased awareness of fairness issues of AI has led to many recent studies in understanding people's fairness perceptions of and reactions to AI models (Wang, Harper, and Zhu 2020; Gemalmaz and Yin 2022; Harrisson et al. 2020). These studies look into the perspectives of different stakeholders, among which a key stakeholder is *decision subjects*, the people who are actually subject to the AI models' decisions. For example, Wang, Harper, and Zhu (2020) found that when decision subjects only interacted with an AI model once, both the model's unbiased decisions across different groups of subjects and the model's favorable decisions towards subjects' own group resulted in an increase in the perceived fairness of the AI model.

Meanwhile, recent research on the long-term dynamics and implications of fairness in AI (Liu et al. 2018, 2020; Zhang et al. 2020; Zhang and Liu 2021; Zhang et al. 2019) has drawn the community's attention to the fact that in reality, decision subjects could often interact with an AI model

*repeatedly* over a long term. Moreover, in each interaction, decision subjects may no longer passively accept the AI model's decisions on them as is. Rather, the AI model's decisions on subjects may shape how they *actively and strategically respond to the AI model*. For instance, one strategic response decision subjects could take in their repeated interactions with the AI model is to decide whether to continue interacting with it and be subject to its decisions (Zhang et al. 2019; Gemalmaz and Yin 2022)—decision subjects have the freedom to quit using an AI model if they wish so. As another example, decision subjects could also respond to AI decisions on them by investing in effort to improve themselves, hoping for an increased chance of receiving the favorable decision in the future (Zhang et al. 2020; Liu et al. 2020)—job applicants could take additional courses about a skill, and loan applicants could explore options to increase their credit scores, both aiming to improve their "*qualification*" for the favorable decision (i.e., getting the job offer or the loan approval). The possibility to improve one's qualification may encourage decision subjects to take a long-term view to think about the future when responding to the AI model at present, and it may indirectly influence their willingness to continue interacting with the AI model.

As decision subjects could repeatedly and strategically respond to AI decisions in many real-world scenarios, understanding their reactions to AI models with different fairness properties and what they perceive as "fair" become critical again. Specifically, we ask that when decision subjects can strategically and repeatedly respond to AI decisions:

- **RQ1:** How will the AI model's fairness properties (both across groups and on the subject's group) affect decision subjects' engagement with the model (e.g., willingness to improve themselves and to be subject to AI decisions)?
- **RQ2:** How will the AI model's fairness properties affect decision subjects' perceived fairness of the model?

Predicting answers to these questions turns out to be very challenging. In terms of the willingness to improve one's qualification for the favorable decision, it is possible that decision subjects decide how much to improve themselves solely based on whether doing so increases their utility, and is not affected by the AI model's decision fairness at all. However, one may also speculate that the AI model's biases against a subject's group could affect their drive to improve themselves. For example, if the AI model consistently

places a subject's group at a disadvantage position in its decisions, individuals in that group might feel they are being treated as "second-class citizens". This feeling could diminish their motivation to improve their qualifications. On the other hand, recognizing the bias, they might be even more determined to improve, seeing it as the sole avenue to increase their odds of favorable decisions and level the playing field with people from other groups. Without a clear hypothesis on how the AI model's decision fairness affects decision subjects' willingness to improve their qualification, predicting how their willingness to keep interacting with the AI model or their perceived fairness of the AI model are affected by the AI model's fairness properties also becomes difficult. This is because both retention and fairness perceptions can be highly influenced by the final qualification level that decision subjects could reach.

To complicate things further, answers to these questions may also vary across different contexts. For example, one may conjecture that if improving one's chance of getting the favorable decision is particularly difficult for those who really "need" the improvement (i.e., those with relatively low qualification), the impact of AI fairness on decision subjects may be more salient, as the hope of changing one's fate through efforts and self-improvement is limited. Similarly, it is also possible that the impact of AI fairness on decision subjects is larger if AI exhibits discriminatory behavior on some salient protected social attributes, triggering people's strong emotional attachment to their own group identities. Formally, one may ask that when decision subjects can strategically and repeatedly respond to AI decisions:

- **RQ3:** Do answers to RQ1–RQ2 change if the difficulty for decision subjects to improve their qualification vary with their current qualification level in different ways?
- **RQ4:** Do answers to RQ1–RQ2 change when the AI model's fairness properties is/is not discussed with respect to groups defined by protected social attributes, such as gender?

To answer these questions, we conducted three exploratory human-subject studies on Amazon Mechanical Turk ( $N = 368, 713, 416$  for the three studies, respectively). In all three studies, subjects completed a simulated loan application task that was designed to mirror real-world loan application scenarios where the loan decisions are made by an AI model. Subjects were free to decide how many times to apply for a loan from the AI model, and whether to improve their own qualification (i.e., their credit score) before each application. In each study, we created two treatments to reflect that the AI model may or may not show systematic bias towards one group over the other in granting loans. Study 1 was designed to answer **RQ1** and **RQ2**, so the difficulty for subjects to improve their qualification does not vary with their current qualification level, and subject's group identity was randomly assigned. To answer **RQ3**, we slightly varied the design of Study 2 so that the difficulty of qualification improvement either increased or decreased with the subject's current qualification level. Finally, to answer **RQ4**, subject's group identity in Study 3 was decided by their self-reported gender rather than a randomly assigned value.

Our results show that when decision subjects could repeatedly and strategically respond to the AI model's decisions on them, their engagement with the model—including their willingness to improve their qualification and willingness to keep interacting with the model—are *not* influenced by the AI model's decision fairness. This holds true both when the difficulty of qualification improvement changes with one's current qualification level in different ways, and when the AI model's fairness is/is not examined with respect to salient protected attributes like gender. However, we find that despite the possibility of strategic responses, decision subjects still perceive the AI model as less fair if the model biases against the subjects' group by placing them at a disadvantaged position in receiving the favorable decision. In other words, when decision subjects can strategically and repeatedly respond to AI decisions, their level of engagement with an AI model does *not* reflect either the model's group-level decision fairness or their perceived fairness of the model. We conclude by providing possible explanations for our findings and discussing their implications,

## Related Work

The complexity of the notion of "AI fairness" has inspired many studies on understanding whether and when do humans perceive an AI model as "fair" in decision making (Yaghini, Krause, and Heidari 2021; Saxena et al. 2019; Harrison et al. 2020; Srivastava, Heidari, and Krause 2019). Of particular relevance to our study are a few recent works on understanding *decision subjects'* fairness perceptions of an AI model that makes decisions about them. Earlier studies typically focus on one-shot interaction scenarios where decision subjects would only receive a decision from an AI model once. For example, Yurrita et al. (2023) studied the influence of explanations, human oversight, and contestability on fairness perceptions of the decision subjects in a one-shot interaction with the AI model for loan approvals. They found that while explanations and contestability significantly impacted fairness perceptions, human oversight showed minimal effect. In another study involving one-shot interaction with the AI model, decision subjects perceive an AI model as fairer both if the AI model makes a favorable decision on them and if the AI model is not biased towards or against any particular group (Wang, Harper, and Zhu 2020).

More recently, there is a line of theoretical works on "long-term fairness" (Liu et al. 2018; Hu and Chen 2018; Zhang et al. 2019; D'Amour et al. 2020; Mouzannar, Ohannessian, and Srebro 2019; Heidari, Nanda, and Gummadi 2019) emphasizing that in the real world, decision subjects often engage in *repeated* interactions with the AI model, and the dynamics between the AI model's decisions on subjects and subjects' *strategic* reactions to those decisions could create feedback loops. One domain that is frequently studied in the long-term fairness literature is loan lending—For a loan applicant characterized by a profile  $x$ , the bank may use its AI-based loan approval system to make loan lending decisions on the applicant. Moreover, the applicant may strategically respond to these decisions by staying or leaving the system and/or changing their profile  $x$ , which may further change the AI model's training data and impact the model's

decision-making policy in the future. This shift from one-shot to long-term, repeated interaction has inspired some empirical studies looking into how decision subjects react to and perceive AI models in their long-term interactions with AI. For example, it was found that when decision subjects had the choice to leave the AI-based decision system at any time in their repeated interactions, their willingness to stay in the system and their perceived fairness of the system are significantly affected by whether the system is in favor of the subject's own group (Gemalmaz and Yin 2022). Compared to earlier works, in this study, we take into account another key strategic action that decision subjects could take in their repeated interactions with AI. That is, decision subjects could also freely decide whether to improve their qualifications for receiving future favorable decisions from AI. In the real world, these qualification improvement attempts are usually realized through the adjustment of decision subjects' input attributes, possibly as the decision subjects follow the algorithmic recourse plans suggested by the AI model to change their situation towards receiving a more favorable decision (Ustun, Spangher, and Liu 2019).

We believe that the addition of qualification improvement in the set of decision subjects' strategic actions brings new perspectives for re-examining the relationship between AI's decision fairness and decision subjects' reactions to and perceptions of it. Indeed, the possibility to improve their qualification for future favorable decisions may shift decision subjects' attention from *focusing on the present* to *thinking about the (long-term) future*, which may have complicated implications on how decision subjects would react to the AI model at present. For example, because of their future thinking, decision subjects may change how much risks they are willing to take or how they treat the immediate and future rewards (Thorstad and Wolff 2018; Hershfield 2011). It may also change the ways that decision subjects weigh utility-related considerations (e.g., how many favorable decisions can I get from the AI model in the long run?) and fairness-related considerations (e.g., is the AI model's decision on me fair?) in deciding their engagement with the AI model. It is even possible for decision subjects to change how they define "fairness" for AI. This is because the possibility of qualification improvement may allow decision subjects to evaluate the AI fairness not only through "*social comparison*" (e.g., does AI make similar decisions on my group and other groups? Festinger (1954)) but also through "*temporal self-comparison*" (e.g., does AI grant more favorable decisions to me after my qualification is improved? Albert (1977)).

A few theoretical works have examined the implications of AI fairness in scenarios where decision subjects can respond to AI decisions strategically by improving their qualifications (Zhang et al. 2020; Liu et al. 2020; Mouzannar, Ohannessian, and Srebro 2019). However, none of these works focus on *empirically* characterizing how decision subjects would respond to AI models with different levels of fairness to determine their qualification transitions. Rather, they make simplified assumptions about subjects' behavior in their theoretical derivations. For example, Liu et al. (2020) assumed that decision subjects decide about their qualification improvement based on a cost-benefit analysis by ex-

amining whether the increase in utility after the qualification is improved is larger than the cost of improvement. Zhang et al. (2020) considered whether decision subjects received a favorable decision from the AI model in one interaction as the key influencing factor for them to decide whether to improve their qualification in the next interaction. Different from these theoretical works, we use experiment with real human subjects to provide empirical insights into whether and how the AI model's decision fairness affects decision subjects' qualification transitions in repeated interactions, and their engagement with/perceived fairness of the AI model in general. Thus, results of our study may help verify or reject assumptions made in previous works.

## Study 1

To understand how an AI model's decision fairness affects decision subjects' repeated and strategic interactions with the AI model, we conducted a series of human-subject experiments<sup>1</sup>. In Study 1, we aim to first answer **RQ1–RQ2** in an environment where (1) the AI model's fairness properties are *not* examined with respect to a salient protected attribute, and (2) the difficulty for a decision subject to improve their qualification for the favorable decision does *not* change with the subject's current qualification level.

### Experimental Design

**Tasks.** Subjects in our experiment were asked to complete a simulated loan application task, which was carefully designed to mimic the real-world loan application scenario that is often used in theoretical studies of the long-term dynamics of AI fairness (Liu et al. 2018; Zhang et al. 2020; D'Amour et al. 2020) and experimental studies of fairness perceptions of AI (Gemalmaz and Yin 2022; Yurrita et al. 2023). Specifically, each subject was assigned with a randomly-generated persona of a small business owner containing 5 attributes; this persona was used as the subject's loan application "profile". We highlight two key attributes in the persona:

- **Group identity:** The subject's group membership, which was set to be either "red" or "blue". This is to reflect that in Study 1, the AI model's fairness properties across groups are *not* defined on salient protected attributes.
- **Initial credit score level:** The subject's credit score level at the beginning of the experiment, which was taken from the set  $\{300\text{--}350, \dots, 500\text{--}550, \dots, 700\text{--}750\}$ ; the subject could later decide to "improve" their credit score with some cost (see details below).

