FAIRNESS IN THE MACHINE LEARNING PIPELINE: A HUMAN-IN-THE-LOOP PERSPECTIVE

by

Meric Altug Gemalmaz

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Computer Science

West Lafayette, Indiana

August 2025

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ming Yin, Chair

Department of Computer Science

Dr. Clifton W. Bingham

Department of Computer Science

Dr. Dan Goldwasser

Department of Computer Science

Dr. Sooyeon Jeong

Department of Computer Science

Approved by:

Dr. Voicu S. Popescu

To my family, friends, and the many good people around the world who have made me who I am today.

ACKNOWLEDGMENTS

Foremost, I would like to begin by thanking my advisor, **Ming Yin**. It's been a long journey, Ming—thank you so much for your mentorship over the years. Thank you for teaching me how to handle uncertainty, overcome doubt, and face challenges with clarity. In the midst of the unknown, you always had a way of showing me the right course of action—how to step back, see the big picture, and make thoughtful decisions. From 2018, when I was an undergraduate taking your seminar course to understand graduate school, all the way to the PhD and graduation, it has truly been a privilege to have your guidance. I am incredibly grateful for the knowledge and wisdom you have shared with me. I will carry it with me into my career and into the challenges life may bring. Beyond research, your guidance prepared me to navigate the complexities of life, and for that, I am truly thankful. Thank you so much for everything, Ming. I will never forget our research lab. It's time for all of us to move on and flourish in our respective careers—thank you for empowering us to do that and for encouraging us to pursue our dreams.

To my committee member **Sooyeon Jeong**—thank you so much, Sooyeon. The fourth chapter manuscript wouldn't have looked this good without your valuable feedback. Thank you for your insights on the introduction and other parts of this thesis as well. Additionally, I appreciate all the advice you've given me on research and academic presentation, and also for giving me the opportunity to guest lecture in your course. The tips and guidance you provided will stay with me for years to come. I would also like to thank my other committee members, **Chris Clifton** and **Dan Goldwasser**. Starting with Chris, I truly appreciate your feedback on my third and fifth chapters. Your suggestions helped me make strong design decisions and carry out successful experiments. Thank you for sharing your years of experience and helping me stay on the right track. As for Dan, thank you for challenging me to think outside the box and encouraging me to look beyond my immediate research to consider broader problem-solving approaches. I am deeply grateful for your invaluable feedback and guidance.

To my labmates: First, I would especially like to thank **Amy Rechkemmer**. Thank you, Amy, for your friendship, mentorship, and support over these years. You've helped

me in countless ways, and I will always be grateful for your presence and support through various stages of my academic life. And, know that I will never forget our great memories at Purdue CS. Thank you so much for everything! Second, **Chun-Wei Chiang**—your presentation skills are truly impressive. Thank you for teaching me how to prepare effective slides and deliver strong presentations. I'm grateful for the time and feedback you offered me. Third, **Xiaoni Duan**—thank you for everything. I've always admired your calm presence, even in the most stressful situations. Your ability to say "it's okay" and guide everyone through challenges left a lasting impression on me. I will always remember your positivity and friendship. Fourth, **Syed Hasan Amin Mahmood**—thank you, Hasan, for your friendship and for the amazing resume and CV tips you shared with me in my final year. Few people would take the time to sit down and spend hours helping a friend rework their LinkedIn or job materials. I will never forget your help during one of the most stressful periods of my life. Fifth, **Xinru Wang**—thank you for your insightful seminars on the job search process and your helpful tips about industry careers. Lastly, but definitely not least, I would like to thank **Yizhou Harry Tian** for his friendship over the past two years. Thanks to all of you—graduate school would not have been the same without you. I will carry a part of each of you with me as I move forward. Good luck, my friends. Take care, and may your life be filled with amazing moments in the years to come!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Machine learning (ML) models are now ubiquitous in many fields, fitting into everyday routines and essential industrial processes. However, the human-led world is inherently marked by bias, and unfortunately, this bias can infiltrate ML models, leading them to act in ways that may not be fair or just. To better understand bias and fairness issues in ML models, adopting a human-in-the-loop perspective is essential, as humans play a key role in the development of the ML pipeline, appearing at its various stages. For instance, people can serve as annotators contributing their data for tasks like data labeling or influence the model through their daily decisions, such as responding to specific situations in life, which are then compiled into datasets to train these models. Subsequently, once these models are deployed in decision-making scenarios, such as determining loan approvals, they make crucial decisions on people, who then take the role of interacting with the models in more nuanced and strategic ways. If we examine the human roles in the ML pipeline more closely, we find that in each role they assume, humans significantly affect bias and fairness issues in the development of the ML pipeline. This is because bias and fairness concerns both originate from and directly impact humans themselves, and their strategic reactions to ML models can further influence the long-term fairness and biases of these models. Consequently, understanding and addressing human biases throughout the ML model development cycle, particularly at each stage of the pipeline itself, is essential. Careful monitoring of the models' statuses and tackling their stage-specific biases can allow us to identify and reduce the potential for unfair decisions, thereby ensuring the fair implementation of ML models. Therefore, this dissertation examines the ML pipeline through a human-in-the-loop lens to better understand and model how human roles contribute to fairness issues within ML models.