In addition to the above two attributes, the subject's persona also included three other attributes—their number of years of having a credit history, their home ownership status (e.g., rent or own), and the type of small business they ran (e.g., healthcare, construction). For each attribute in a subject's persona, we uniformly randomly sampled a value from the set of all candidate values for that attribute. After getting their assigned profile, the subject was asked to use it to apply for loans from their "local bank" to support their

<sup>1</sup>All of our experiments were approved by the IRB of the authors' institution.

business, and they were explicitly told that the bank utilized an *AI model* to analyze loan application profiles and make loan approval decisions. The subject was also informed that their credit score level would be used by the AI model as a primary determinant of their “qualification” for the loans—credit scores of 650 or higher were generally considered as “high”, and higher credit scores were associated with higher chance of getting loans approved.

In the experiment, the subject started the experiment with 600 “coins” in their account. Each loan application cost the subject 50 coins, and the subject would gain 100 coins if the application got approved or nothing otherwise. Each subject was asked to apply for a loan from the bank for at least once to get a sense of the AI model’s decision fairness. After that, they could interact with the AI model for at most 9 more rounds. In each round, the subject had the freedom to choose from one of three actions:

- **Improve and apply:** The subject would first attempt to improve their credit score to the next level (e.g., from 600–650 to 650–700) with a cost of 5 coins<sup>2</sup>, and then apply for a loan with a cost of 50 coins. The credit score improvement attempt was *not* guaranteed to be successful (see details below)—if successful, subjects would be notified, and the AI model would use the updated credit score to make the loan approval decision.
- **Apply without improvement:** The subject would directly apply for a loan with a cost of 50 coins without attempting to improve their credit score level. This action was provided to subjects to reflect that in reality, whether and when to improve one’s qualification is a voluntary (and strategic) decision.
- **Not apply any more:** The subject would not apply for loans any more and would be redirected to the end of the experiment. This action was provided to subjects to reflect that in reality, whether to continue interacting with an AI model is a voluntary (and strategic) decision.

Since Study 1 concerns an environment where the difficulty for decision subjects to improve their qualification does *not* change with their current qualification level, we set the “success rate” for subjects across *all* credit levels to progress to the next level at a constant value (i.e., 44%). That is, if the subject attempted to improve their qualification in one round, whether they could successfully progress to the next credit level would be stochastically decided by this success rate<sup>3</sup>. Once the subject successfully progressed to the next credit level, they would at least maintain that level, and possibly progress to even higher levels if they decided to make additional improvement attempts in future rounds.

At the end of each round, if the subject applied for a loan in that round, the AI model’s approval or denial decision on

---

<sup>2</sup>Via a simulation study, we found that an improvement cost of 5 coins is an intermediate level of cost that requires subjects to carefully deliberate about whether and when to improve their qualification. If the improvement cost was too low (or high), subjects may simply opt for always improving their qualification (or never improving their qualification). See Appendix J for more details.

<sup>3</sup>Subjects were not explicitly told about this success rate but could experience it through their actual improvement attempts.

the subject would be revealed to them, without providing explanations on why it makes this decision. To enable subjects to perceive the AI model’s decision fairness, a flowchart summary of the AI model’s decisions in this round on applicants of both red and blue groups would also be provided to the subject (see Figure A1 in appendix for an example).

Figure 1 illustrates the process of the simulated loan application task. We note that the designs of this task reflect a few key characteristics of the real-world repeated interactions between decision subjects and an AI model: (1) participating in the decision making process (thus triggering the usage of the AI model) is costly; (2) receiving a favorable decision from the AI model is rewarding; (3) the improvement of qualification is costly and has uncertainty; and (4) decision subjects have the freedom to respond to the AI model’s decisions by deciding whether to improve their qualification and/or whether to continue being subject to the AI model’s decision. By assigning each subject with a persona of a small business owner, our experiment reflects a scenario where the real-world decision subjects will often interact with the AI-based decision systems *repeatedly*, as small business owners often need to apply for loans at different time points to meet their business’s various financing needs.

**Treatments.** We created two treatments in Study 1:

- **Fair AI model:** In this treatment, the bank’s AI model was fair towards subjects in both the red group and the blue group. In particular, as shown in Table 1a, *regardless of the subject’s group identity*, the AI model’s loan approval rate always started from 15% for subjects with the lowest credit level (i.e., 300–350), and the approval rate increased by 7% as the subject’s credit score went one level up, with a highest possible approval rate of 85% for subjects with the highest credit level (i.e., 800–850).
- **Unfair AI model:** In this treatment, the bank’s AI model was unfair and *systematically biased against subjects in the blue group*. Table 1b shows this model’s approval rate for subjects in the red group, while Table 1c shows this model’s approval rate for subjects in the blue group. As shown in the tables, for every credit level, the AI model’s approval rate for red group subjects with this credit level was always 20% higher than that for blue group subjects<sup>4</sup>. Same as that in the previous treatment, the AI model’s increment in approval rate for each increased credit level was kept at 7% for both red and blue groups.

So, if the subject decided to apply for a loan in one round, the AI model’s loan approval decision on them would be stochastically decided by the approval rate for their most updated credit level, given the subject’s treatment assignment and group identity. As described earlier, at the end of each round where the subject applied for a loan, we also presented to the subject a summary of the AI model’s decisions across applicants in different groups. In particular, we told the subject that there are a total of  $N \sim U[12000, 12200]$  applicants

---

<sup>4</sup>Note that since the subject’s group identity was sampled uniformly randomly from the two candidate values, for each credit level, the expected chance for a subject getting their loans approved in the “unfair AI” treatment was still the same as a subject with the same credit level in the “fair AI” treatment.

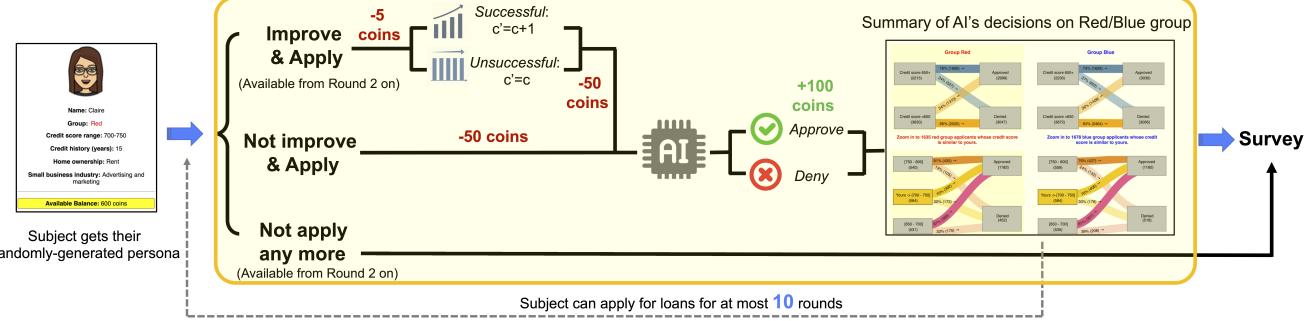


Figure 1: An illustration of the process of the loan application task. Here,  $c$  is the subject’s current qualification level, while  $c'$  denotes the qualification level after an improvement attempt, which can either remain the same or advance to the next level.

Credit/Decision	Approve	Deny
800–850	85%	15%
750–800	78%	22%
$\vdots$	$\vdots$	$\vdots$
350–400	22%	78%
300–350	15%	85%

(a) Fair AI: Red/Blue group

Credit/Decision	Approve	Deny
800–850	95%	5%
750–800	88%	12%
$\vdots$	$\vdots$	$\vdots$
350–400	32%	68%
300–350	25%	75%

(b) Unfair AI: Red group

Credit/Decision	Approve	Deny
800–850	75%	25%
750–800	68%	32%
$\vdots$	$\vdots$	$\vdots$
350–400	12%	88%
300–350	5%	95%

(c) Unfair AI: Blue group

Table 1: The AI model’s probability of approving/rejecting loan applications in different treatments.

who applied for loans from the bank in this round. For each of these  $N$  applicants, we randomly generated their persona, and then simulated the AI model’s decision on them using the approval rates defined by *the subject’s assigned treatment*. We then visualized the AI model’s decisions on all  $N$  applicants using two sets of flowcharts. The first set showed the AI model’s approve/deny decisions for applicants with-/without “high” credit scores (i.e., a score of at least 650). The second set zoomed in to applicants with similar credit scores as the subject and showed the AI model’s approve-/deny decisions for applicants whose credit level was one level higher than, the same as, or one level lower than the subject (see Figure A1 in appendix)<sup>5</sup>. Note that within each set, the AI model’s decisions on red/blue group applicants were shown in separate flowcharts; this enables subjects to make direct comparisons across the two groups of applicants to determine the AI model’s decision fairness.

**Procedure.** Our experiment was made available as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (MTurk). Only U.S. workers with a HIT approval rate of at least 95% and a total of at least 1000 approved HITs were eligible for taking this HIT.

Upon arrival, the subject first created a nickname and selected an avatar for their persona. Next, we provided an interactive tutorial explaining each attribute’s meaning on the subject’s randomly-assigned persona profile. Additionally, this tutorial explained to subjects what they were asked to do in each round, how to use the interface, and how to interpret the summary information of the AI model’s decisions

<sup>5</sup>We used flowcharts to visualize the AI model’s decisions because prior research has showed their effectiveness in aiding the comprehension of algorithmic model performance among non-experts (Shen et al. 2020). We also provided textual explanations to help subjects interpret the numbers shown in the flowcharts.

as displayed in the flowcharts. We then used a quiz of 5 questions to test subjects’ understanding of the task procedure and ability to interpret flowcharts. Subjects were only allowed to advance to the actual experiment after correctly answering all 5 questions.

In the actual experiment, the subject was first randomly assigned to one of the two treatments. Then, the subject went through the simulated loan application task. Note that throughout the task, the subject’s credit level would be updated if the subject’s improvement attempt was successful, and the subject’s account balance would be updated depending on the loan application outcomes. Once the subject finished the loan application task, they were redirected to our post-experiment survey. The survey included questions about the subject’s demographics (e.g., race, age, education) and perceived fairness of the AI model. We also measured a few other characteristics of the subject that we conjectured to influence their behavior in interacting with the AI model, such as their risk attitude. All questions except for the demographics were presented as 5-point Likert scale questions in which the subject needed to indicate their agreement with a set of statements from 1 (strongly disagree) to 5 (strongly agree). The subject could also explain why they considered the AI model they encountered in the experiment as fair or unfair using free-form texts. The complete list of survey questions are provided in Appendix B.

Subjects were told that their payment in this experiment was composed of a *base payment* of \$2 and a *bonus payment* that would be decided by their account balance at the end of the experiment. Upon survey completion, we converted the remaining balance in the subject’s account to their bonus payment using a 500 coins to \$2 ratio. The maximum bonus a subject could earn was \$4.40. Subjects spent a median time of 27 minutes on our experiment and received a median payment of \$4.30, resulting in an hourly wage of \$9.60. We

also included three filtering procedures to filter out potential spammers (see Appendix C for details). A subject's data was only considered valid if they passed all filtering procedures.

## Analysis Methods

We used regression analyses to answer our research questions. Specifically, we used three dependent variables to quantify decision subjects' engagement with an AI model (**RQ1**) and their fairness perceptions of it (**RQ2**):

- **Engagement-Improvement:** The number of qualification improvement attempts the subject made in the loan application task; a higher value indicates a higher level of willingness for the subject to improve themselves.
- **Engagement-Retention:** The number of times that the subject applied for a loan; a higher value indicates a higher level of willingness for the subject to continue being subject to the AI model's decisions.
- **Perceived Fairness:** The subject's rating to six statements (e.g., "The bank's AI system is fair to manage loan applications.") adapted from Wang, Harper, and Zhu (2020) regarding their perceptions of the AI model's fairness; the higher the rating, the fairer the subject found the AI model to be (the max rating is 30).

The independent variable we included in the regression reflects the AI model's "*fairness properties*," which was operationalized in two ways. First, to explore how subjects' engagement with and perceived fairness of AI vary with the AI model's decision fairness *across groups*, we used a binary variable **Fair AI** to represent if the AI model treats the two groups of loan applicants in a similar way (i.e., it was set to 1 for subjects in the fair AI treatment and 0 otherwise). Second, to understand how subjects' engagement with and perceived fairness of AI are affected by the AI model's decision fairness *on the subject's own group*, we used two other binary variables in our regressions—**Advantaged** and **Disadvantaged**, representing if the subject's group was favored or disfavored by the AI model. That is, Advantaged (Disadvantaged) was set to 1 only for red (blue) group subjects in the unfair AI treatment.

Finally, to control for the influences on dependent variables beyond those brought up by the independent variables, we considered a few characteristics of the subjects and included them as covariates:

- **Initial Credit Score:** The credit score level that was assigned to the subject at the beginning of the experiment. We conjectured that subjects with higher credit levels make fewer improvement attempts as there is smaller room of improvement for them. However, they might be more willing to interact with the AI model and even perceive it as fairer, because they were more likely to receive loan approval decisions from the AI model.
- **Fairness Sensitivity:** The degree to which the subject values fairness as a core principle, which was measured in the post-experiment survey through soliciting the subject's opinions on a set of statements (e.g., "I would stop using an AI system if it is unfair, even if it tends to be in favor of me") adapted from Gemalmaz and Yin (2022). We conjectured that subjects' fairness sensitivity may af-

flect how they react to the AI model's decision fairness.

- **Empathy:** The subject's empathy level, which was measured in the post-experiment survey through soliciting the subject's agreement with a set of statements (e.g., "I get a strong urge to help when I see someone who is upset") adapted from Spreng et al. (2009). We conjectured that individuals with higher levels of empathy may exhibit stronger concerns regarding the fairness of an AI model, as they are concerned with the well-being of others even if they are not personally affected by the AI model's bias.
- **Risk Attitude:** The subject's risk attitude, which was measured through soliciting the subject's opinions on a set of statements (e.g., "I like to do frightening things.") adapted from Kam (2012). We conjectured that subjects who were more risk-seeking might be more willing to take actions to improve their qualification or continue interacting with the AI model. Previous research has also found that people with higher risk-taking tendencies tend to perceive AI systems as fairer compared to those who are less risk-takers (Nakao et al. 2022).

We fit our experimental data into regression models to predict subjects' engagement (**RQ1**)—including their retention and improvement—and fairness perceptions (**RQ2**). To emphasize the exploratory nature of this study, we followed the interval estimate method (Cumming 2014; Dragicevic 2016) in our analysis. That is, our regression results are interpreted via the estimated coefficient values for the independent variables as well as their 95% bootstrap confidence intervals ( $R = 1000$ ). When the 95% bootstrap confidence interval of the coefficient for an independent variable does not include zero, we consider the effect of the independent variable to be reliable (Cumming 2014).

## Experimental Results

368 subjects participated in our experiment and passed all filtering procedures (see Appendix D for the demographics)<sup>6</sup>. In the following, we analyzed the full dataset collected from these 368 subjects to answer our research questions.

**RQ1: Impacts of AI's decision fairness on engagement.** First, we look into that in decision subjects' strategic, repeated interactions with an AI model, how the AI model's decision fairness affects their engagement with the model, including their willingness to improve their qualification and their willingness to be subject to the AI model's decisions (see Figure A2 in Appendix E for the histograms of the number of improvement attempts/loan applications made by subjects in different treatments or different groups).

Regarding the willingness to improve their qualification, the average number of times that improvement attempts

<sup>6</sup>We conducted a priori power calculations using the G\*Power software (Faul et al. 2007). For a regression model, we assumed an effect size  $f^2 = 0.05$ , an  $\alpha$  error probability of 0.05, and a desired power level of 0.95. Results suggest that 262 subjects are needed for achieving the desired level of treatment effect for the independent variable "Fair AI", while 312 subjects are needed for achieving the desired level of effect for the independent variables "Advantaged" and "Disadvantaged". Thus, in each (sub)experiment of our studies, we targeted at recruiting at least 312 subjects.

	Improvement		Retention		Perceived Fairness	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fair AI	-0.03 [-0.73, 0.62]	0.49 [+0.32, 1.38]	-0.42 [-1.18, 0.3]	0.77 [-0.18, 1.76]	0.59 [-0.47, 1.69]	0.27 [-0.95, 1.53]
Advantaged				0.08 [-0.88, 1.01]		<b>-1.41<sup>§</sup></b> [-2.78, -0.11]
Disadvantaged						
Risk attitude	0.05 [-0.03, 0.14]	0.05 [-0.04, 0.12]	0.04 [-0.04, 0.17]	0.07 [-0.05, 0.16]	0.06 <sup>§</sup> [0.2, 0.5]	0.36 <sup>§</sup> [0.19, 0.49]
Fairness Sensitivity	-0.00 [-0.16, 0.14]	-0.01 [-0.16, 0.14]	-0.03 [-0.2, 0.14]	-0.03 [-0.2, 0.13]	-0.11 [-0.38, 0.12]	-0.13 [-0.38, 0.1]
Empathy	0.13 <sup>§</sup> [+0.00, 0.24]	0.13 <sup>§</sup> [+0.00, 0.24]	0.12 [-0.02, 0.25]	0.12 [-0.02, 0.25]	-0.11 [-0.31, 0.09]	-0.11 [-0.3, 0.09]
Initial Credit Score	-0.13 <sup>§</sup> [-0.26, -0.01]	-0.13 <sup>§</sup> [-0.25, -0.00]	0.22 <sup>§</sup> [+0.07, 0.37]	0.22 <sup>§</sup> [0.08, 0.38]	0.19 <sup>§</sup> [+0.00, 0.37]	0.20 <sup>§</sup> [0.02, 0.39]
Constant	2.49 <sup>§</sup> [0.13, 4.86]	2.54 <sup>§</sup> [0.22, 4.85]	3.76 <sup>§</sup> [1.1, 6.36]	3.42 <sup>§</sup> [0.74, 6.07]	17.51 <sup>§</sup> [13.34, 21.75]	18.27 <sup>§</sup> [14.19, 22.39]

Table 2: Regression models predicting decision subjects' improvement, retention, and perceived fairness based on the AI model's decision fairness for Study 1. Coefficients and their 95% bootstrap confidence intervals are reported. A superscript <sup>§</sup> indicates that the estimated coefficient is reliably different from zero. Reliable effects of the independent variables of interests are bolded.

were made was  $M_{\text{fair}} = 4.11$  ( $SD = 3.43$ ) and  $M_{\text{unfair}} = 4.11$  ( $SD = 3.26$ ) for subjects in the fair AI and unfair AI treatments, respectively. Model 1 in Table 2 examines whether the AI model's fairness level *across groups* has any impact on subjects' willingness to improve their qualification. Here, the estimated coefficient for the independent variable "Fair AI" was not reliably different from zero ( $\beta = -0.03[-0.73, 0.62]$ ). This suggests that subjects' average level of willingness to improve their qualification is *not* impacted by the AI model's decision fairness across groups. Moreover, as shown in Table 2 (Model 2), we also find that the AI model's decision fairness *on the subject's own group* does not impact their willingness to improve, as neither of the coefficients associated with "Advantaged" and "Disadvantaged" are reliably different from zero.

Similar observations can also be made for decision subjects' retention. On average, subjects in the fair AI treatment interacted with the AI model for  $M_{\text{fair}} = 6.23$  ( $SD = 3.93$ ) rounds, while subjects in the unfair AI treatment interacted with the AI model for  $M_{\text{unfair}} = 6.58$  ( $SD = 3.72$ ) rounds. Regression results are shown in the middle panel of Table 2 (Models 3 and 4). We find that once decision subjects can strategically react to the AI model's decisions on them, on average, they are equally willing to keep interacting with the AI model regardless of the model's decision fairness, both across groups and specifically towards their group.

Interestingly, across Models 1–4 in Table 2, we also notice that subjects with higher initial credit score levels consistently made fewer improvement attempts, but interacted with the AI model for more rounds. This is expected and consistent with our conjecture.

**RQ2: Impacts of AI's decision fairness on fairness perceptions.** Subjects in the fair AI treatment reported an average fairness rating of 18.77 ( $SD = 5.15$ ) for the AI model. Meanwhile, subjects in the unfair AI treatment reported an average fairness rating of 18.02 ( $SD = 5.40$ ), with those who were placed at the advantaged position (i.e., the red group) reported an average rating of 19.0 ( $SD = 4.75$ ) and those who were placed at the disadvantaged position (i.e., the blue group) reported an average rating of 17.0 ( $SD =$

5.81). Table 2 (Models 5 and 6) shows the results of the regression analysis. While the AI model's fairness level across different groups still does not appear to reliably influence decision subjects' average level of perceived fairness of the model (Model 5), we do find that when the AI model systematically biases against the subject's group, the subject perceives the model as less fair (i.e.,  $\beta = -1.41[-2.78, -0.11]$  for "Disadvantaged" in Model 6). Also, consistent with our conjecture, we also observed that subjects who were more risk-seeking or had a higher initial credit score level tended to perceive the AI model as fairer.

## Study 2

In Study 1, we have answered **RQ1** and **RQ2** in an environment where the difficulty for decision subjects to improve their qualification does not vary with their current qualification levels. In Study 2, we explore the generalizability of our Study 1 results in different environments where the qualification improvement difficulty varies with one's current qualification level and therefore answer **RQ3**.

### Experimental Design

In Study 2, we conducted two sub-experiments simultaneously. The design of each sub-experiment was identical to the experiment designed for Study 1 with only one exception: In Study 1, the success rate for a subject to progress to the next credit level after they make an improvement attempt was set at a fixed value, while we vary this value in different ways in each sub-experiment of Study 2:

- **Easy to hard:** In this sub-experiment, it is easier for lowly-qualified subjects to improve their qualification than highly-qualified subjects. Specifically, the success rate for subjects with the lowest credit level (i.e., 300–350) in an improvement attempt was set to 80% and the success rate decreased by 8% for each increased level of credit score. The lowest possible success rate was 8% for subjects with the highest credit level who could still make an improvement attempt (i.e., 750–800)<sup>7</sup>.
- **Hard to easy:** In this sub-experiment, it is more difficult for lowly-qualified subjects to improve their qualification than highly-qualified subjects. Specifically, the success rate in an improvement attempt ranged from 8% (for subjects with the 300–350 credit level) to 80% (for subjects with the 750–800 credit level), and it increased by 8% for each increased level of credit score.

The experimental procedure was the same as Study 1 except that: (1) Workers who had previously participated in the experiment in Study 1 were excluded from participating in this new experiment; (2) Workers were randomly assigned to one of the two sub-experiments as defined above.

### Experimental Results

A total of 713 subjects participated in our experiment and passed all filtering procedures ("Easy to hard": 328 subjects,

<sup>7</sup>Note that the average success rate across subjects of different credit levels was still 44%, which was the same as the constant success rate used in Study 1.

	Improvement				Retention				Perceived Fairness			
	Easy to hard		Hard to easy		Easy to hard		Hard to easy		Easy to hard		Hard to easy	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Fair AI	0.18 [-0.56, 0.96]	-0.00 [-0.57, 0.57]	-0.30 [-0.99, 0.42]	0.29 [-0.47, 1.14]	0.18 [-0.84, 1.1]	-0.18 [-1.08, 0.66]	0.01 [-0.68, 0.7]	0.34 [-0.58, 1.31]	-0.06 [-1.3, 1.17]	<b>1.45<sup>\$</sup></b> [0.5, 2.46]	-0.58 [-1.89, 0.64]	
Advantaged	0.11 [-0.9, 1]	-0.30 [-0.99, 0.42]	0.24 [-0.47, 0.94]	0.29 [-1.78, 0.19]	-0.79 [-0.7, 0.97]	0.12 [-1.94, 0.61]	-0.18 [-1.94, 0.61]	-0.63 [-3.35, -0.89]	-0.06 [-1.3, 1.17]	-2.14 <sup>\$</sup> [-3.35, -0.89]	-0.58 [-1.89, 0.64]	
Disadvantaged	-0.46 [-1.44, 0.45]	0.24 [-0.47, 0.94]	0.08 <sup>\$</sup> [+0.00, 0.17]	0.08 [+0.00, 0.17]	0.15 <sup>\$</sup> [0.05, 0.25]	0.16 <sup>\$</sup> [0.06, 0.26]	0.18 <sup>\$</sup> [0.09, 0.28]	0.18 <sup>\$</sup> [0.08, 0.28]	0.34 <sup>\$</sup> [0.2, 0.49]	0.35 <sup>\$</sup> [0.21, 0.49]	0.31 <sup>\$</sup> [0.19, 0.44]	0.32 <sup>\$</sup> [0.19, 0.44]
Risk attitude	0.13 <sup>\$</sup> [0.04, 0.24]	0.14 <sup>\$</sup> [0.05, 0.24]	-0.03 [-0.16, 0.09]	-0.03 [-0.16, 0.09]	-0.13 <sup>\$</sup> [-0.28, -0.00]	-0.14 <sup>\$</sup> [-0.28, -0.01]	-0.04 [-0.19, 0.11]	-0.04 [-0.19, 0.11]	0.25 <sup>\$</sup> [0.05, 0.51]	0.25 <sup>\$</sup> [0.04, 0.51]	0.10 [-0.11, 0.31]	0.11 [-0.11, 0.31]
Fairness Sensitivity	-0.10 [-0.25, 0.03]	-0.10 [-0.26, 0.04]	-0.03 [-0.16, 0.09]	-0.03 [-0.16, 0.09]	-0.12 <sup>\$</sup> [-0.28, -0.00]	-0.14 <sup>\$</sup> [-0.28, -0.01]	-0.12 <sup>\$</sup> [-0.19, 0.11]	-0.12 <sup>\$</sup> [-0.19, 0.11]	-0.14 [-0.34, 0.06]	-0.15 [-0.35, 0.05]	-0.20 <sup>\$</sup> [-0.35, -0.03]	-0.20 <sup>\$</sup> [-0.36, -0.03]
Empathy	0.13 <sup>\$</sup> [+0.00, 0.25]	0.12 [-0.00, 0.24]	0.10 [-0.00, 0.22]	0.10 [-0.00, 0.22]	0.16 <sup>\$</sup> [0.03, 0.29]	0.14 <sup>\$</sup> [0.02, 0.28]	0.12 <sup>\$</sup> [+0.00, 0.27]	0.12 <sup>\$</sup> [+0.00, 0.27]	-0.14 [-0.34, 0.06]	-0.15 [-0.35, 0.05]	-0.20 <sup>\$</sup> [-0.35, -0.03]	-0.20 <sup>\$</sup> [-0.36, -0.03]
Initial Credit Score	-0.11 [-0.26, 0.04]	-0.11 [-0.27, 0.03]	-0.24 <sup>\$</sup> [-0.34, -0.13]	-0.24 <sup>\$</sup> [-0.34, -0.12]	0.16 <sup>\$</sup> [0.01, 0.32]	0.16 <sup>\$</sup> [0.01, 0.31]	0.25 <sup>\$</sup> [0.12, 0.38]	0.25 <sup>\$</sup> [0.12, 0.39]	0.13 [-0.08, 0.34]	0.13 [-0.08, 0.33]	0.24 <sup>\$</sup> [0.03, 0.41]	0.23 <sup>\$</sup> [0.03, 0.41]
Constant	3.48 <sup>\$</sup> [0.98, 5.89]	3.79 <sup>\$</sup> [1.3, 6.23]	2.95 <sup>\$</sup> [0.97, 4.91]	2.94 <sup>\$</sup> [0.99, 4.84]	4.19 <sup>\$</sup> [1.56, 6.74]	4.71 <sup>\$</sup> [2.1, 7.15]	2.78 <sup>\$</sup> [0.21, 5.06]	2.78 <sup>\$</sup> [0.31, 4.99]	14.64 <sup>\$</sup> [11.23, 18.23]	15.12 <sup>\$</sup> [11.64, 16.81]	16.23 <sup>\$</sup> [13.01, 19.36]	17.71 <sup>\$</sup> [14.41, 20.71]

Table 3: Regression models predicting decision subjects’ improvement, retention, and perceived fairness based on the AI model’s decision fairness in the two sub-experiments of Study 2. Coefficients and their 95% bootstrap confidence intervals are reported. A superscript <sup>\$</sup> indicates that the estimated coefficient is reliably different from zero. Reliable effects of the independent variables of interests are bolded.

“Hard to easy”: 385 subjects). Demographic details of these subjects are reported in Appendix D. Below, we analyzed the full dataset collected from these subjects to answer **RQ3** by re-examining **RQ1–RQ2** on this dataset using the same analysis methods as in Study 1.

**Re-examining RQ1: Impacts of AI’s decision fairness on engagement.** For the “Easy-to-hard” sub-experiment, on average, subjects made 4.86 ( $SD = 3.63$ ) improvement attempts and stayed for 7 ( $SD = 3.77$ ) rounds in the loan application tasks when they interacted with the fair AI model. Meanwhile, subjects in the unfair AI model treatment made an average of 4.70 ( $SD = 3.49$ ) improvement attempts and interacted with the AI model for 6.67 ( $SD = 3.68$ ) rounds. The regression results, as reported in Table 3 (Models 1, 2, 5, 6), suggest that in this sub-experiment, decision subjects’ engagement with the AI model was not reliably influenced by either the AI model’s decision fairness across groups or on the specific group that the subject belonged to. This is also true for subjects in the “Hard-to-easy” sub-experiment—Again, in the regression models learned for this sub-experiment (Models 3, 4, 7, 8 in Table 3), we did not detect any reliable effect of the AI model’s decision fairness on subjects’ willingness to improve their qualification or continue interacting with the model (see Appendix F and G for additional figures and statistics).

**Re-examining RQ2: Impacts of AI’s decision fairness on fairness perceptions.** Table 3 (Models 9–12) shows how the AI model’s decision fairness affects decision subjects’ perceived fairness of the AI model in the two sub-experiments of Study 2. We note that in the “Easy to hard” sub-experiment, subjects’ perceived fairness of the AI model is *not* reliably impacted by either the AI model’s group-level fairness or the AI model’s bias on the subject’s own group. In contrast, in the “Hard to easy” sub-experiment, we find that subjects’ perceived fairness of the AI model is affected by the AI model’s fairness properties (e.g., fair AI treatment:  $M = 19.37$ ,  $SD = 4.46$ ; unfair AI treatment, red group:  $M = 18.75$ ,  $SD = 5.34$ ; unfair AI treatment, blue group:  $M = 17.26$ ,  $SD = 5.75$ ). In particular, in this sub-experiment, decision subjects’ average perceived fairness of

the AI model increased when the AI model made fair decisions across different groups (i.e., the coefficient for “Fair AI” in Model 11 is  $\beta = 1.45[0.5, 2.46]$ ). In addition, those subjects who have been placed at the disadvantaged position by the AI model also perceived the model to be less fair (i.e., the coefficient for “Disadvantaged” in Model 12 is  $\beta = -2.14[-3.35, -0.89]$ ).

### Study 3

Finally, to answer **RQ4**, we conduct Study 3 where the AI model’s fairness is examined with respect to a protected attribute, gender, and subjects’ group identities in the study were determined by their self-reported, real-world gender.

### Experimental Design

In Study 3, we adopted the experimental design from Study 1 and made a few minor changes. First, instead of assigning a fictional group identity (red or blue) to each subject, we asked subjects to self-report their gender at the beginning of the experiment and used it as their group identity in the experiment. Second, in the “fair AI” treatment, the AI model employed provided equal decisions to male and female subjects in granting loans; the approval rate of this AI model to both male and female was determined by Table 1a. In contrast, in the “unfair AI” treatment, the AI model exhibited gender bias and favored male over female subjects in granting loans, and its approval rate for male and female was determined by Table 1b and Table 1c, respectively. This unfair AI model was designed to mirror the AI model’s gender biases against females in the real world that were increasingly revealed by researchers (Dastin 2018; Bolukbasi et al. 2016; Buolamwini and Gebru 2018; Nadeem, Marjanovic, and Abedin 2022). Finally, we also modified the flowcharts that were shown to subjects at the end of each round, so that they summarized the AI model’s decisions on simulated loan applicants who were grouped by their gender (instead of the red vs. blue group identities as used in Study 1).

The procedure of Study 3 was also largely the same as Study 1, except for a few minor changes: (1) Previous participants from Studies 1 and 2 were excluded. (2) The AI model, as discussed earlier, determined loan approvals based

on the subject’s self-reported gender and their credit level. (3) We also incorporated a deceptive component in the study to capture subjects’ genuine reactions towards a possibly biased AI model while avoiding actually paying female subjects systematically less in our experiment. In particular, at the beginning of the experiment, subjects were told that their bonus payment in the experiment was proportional to the final balance in their account, with every 500 coins translating to \$1.50. However, in reality, to prevent gender bias in study payments, each subject was given a fixed bonus of \$3.30. In other words, all subjects of Study 3 eventually received a total payment of \$5.30.

## Experimental Results

A total of 416 subjects participated in Study 3 and passed the filtering procedure. Among them, 55% self-identified as male, and 45% self-identified as female (see Appendix D for more demographics). In the following, we used the full dataset from these subjects to re-do the same kind of analysis that we conducted for **RQ1–RQ2**, in order to examine whether our results of Study 1 still hold true when the AI model’s fairness properties was examined with respect to a protected social attribute, i.e., gender (**RQ4**).

**Re-examining RQ1: Impacts of AI’s decision fairness on engagement.** Regarding subjects’ engagement with the AI model in this study, we found that male subjects made 4.28 ( $SD = 3.44$ ) improvement attempts and stayed for 6.83 ( $SD = 3.73$ ) rounds on average, while female subjects made 3.98 ( $SD = 3.17$ ) improvement attempts and stayed for 6.46 ( $SD = 3.82$ ) rounds. The regression results are presented in Table 4 (Models 1–4). Again, we find that, on average, decision subjects’ willingness to improve themselves and willingness to keep interacting with the AI model are not reliably affected by the AI model’s fairness level across groups or its decision fairness towards the subjects’ group. In other words, our Study 1 results still hold true even when the AI model’s fairness is examined with respect to salient protected attributes like gender.

**Re-examining RQ2: Impacts of AI’s decision fairness on fairness perceptions.** In Study 3, the average rating of the perceived fairness of the AI model was 18.90 ( $SD = 5.33$ ) for subjects in the fair AI model treatment, and 18.24 ( $SD = 5.75$ ) for subjects in the unfair AI model treatment. Within the unfair AI model treatment, the average perceived fairness rating of the AI model was 18.92 ( $SD = 5.64$ ) and 17.30 ( $SD = 5.79$ ) for the male and female subjects, respectively. The regression results, as presented in Table 4 (Models 5 and 6), again suggest that female subjects perceived the AI model as less fair if the AI model systematically biased against them. This is evidenced by the reliably negative coefficient estimated for the independent variable “*Disadvantaged*” in Model 6 (i.e.,  $\beta = -1.38[-2.65, -0.16]$ ). In contrast, male subjects who were being placed at an advantaged position by the unfair AI model did not show any decrease in their perceived fairness of the model, despite the AI model exhibited a clear gender bias.

	Improvement		Retention		Perceived Fairness	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fair AI	0.12 [-0.49, 0.78]		-0.45 [-1.17, 0.35]		0.45 [-0.52, 1.46]	
Advantaged		-0.04 [-0.83, 0.68]		0.64 [-0.24, 1.46]		0.23 [-1, 1.41]
Disadvantaged		-0.22 [-1.08, 0.5]		0.18 [-0.8, 1.08]		<b>-1.38<sup>\$</sup></b> <b>-2.65, -0.16</b>
Risk attitude	0.09 <sup>\$</sup> [0.01, 0.17]	0.09 <sup>\$</sup> [0.01, 0.17]	0.13 <sup>\$</sup> [0.04, 0.22]	0.13 <sup>\$</sup> [0.04, 0.22]	0.41 <sup>\$</sup> [0.28, 0.54]	0.41 <sup>\$</sup> [0.27, 0.54]
Fairness Sensitivity	0.04 [-0.1, 0.17]	0.04 [-0.1, 0.17]	0.02 [-0.14, 0.17]	0.02 [-0.14, 0.17]	-0.14 [-0.38, 0.12]	-0.13 [-0.37, 0.13]
Empathy	0.05 [-0.05, 0.16]	0.05 [-0.05, 0.16]	0.01 [-0.11, 0.13]	0.01 [-0.11, 0.13]	-0.01 [-0.19, 0.16]	-0.01 [-0.18, 0.16]
Initial Credit Score	-0.23 <sup>\$</sup> [-0.35, -0.11]	-0.23 <sup>\$</sup> [-0.35, -0.11]	0.11 [-0.03, 0.26]	0.12 [-0.03, 0.26]	0.36 <sup>\$</sup> [0.17, 0.56]	0.37 <sup>\$</sup> [0.18, 0.56]
Constant	3.14 <sup>\$</sup> [0.9, 5.22]	3.23 <sup>\$</sup> [1.06, 5.33]	5.04 <sup>\$</sup> [2.62, 7.55]	4.53 <sup>\$</sup> [2.17, 7.06]	15.21 <sup>\$</sup> [11.43, 19.09]	15.44 <sup>\$</sup> [11.73, 19.28]

Table 4: Regression models predicting decision subjects’ improvement, retention, and perceived fairness based on the AI model’s decision fairness for Study 3. Coefficients and their 95% bootstrap confidence intervals are reported. A superscript <sup>\$</sup> indicates that the estimated coefficient is reliably different from zero. Reliable effects of the independent variables of interests are bolded.

## Conclusions and Discussions

In this paper, we conducted exploratory randomized human-subject experiments to investigate how an AI model’s decision fairness affect decision subjects’ engagement with and fairness perceptions of the model when they could repeatedly and strategically respond to the AI model’s decisions on them. A key finding of our study is that when decision subjects can repeatedly and strategically respond to AI decisions, their engagement with the AI model does *not* reflect either the AI model’s decision fairness—as decided by various fairness definitions—or their perceived fairness of the AI model. In this section, we reflect on our findings, and address the limitations and future work of our study.

**Similarity and difference of our results with earlier findings.** In this study, we find that when decision subjects can repeatedly and strategically respond to AI decisions, their perceived fairness of the model is still affected by the model’s biases against their *own* group. This is largely consistent with previous findings (Wang, Harper, and Zhu 2020; Gemalmaz and Yin 2022), which suggests that decision subjects have a degree of “outcome favorability bias” in their fairness perceptions of AI models (Wang, Harper, and Zhu 2020). This also indicates that social comparison across groups is still a key factor that contributes to decision subjects’ perceived fairness of an AI model, despite the possibility to improve one’s qualification may have shifted some of their attention towards temporal self-comparison. Indeed, in subjects’ open-ended text responses in the exit survey explaining why they perceived the AI model as fair or unfair, we find evidence suggesting decision subjects’ perceived fairness of the AI model may be affected by both social comparison (e.g., “*people in red and blue groups with the exact same credit score ranges had different chances of getting approved*”) and temporal self-comparison (“*when my credit score improved, I kept getting approved for loans*”). See Appendix I for more detailed analysis.

On the other hand, in terms of decision subjects’ engagement with the AI model, our study shows that their willingness to keep interacting with the model is not affected by the AI model’s biases towards/against their own group. This

is different from findings in prior work (Gemalmaz and Yin 2022) when the only strategic action decision subjects could take is to stop interacting with the AI model. We speculate that this is because in our study, the qualification improvement opportunities may have led decision subjects to prioritize “utility-related considerations” over “fairness-related considerations” when deciding how to engage with the AI model. Indeed, for those decision subjects whose group is disfavored by the AI model, continuing to improve themselves and interact with the AI model but receiving less favorable decisions from AI is certainly not ideal—as reflected in their perceived fairness of the AI model—but from a pure utility point of view, it may be better than simply “boycotting” the AI model and leaving the system with nothing.

**Explaining findings of Study 2 and 3.** In Study 2, we find that the AI model’s decision fairness significantly affected subjects’ perceived fairness of AI only in the “Hard to easy” sub-experiment, but not in the “Easy to hard” sub-experiment. We speculate that this is because subjects’ perception of an AI model’s fairness is influenced by both the model’s decisions on their present-self, and the model’s anticipated decisions on their ideal future-self after they improve their qualification. For subjects with high initial qualifications, their primary focus might be on their present-self (as they have limited room for further improvement), hence their perceived fairness of AI might not be significantly affected by the difficulty of qualification improvement. However, for subjects with low initial qualifications, the emphasis they put on their present-self and future-self might differ based on their belief in whether they could successfully achieve their ideal future-self. For example, in the “Easy to hard” sub-experiment, subjects with lower qualifications might be more optimistic about reaching their ideal future-self, hence they may focus more on temporal self-comparison when assessing the AI model’s fairness. In contrast, in the “Hard to easy” sub-experiment, subjects with lower qualifications found it difficult to improve their qualification. Thus, they might focus more on their present-self, engaging in social comparisons, and have a heightened sensitivity for the AI model’s biases.

Moreover, in Study 3, we found that decision subjects’ engagement with AI was not affected by the AI model’s decision fairness even when fairness is examined on subjects’ real-world gender groups. Prior research has pointed out the frequent challenges women encounter due to societal biases in the real world (Dastin 2018; Moss-Racusin et al. 2012; Elsesser 2016; Lee 2015). Thus, we speculate that female subjects in our study might find the gender biases exhibited by the AI model in our experiment to be “familiar”. Thus, they may have decided to adapt to such bias based on their past experiences of dealing with real-world bias in order to maximize the utility they may obtain from the model, especially given that they see little possibility of changing how the model works by disengaging with the unfair AI model. Meanwhile, we also found that male subjects in Study 3 did not consider the AI model as less fair when they were placed at the advantaged position by an AI model with gender bias. This again suggests that male subjects may have suffered from the outcome favorability bias when forming their fair-

ness perceptions of an AI model (Wang, Harper, and Zhu 2020), and highlights the importance of identifying a representative and diverse population of decision subjects when evaluating the perceived fairness of AI models.

**Implications of our findings.** Our findings imply that when decision subjects can repeatedly and strategically respond to an AI model, the equality in engagement across different groups of decision subjects should *not* be used as a proxy indicator of a fair AI model. Instead, responsible AI practitioners should delve deeper into truly understanding decision subjects’ perceptions of and satisfaction with the AI model to develop fair AI-based decision systems. On the other hand, compared to previous findings (Gemalmaz and Yin 2022; Wang, Harper, and Zhu 2020), results of our study highlight that providing decision subjects with opportunities to improve their qualification can effectively maintain user retention in the case that the AI model exhibits a degree of unfairness across groups, especially for subjects whose group is placed at the disadvantaged position by the AI model. This means that when fairness concerns have been identified in AI-based decision systems, actively providing decision subjects with information and guidance on qualification improvement (e.g., via providing algorithmic recourse plans) might be used as a temporary solution to maintain user retention. This may buy the system developers some time to address the fairness concerns of the AI models without losing a significant sector of users. However, we emphasize that this short-term relief should not replace more fundamental and comprehensive solutions which address the root cause fairness problems of the AI model and truly improve decision subjects’ satisfactions with the model. In fact, in a real-world environment with competitions (e.g., multiple banks providing loan lending services), it may be users’ fairness perceptions and satisfaction that play the key role in shaping their long-term retention.

**Limitations and future work.** We acknowledge that our findings were constrained by our study setup. For example, the ways that subjects were incentivized in our study, albeit designed to reflect the uncertainty, gains, and losses that they would encounter in the real-world loan lending scenarios, may have nudged our subjects into more rational thinking and made them put a higher weight on utility-related considerations when determining how to engage with the AI model. The subject population that we engaged through the online platform (i.e., MTurk) may also lean more towards rational thinking, as one of their primary goals is to maximize their earnings. Our study was also constrained by a few “parameters” of our experimental tasks (e.g., cost for qualification improvement attempt, reward for favorable decision). Future research could systematically vary these “parameters” and investigate how they affect decision subjects’ behavior and perceptions. Additionally, the AI model used in our experiment is not updating over time as new data gets generated from decision subjects’ interactions with the model. Empirically investigating the feedback loop between AI model updates and user reactions over time could be another critical direction to explore in the future for advancing our understanding of the long-term impact of AI fairness.

## References

- Albert, S. 1977. Temporal comparison theory. *Psychological review*, 84(6): 485.
- Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520.
- Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.
- Cumming, G. 2014. The new statistics: Why and how. *Psychological science*, 25(1): 7–29.
- D’Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 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.
- Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women.
- Dragicevic, P. 2016. Fair statistical communication in HCI. *Modern statistical methods for HCI*, 291–330.
- Elsesser, K. M. 2016. Gender bias against female leaders: A review. *Handbook on well-being of working women*, 161–173.
- Faul, F.; Erdfelder, E.; Lang, A.-G.; and Buchner, A. 2007. G\* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2): 175–191.
- Festinger, L. 1954. A theory of social comparison processes. *Human relations*, 7(2): 117–140.
- Gemalmaz, M. A.; and Yin, M. 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, 295–306. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.
- Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B. 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.
- Heidari, H.; Nanda, V.; and Gummadi, K. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *International Conference on Machine Learning*, 2692–2701. PMLR.
- Hershfield, H. E. 2011. Future self-continuity: How conceptions of the future self transform intertemporal choice. *Annals of the New York Academy of Sciences*, 1235(1): 30–43.
- Hu, L.; and Chen, Y. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, 1389–1398.
- Kam, C. 2012. Risk Attitudes and Political Participation. *American Journal of Political Science*, 56.
- Lee, C. 2015. Gender bias in the courtroom: Combating implicit bias against women trial attorneys and litigators. *Cardozo JL & Gender*, 22: 229.
- Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.
- Liu, L. T.; Wilson, A.; Haghtalab, N.; Kalai, A. T.; Borgs, C.; and Chayes, J. 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.
- Mesa, N. 2021. Can the criminal justice system’s artificial intelligence ever be truly fair?
- Moss-Racusin, C. A.; Dovidio, J. F.; Brescoll, V. L.; Graham, M. J.; and Handelsman, J. 2012. Science faculty’s subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41): 16474–16479.
- Mouzannar, H.; Ohannessian, M. I.; and Srebro, N. 2019. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 359–368.
- Nadeem, A.; Marjanovic, O.; and Abedin, B. 2022. Gender bias in AI-based decision-making systems: a systematic literature review. *Australasian Journal of Information Systems*, 26.
- Nakao, Y.; Stumpf, S.; Ahmed, S.; Naseer, A.; and Strapelli, L. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3): 1–30.
- Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 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*, AIES ’19, 99–106. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Shen, H.; Jin, H.; Cabrera, Á. A.; Perer, A.; Zhu, H.; and Hong, J. I. 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): 1–22.
- Spreng, R. N.; McKinnon, M. C.; Mar, R. A.; and Levine, B. 2009. The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, 91(1): 62–71.
- Srivastava, M.; Heidari, H.; and Krause, A. 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.
- Thorstad, R.; and Wolff, P. 2018. A big data analysis of the relationship between future thinking and decision-making.

*Proceedings of the National Academy of Sciences*, 115(8): E1740–E1748.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 10–19.

Wang, R.; Harper, F. M.; and Zhu, H. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *CoRR*, abs/2001.09604.

Yaghini, M.; Krause, A.; and Heidari, H. 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.

Yurrita, M.; Draws, T.; Balayn, A.; Murray-Rust, D.; Tintarev, N.; and Bozzon, A. 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*, 1–21.

Zhang, X.; Khalilgarekani, M.; Tekin, C.; 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.

Zhang, X.; and Liu, M. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, 525–555. Springer.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.

# Appendices

## A Example Flowchart

Figure A1 shows an example flowchart that participants saw in our experiment, which illustrates the AI model's decisions on loan applicants of different groups in one round.

## B A Complete List of Questions Used in the Post-Experiment Survey

In the post-experiment survey, the questions listed below were used. Please note that for any negative statements, we reversed the rating when calculating the scores for specific measurements, such as risk attitude.

### 1. DEMOGRAPHICS QUESTIONS

- **What is your age?**

- (a) 18-24 years old
- (b) 25-34 years old
- (c) 35-44 years old
- (d) 45-54 years old
- (e) 55-64 years old
- (f) 65-74 years old
- (g) 75 years or older

- **Which race or ethnicity best describes you? (Please choose only one.)**

- (a) American Indian or Alaskan Native
- (b) Asian / Pacific Islander
- (c) Black or African American
- (d) Hispanic
- (e) White / Caucasian
- (f) Multiple ethnicity/ Other

- **Please select the state you are in:**

- (a) Alabama
- (b) Alaska
- (c) Arizona
- (d) ...

- **In general, would you describe your political party of affiliation as:**

- (a) Democrat
- (b) Republican
- (c) Independent

- **In general, would you describe your political view as:**

- (a) Very liberal
- (b) Liberal
- (c) Somewhat liberal
- (d) Moderate
- (e) Somewhat conservative
- (f) Conservative
- (g) Very conservative

- **What is the highest degree or level of school you have completed (if currently enrolled, highest degree received)?**

- (a) No schooling completed

- (b) Nursery school to 8th grade
- (c) Some high school, no diploma
- (d) High school graduate, diploma or the equivalent (for example: GED)
- (e) Some college credit, no degree
- (f) Trade/technical/vocational training
- (g) Associate degree
- (h) Bachelor's degree
- (i) Master's degree
- (j) Professional degree
- (k) Doctorate degree

### 2. Sensitivity to Fairness Questions

For the rest of the statements below about you (until the end of the survey), please indicate how much you agree with it.

- **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). (Positive)**

- (a) Strongly disagree
- (b) Disagree
- (c) Neutral
- (d) Agree
- (e) Strongly agree

- **I would only use an AI system if it is fair to everyone. (Positive)**

- (a) Strongly disagree
- (b) Disagree
- (c) Neutral
- (d) Agree
- (e) Strongly agree

- **I would stop using an AI system if it is unfair, even if it tends to be in favor of me. (Positive)**

- (a) Strongly disagree
- (b) Disagree
- (c) Neutral
- (d) Agree
- (e) Strongly agree

- **When I decide whether to use an AI system or not, I seldom think about whether the system is fair. (Negative)**

- (a) Strongly disagree
- (b) Disagree
- (c) Neutral
- (d) Agree
- (e) Strongly agree

### 3. Risk Attitude Questions

- **If I were betting on horses and were a big winner in the third or fourth race, I would be more likely to stop playing and take my winnings. (Negative)**

- (a) Strongly disagree
- (b) Disagree
- (c) Neutral
- (d) Agree

AI system's decisions on all 12105 applicants in this round:

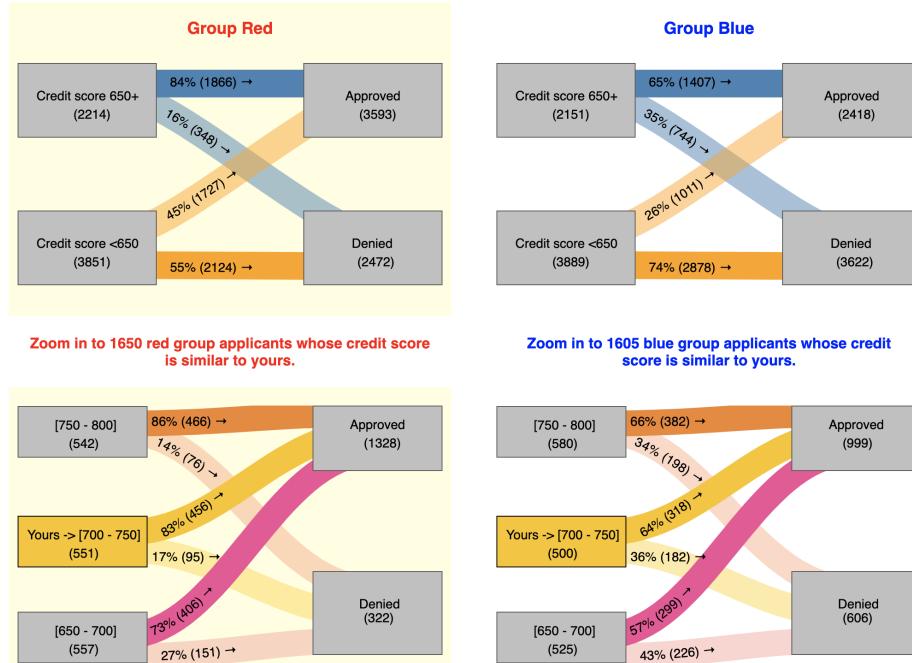


Figure A1: An example of the flowchart that subjects in the unfair AI treatment saw in the experiment, which summarizes the AI model's decisions on different groups of applicants in the past round. Subjects could see the frequency of the AI model approving/denying loans both for applicants with/without "high" credit scores (i.e., a score of at least 650), and for applicants with similar credit scores as themselves (i.e., applicants with the same credit scores as themselves or one level above/below themselves).

- (e) Strongly agree
  - **I like to do frightening things. (Positive)**
    - (a) Strongly disagree
    - (b) Disagree
    - (c) Neutral
    - (d) Agree
    - (e) Strongly agree
  - **I like new and exciting experiences, even if I have to break the rules. (Positive)**
    - (a) Strongly disagree
    - (b) Disagree
    - (c) Neutral
    - (d) Agree
    - (e) Strongly agree
  - **I prefer friends who are exciting and unpredictable. (Positive)**
    - (a) Strongly disagree
    - (b) Disagree
    - (c) Neutral
    - (d) Agree
    - (e) Strongly agree
  - **In general, it is very easy for me to accept taking risks. (Positive)**
    - (a) Strongly disagree
    - (b) Disagree

- (c) Neutral
- (d) Agree
- (e) Strongly agree

#### 4. Empathy Questions

- **I remain unaffected when someone close to me is happy. (Negative)**
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- **I enjoy making other people feel better. (Positive)**
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- **I get a strong urge to help when I see someone who is upset. (Positive)**
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree

- When I see someone being treated unfairly, I do not feel very much pity for them. (Negative)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- When I see someone being taken advantage of, I feel kind of protective towards them. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree

## 5. Perceptions of AI Fairness Questions

- The bank's AI system is fair. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- The bank's AI system is fair to loan applicants. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- The bank's AI system is fair to manage loan applications. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- The decisions that the bank makes as a result of the AI system will be fair. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- The bank's AI system will lead the bank to make great loan lending decisions. (Positive)
  - (a) Strongly disagree
  - (b) Disagree
  - (c) Neutral
  - (d) Agree
  - (e) Strongly agree
- The bank's AI system will make mistakes. (Negative)
  - (a) Strongly disagree
  - (b) Disagree

- (c) Neutral
- (d) Agree
- (e) Strongly agree

- **Free-form text:** Based on your responses to the survey questions above, it seems like that you feel that in this game, the bank's AI-powered system is X to loan applicants in general. If this is correct, please briefly explain why you believe so; if this is incorrect, please briefly describe what do you think of the bank's AI system in terms of its level of fairness. (Add up the ratings for the "Perceptions of AI Fairness Questions." If the total is less than 15, replace X with "unfair." If the total is equal to or greater than 15, replace X with "fair.")

## C Filtering Procedures Used in the Experiment

To ensure that our experimental data was provided by genuine human subjects rather than bots or spammers, we implemented a few protective procedures. First, we incorporated both Google's reCAPTCHA v3<sup>8</sup> and a honeypot CAPTCHA (i.e., a CAPTCHA that is hidden in the HTML, thus invisible to real human subjects but visible to bots) in the web application of our experiment to filter out bots. Second, we included an attention check question in the post-experiment survey, which instructed the subject to select a pre-defined option, to filter out inattentive subjects. Finally, we manually checked the subject's responses to open-ended questions in the survey, and filtered out potential spammers (e.g., subjects who provided identical responses or responses that were not comprehensible). A subject's data was only considered valid if it can pass all these three filtering procedures.

## D Participant Demographics

%	Study 1	Study 2 (Easy to hard)	Study 2 (Hard to easy)	Study 3
<b>Gender</b>				
Male	54%	51%	57%	55%
Female	45%	48%	42%	45%
Other	1%	1%	1%	0%
<b>Age</b>				
18–24	17%	12%	15%	15%
25–34	29%	26%	27%	26%
35–44	24%	26%	25%	25%
45+	30%	36%	33%	34%
<b>Race</b>				
White	76%	80%	77%	82%
Black	7%	6%	8%	6%
Asian	8%	6%	6%	6%
Hispanic	3%	3%	4%	3%
Other	6%	5%	5%	3%

Table A1: Demographics of the subjects in each experiment. The total number of subjects in each experiment is: Study 1: 368 participants, Study 2 (Easy to hard): 328 participants, Study 2 (Hard to easy): 385 participants, and Study 3: 416 participants.

<sup>8</sup><https://www.google.com/recaptcha/about/>

## E Study 1: Data Details

Variable	Unfair Red (88)			Unfair Blue (92)			Fair Red (92)			Fair Blue (96)		
	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Improvement	4.60	3.41	[3.88, 5.32]	3.63	3.04	[3.00, 4.26]	3.79	3.41	[3.09, 4.50]	4.41	3.44	[3.71, 5.10]
Retention	6.91	3.78	[6.11, 7.71]	6.26	3.66	[5.50, 7.02]	6.04	3.91	[5.23, 6.85]	6.41	3.95	[5.61, 7.21]
Fairness Perception	19.0	4.75	[18.0, 20.0]	17.0	5.81	[15.84, 18.2]	18.7	5.45	[17.6, 19.9]	18.8	4.87	[17.8, 19.8]

Table A2: Descriptive Statistics of Subjects by Group (Study 1)

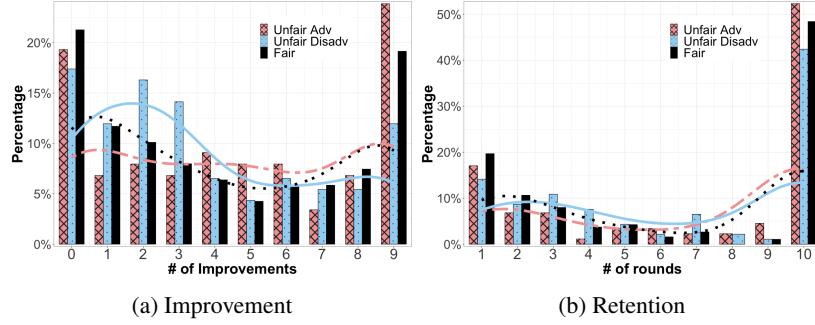


Figure A2: Distributions of (a) the number of improvement attempts that subjects made, and (b) the number of rounds that subjects interacted with the AI model, for subjects who were assigned to the fair AI model treatment, the red (and advantaged) group of the unfair AI model treatment, and the blue (and disadvantaged) group of the unfair AI model treatment, in Study 1. Curves represent the probability density functions obtained through kernel density estimation.

## F Study 2: Data Details (Easy to hard)

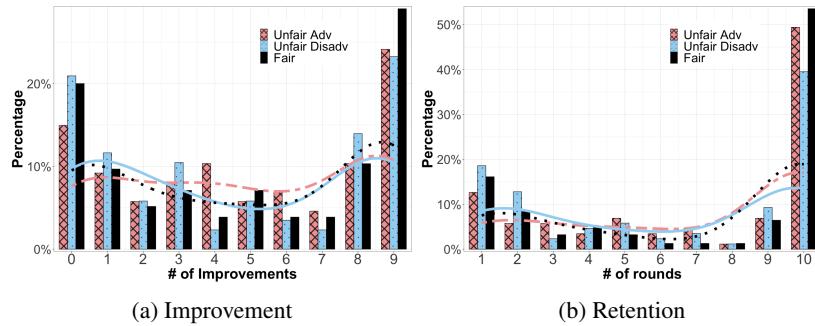


Figure A3: Distributions of (a) the number of improvement attempts that subjects made, and (b) the number of rounds that subjects interacted with the AI model, for subjects who were assigned to the fair AI model treatment, the red (and advantaged) group of the unfair AI model treatment, and the blue (and disadvantaged) group of the unfair AI model treatment, in Study 2 sub-experiment ‘Easy to hard’. Curves represent the probability density functions obtained through kernel density estimation.

Variable	Unfair Red (87)			Unfair Blue (86)			Fair Red (78)			Fair Blue (77)		
	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Improvement	4.89	3.36	[4.17, 5.60]	4.51	3.62	[3.74, 5.29]	5.19	3.54	[4.39, 5.99]	4.53	3.71	[3.69, 5.37]
Retention	7.08	3.52	[6.33, 7.83]	6.26	3.81	[5.44, 7.07]	7.36	3.58	[6.55, 8.17]	6.64	3.94	[5.74, 7.53]
Fairness Perception	18.6	5.25	[17.5, 19.7]	18.9	4.99	[17.8, 20.0]	18.6	4.94	[17.5, 19.7]	19.5	4.45	[18.5, 20.5]

Table A3: Descriptive Statistics of Subjects by Group (Study 2: Easy to hard)

## G Study 2: Data Details (Hard to easy)

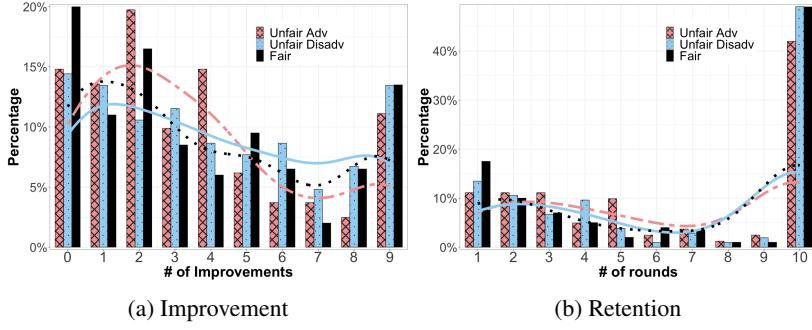


Figure A4: Distributions of (a) the number of improvement attempts that subjects made, and (b) the number of rounds that subjects interacted with the AI model, for subjects who were assigned to the fair AI model treatment, the red (and advantaged) group of the unfair AI model treatment, and the blue (and disadvantaged) group of the unfair AI model treatment, in Study 2 sub-experiment “Hard to easy”. Curves represent the probability density functions obtained through kernel density estimation.

Variable	Unfair Red (81)			Unfair Blue (104)			Fair Red (96)			Fair Blue (104)		
	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Improvement	3.41	2.82	[2.78, 4.03]	4.03	3.07	[3.43, 4.63]	3.26	3.07	[2.64, 3.88]	4.06	3.16	[3.44, 4.67]
Retention	6.28	3.60	[5.49, 7.08]	6.54	3.74	[5.81, 7.27]	6.30	3.92	[5.51, 7.10]	6.57	3.77	[5.83, 7.30]
Fairness Perception	18.8	5.34	[17.6, 19.9]	17.3	5.75	[16.1, 18.4]	19.8	4.02	[19.0, 20.7]	18.9	4.81	[18.0, 19.9]

Table A4: Descriptive Statistics of Subjects by Group (Study 2: Hard to easy)

## H Study 3: Data Details

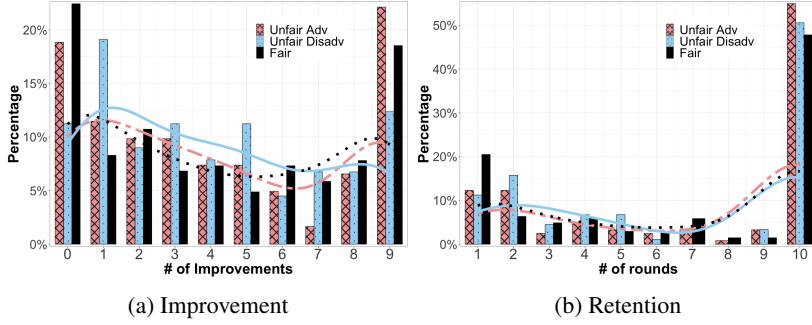


Figure A5: Distributions of (a) the number of improvement attempts that subjects made, and (b) the number of rounds that subjects interacted with the AI model, for subjects who were assigned to the fair AI model treatment, the red (and advantaged) group of the unfair AI model treatment, and the blue (and disadvantaged) group of the unfair AI model treatment, in Study 3. Curves represent the probability density functions obtained through kernel density estimation.

Variable	Unfair Male (122)			Unfair Female (89)			Fair Male (107)			Fair Female (98)		
	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Improvement	4.20	3.41	[3.59, 4.81]	3.98	3.03	[3.34, 4.62]	4.36	3.50	[3.69, 5.04]	3.98	3.31	[3.32, 4.64]
Retention	7.03	3.70	[6.37, 7.70]	6.60	3.78	[5.80, 7.39]	6.60	3.77	[5.87, 7.32]	6.34	3.88	[5.56, 7.12]
Fairness Perception	18.9	5.64	[17.9, 19.9]	17.3	5.79	[16.1, 18.5]	19.4	4.85	[18.5, 20.3]	18.4	5.79	[17.2, 19.5]

Table A5: Descriptive Statistics of Subjects by Group (Study 3)

## I Factors explaining why decision subjects might perceive an AI-based decision system as fair or unfair

To explore factors that drive decision subjects' perceived fairness of an AI model when they can strategically and repeatedly respond to its decisions, we looked into subjects' open-ended text responses in the exit survey, explaining why they perceived the AI model as fair/unfair. Among these responses, we identified evidence suggesting decision subjects' perceived fairness of the AI was influenced by *social comparisons*. For example, a subject in the unfair AI treatment believed the AI model as unfair because "*people in red and blue groups with the exact same credit score ranges had different chances of getting approved*". Meanwhile, subjects' fairness perceptions of the AI model could also be influenced by *temporal self-comparison*. As an example, one subject explained their rationale for perceiving the AI model as fair by stating "*when my credit score improved, I kept getting approved for loans*".

For many subjects, *meritocracy* was a key principle for them to evaluate the AI model's fairness: Observing people with higher credit scores have higher chances of being approved for their loans increased their perceived fairness of the AI model; however, observing people with similar credit scores get different outcomes or some highly-qualified people get loans denied while some lowly-qualified people get loans approved made them feel that the AI model was inconsistent and unfair. On the other hand, some subjects believed the key principle for fairness should be "*prioritize the needed*", and they criticized the AI model in the experiment as unfair because "*the AI is geared towards giving approvals to customers who generally don't need loans in the first place (highest credit score demographic)*". Decision subjects' judgement of the AI's fairness was also influenced by its *transparency*. For example, some subject found the AI model to be unfair because "*I was rejected and there was no information given as to why*", while others commented that they can not definitely assess whether the AI model is fair or unfair because they "*don't understand the internal workings of the AI system well enough*".

Interestingly, there also exist some influencing factors of subjects' perceived fairness of AI wherein different individuals may hold different opinions. One such factor is the *feature* that the AI model uses to make its decisions. For example, some subjects perceived the AI model in the experiment as fair because they believe "*credit scores give an accurate assessment of the likelihood that someone will pay their loans back and it's fair to use those scores when determining credit worthiness*". However, some other subjects considered the AI model as unfair because they thought AI's focus on credit scores implied it failed to "*take many more factors into account*" and credit score itself was "*an inherently unfair system*". Similarly, the fact that AI is pre-programmed is interpreted by different subjects differently—some considered this as an indicator of fairness because "*emotion and bias appears to be left out of AI decision making*", while others believed it leads to unfair decisions because it means "*AI is not capable of compassion*" and "*treating people as*

*numbers is not treating them fairly*".

## J Simulation for Parameter Estimation

To determine the ideal cost of qualification improvement to be used in our experiment, we conduct a simulation by modeling subjects' decision-making process in our experiment as a Markov Decision Process (MDP). Specifically, in our MDP, we identify the following key components:

- **States ( $s$ ):** There are 13 possible states in total—12 “regular” states ( $s_1, s_2, \dots, s_{12}$ ), each corresponding to one credit score level (e.g.,  $s_1$  corresponds to 300–350), and an additional terminal state ( $s_0$ ) representing an exit from the game. At any time  $t$ , the subject is in one of these 13 states.
- **Actions ( $a$ ):** In a regular state, there are three actions that a subject could take:
  - $A1$ : Apply and attempt to improve credit score. (Not available for the highest credit level state  $s_{12}$ )
  - $A2$ : Apply without attempting to improve credit score.
  - $A3$ : Do not apply and exit the game.
- When the subject is in the terminal state ( $s_0$ ), no action is available to them.
- **Transition Probabilities ( $T(s, a, s')$ ):**
  - When  $a = A1$ ,  $T(s, a, s') = P(s'|s, a) = 1 - p_s$  if  $s' = s$  and  $T(s, a, s') = p_s$  if  $s'$  is the level above  $s$ ;  $T(s, a, s') = 0$  for all other  $s'$ . This is to reflect that the success rate for a subject of credit level  $s$  to progress to the next credit level is  $p_s$  (e.g.,  $p_s = 0.44$  for all  $s$  in Study 1 and 3).
  - When  $a = A2$ ,  $T(s, a, s') = P(s'|s, a) = 1$  if  $s' = s$ ;  $T(s, a, s') = 0$  for all other  $s'$ .
  - When  $a = A3$ ,  $T(s, a, s') = P(s'|s, a) = 1$  if  $s' = s_0$ ;  $T(s, a, s') = 0$  for all other  $s'$ .
- **Reward Function ( $R(s, a, s')$ ):**
  - For  $a = A1$ ,  $R(s, a, s') = q_{s'} \cdot 50 - (1 - q_{s'}) \cdot 50 - x$  when  $s' = s$  or  $s'$  is the level above  $s$  ( $q_{s'}$  is the probability of getting loan approval when the credit score is at level  $s'$ ,  $x$  is the cost for improvement). This is to reflect that when the loan is approved, the subject's net profit is 50 (earn 100 coins but paid 50 coins for application), but when the loan is rejected, the subject's net profit is -50 (paid 50 coins for application). Regardless of whether the loan gets approved,  $x$  coins have been paid for qualification improvement.
  - For  $a = A2$ ,  $R(s, a, s') = q_s \cdot 50 - (1 - q_s) \cdot 50$  (since  $s' = s$  for certain).
  - For  $a = A3$ ,  $R(s, a, s') = 0$  (since  $s' = s_0$  for certain).

We considered a set of candidate cost of qualification improvement, i.e.,  $x \in \{1, 5, 10, 20\}$ . For each cost level, we used the value iteration algorithm to solve the corresponding MDP to determine the optimal policy for decision subjects who were interacted with the fair AI model, decision subjects who were interacted with the unfair AI model and were favored by AI, and decision subjects who were interacted

with the unfair AI model and were disfavored by AI. Then, to understand the average engagement strategy that a decision subject with an initial credit level of  $s \in \{s_1, \dots, s_9\}$  would take, we simulated  $N = 1000$  decision subjects who started from  $s$ , and took actions based on the optimal policy for a maximum of 9 rounds (as subjects in our experiments were required to apply for a loan in their first round of interaction with AI, and then can decide to continue interacting with the AI model for up to 9 more rounds). We then averaged across these 1000 simulated subjects to compute the expected number of qualification improvements and the expected number of rounds they would apply for loans for a subject starting from state  $s$ , and we visualized how these expectations change with the subject's initial credit level. Moreover, we repeated this simulation for three different qualification improvement schemes: the constant scheme used in Study 1 and 3, the easy-to-hard scheme used in the "Easy to hard" sub-experiment in Study 2, and the hard-to-easy scheme used in the "Hard to easy" sub-experiment in Study 2.

Figure A6–A9 show our simulation results for  $x = 1, 5, 10, 20$ , respectively. As we can see, when the cost of qualification improvement is very low (e.g.,  $x = 1$ ), subjects assigned to the advantaged group in the unfair AI treatment (constant scheme) almost always attempt to improve their qualification in every round (except for subjects who start with the lowest two credit levels, who will directly exit the game). On the other hand, when the cost of qualification improvement is relatively high (e.g.,  $x = 10$  or  $x = 20$ ), subjects assigned to the disadvantaged group in the unfair AI treatment (constant scheme) never attempt to improve. However, when we choose an intermediate level of cost (e.g.,  $x = 5$ ), the number of times subjects would engage with qualification improvement is more responsive to their initial credit level when subjects were placed at an advantaged position by the AI model. Meanwhile, when subjects were placed at a disadvantaged position by the AI model, subjects from more initial credit levels were willing to make improvement attempts. Based on these considerations, we chose 5 coins as the qualification improvement cost in our experiment eventually.

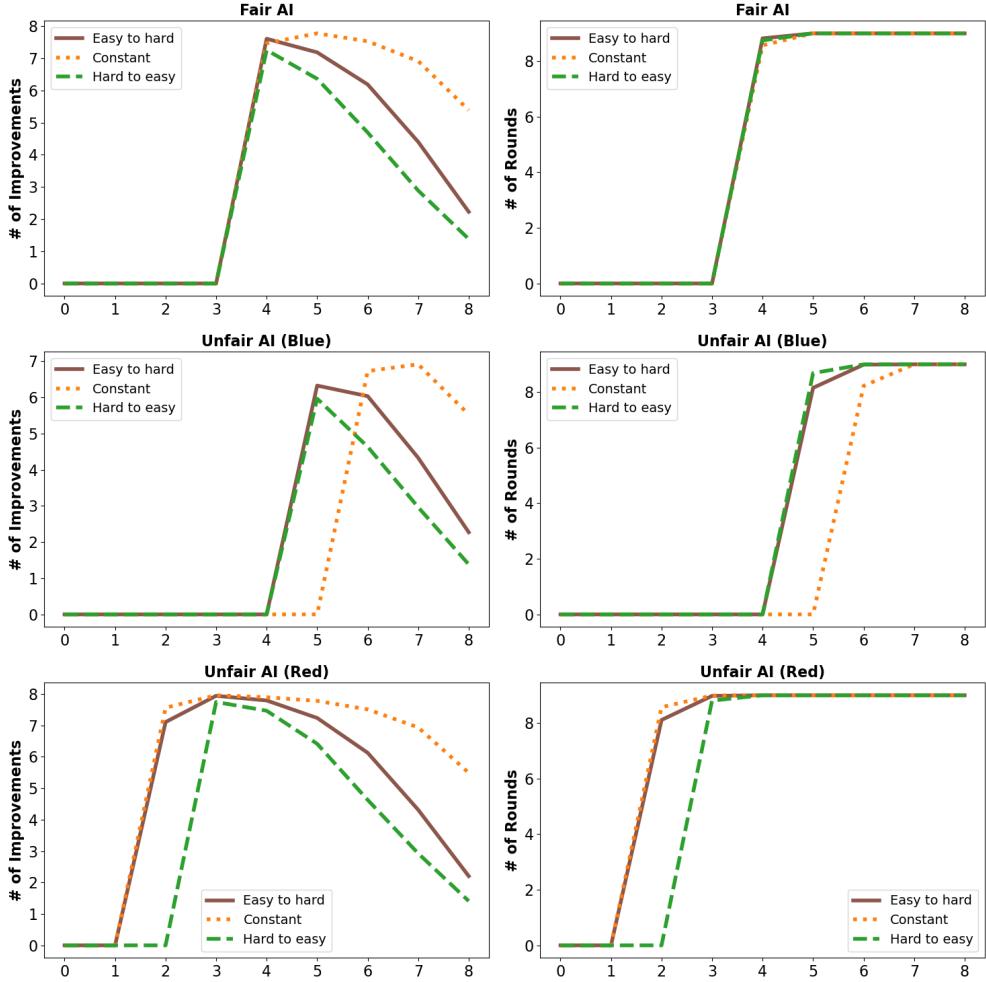


Figure A6: This figure presents the outcomes of implementing a 1-coin improvement cost, considering both Fair and Unfair AI treatments across three studies. The x-axis represents the initial credit score ranges, categorized from 0 to 8, with higher scores indicating higher ranges. In the Fair AI scenario (first row), the figure displays two graphs: the left graph indicates the expected number of improvements, while the right graph shows the expected number of rounds for subjects to keep interacting with the AI model. These are given within each credit score range and for each improvement difficulty used across our three studies. The middle and third rows represent the Unfair AI treatment for blue/female (disadvantaged) and red/male (advantaged) group subjects, similarly displaying “number of times to improve” and “number of rounds to apply for a loan” strategies as in the first row. The various improvement difficulties are color-coded: constant difficulty used in Studies 1 and 3 (orange), hard to easy in Study 2’s sub-experiment (green), and easy to hard in Study 2’s other sub-experiment (brown), illustrating different expected strategies for each improvement difficulty across credit score ranges. A low improvement cost might cause most decision subjects to improve since it is easily accessible.

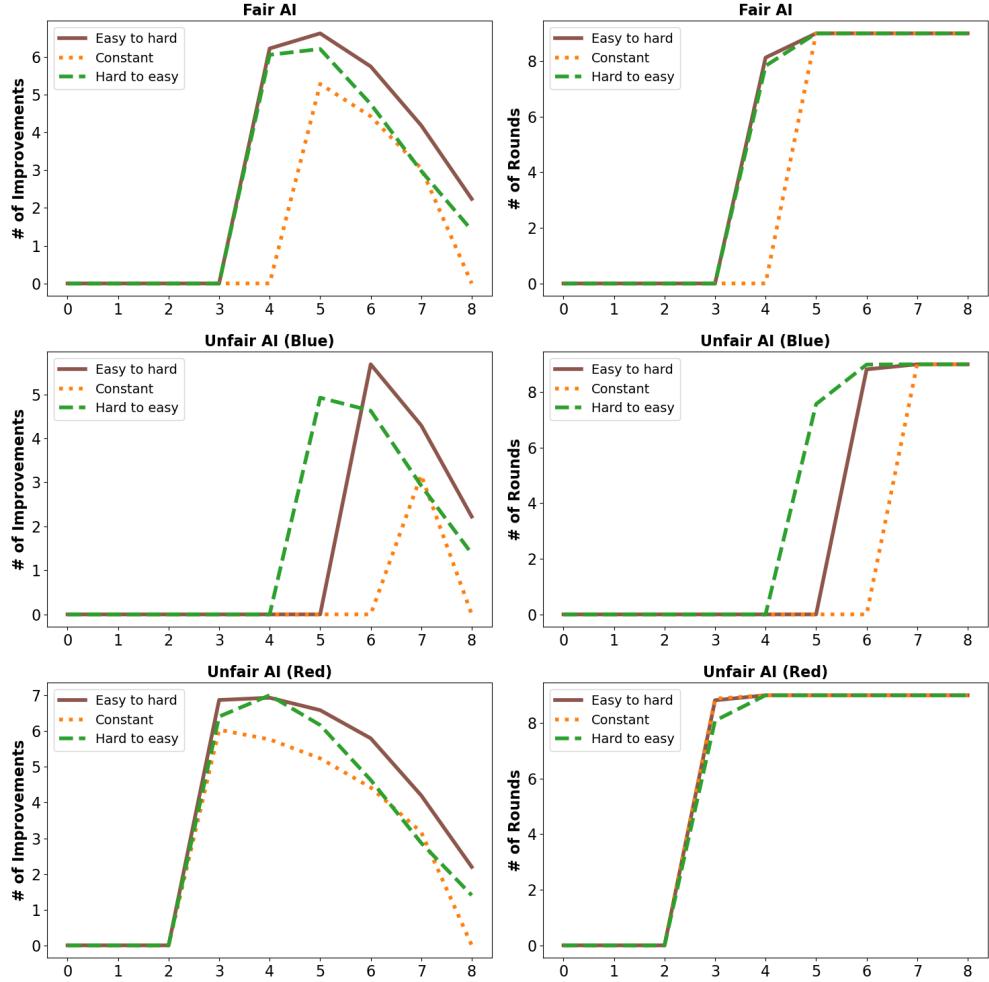


Figure A7: This figure presents the outcomes of implementing a 5-coins improvement cost, considering both Fair and Unfair AI treatments across three studies. The x-axis represents the initial credit score ranges, categorized from 0 to 8, with higher scores indicating higher ranges. In the Fair AI scenario (first row), the figure displays two graphs: the left graph indicates the expected number of improvements, while the right graph shows the expected number of rounds for subjects to keep interacting with the AI model. These are given within each credit score range and for each improvement difficulty used across our three studies. The middle and third rows represent the Unfair AI treatment for blue/female (disadvantaged) and red/male (advantaged) group subjects, similarly displaying “number of times to improve” and “number of rounds to apply for a loan” strategies as in the first row. The various improvement difficulties are color-coded: constant difficulty used in Studies 1 and 3 (orange), hard to easy in Study 2’s sub-experiment (green), and easy to hard in Study 2’s other sub-experiment (brown), illustrating different expected strategies for each improvement difficulty across credit score ranges. 5 coins improvement cost seems to be a better fit for our setting.

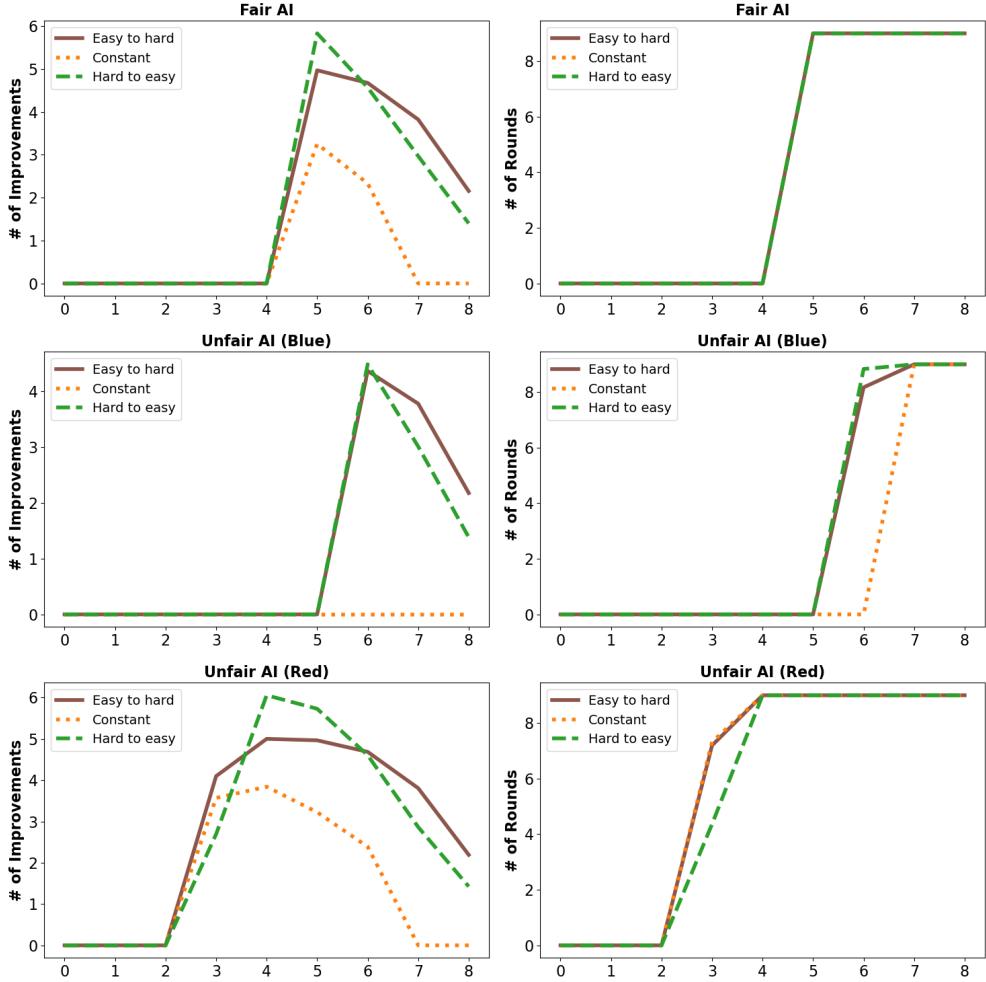


Figure A8: This figure presents the outcomes of implementing a 10-coins improvement cost, considering both Fair and Unfair AI treatments across three studies. The x-axis represents the initial credit score ranges, categorized from 0 to 8, with higher scores indicating higher ranges. In the Fair AI scenario (first row), the figure displays two graphs: the left graph indicates the expected number of improvements, while the right graph shows the expected number of rounds for subjects to keep interacting with the AI model. These are given within each credit score range and for each improvement difficulty used across our three studies. The middle and third rows represent the Unfair AI treatment for blue/female (disadvantaged) and red/male (advantaged) group subjects, similarly displaying “number of times to improve” and “number of rounds to apply for a loan” strategies as in the first row. The various improvement difficulties are color-coded: constant difficulty used in Studies 1 and 3 (orange), hard to easy in Study 2’s sub-experiment (green), and easy to hard in Study 2’s other sub-experiment (brown), illustrating different expected strategies for each improvement difficulty across credit score ranges. A high improvement cost might cause most decision subjects to not improve.