In this dissertation, I present my findings on the comprehension and modeling of human behavior and biases within various stages of the ML pipeline. First, in the data annotation/-collection stage, my exploration reveals that human cognitive biases, such as confirmation bias, will significantly affect the quality of the training data. Building on this insight, I devised a bias-aware algorithm that directly models this bias. The algorithm has demonstrated effectiveness in inferring ground-truth labels for the training data, especially when

subjects exhibit polarized values. Second, in the deployment stage, where ML models are implemented and begin making decisions on people, I investigate how decision subjects (i.e., the people who are subject to ML models' decisions.) strategically react to an ML-based system's fairness across groups or favoritism towards certain groups, particularly within the context of loan lending. This research uncovers the ways in which decision subject's perceptions of fairness and engagement with the system are influenced by the fairness properties of the ML models. Largely, the results unveil that decision subjects' behavior seem to be more influenced by the system's favoritism towards their own group rather than the system's fairness across groups. Finally, I investigate the feedback loop that emerges from the combination of multiple stages—specifically, the deployment stage, where ML models are implemented and make decisions about people, the behavioral data stage, where individuals react to these decisions, and the model update stage, where models are retrained using that behavioral data. In this loop, I examine how fairness evolves over time in ML models that are repeatedly updated using behavioral data that is generated as a reaction to models' previous decisions. Through a simulation study, I model interactions between decision subjects and ML models in a task assignment setting managed by the models themselves. The results reveal how initial fairness can degrade due to feedback loops formed by biased feedback and subject reactions—such as reduced effort or early disengagement in response to perceived unfairness. I further validate these findings through a human-subject experiment in a setting where subjects interact with a real ML model that assigns tasks. Notably, the results reveal that seemingly intuitive heuristics—such as compensating individuals with more tasks after they receive fewer due to unfair treatment—can slow the growth of disparities, but only to a limited extent if the model continues to rely on biased feedback, ultimately exacerbating disparities. Overall, this dissertation explores the human-in-the-loop perspective of the ML pipeline, identifying key areas where bias can be introduced or perpetuated. It emphasizes the importance of carefully examining and controlling each stage to ensure the development of fair ML models in their decision-making. These findings contribute valuable insights into the design of more responsible and equitable ML systems, highlighting the essential role of understanding and addressing human behavior and biases throughout the pipeline.

1. INTRODUCTION

Machine learning (ML) models have become essential across various industries, playing diverse roles, and making significant impacts worldwide. Now, more than ever, these models are woven into the fabric of daily life, offering vast benefits. For instance, OpenAI's ChatGPT [1] acts as a virtual assistant for numerous users. Email systems employ ML, combined with user feedback, to sharpen their spam filter accuracy [2]. In the financial sector, analysts utilize ML for functions like algorithmic trading, fraud detection, and automated financial advising, deriving insights from news feeds and trading patterns [3]. In high-security contexts, image classification is pivotal for detecting weapons or identifying potentially dangerous individuals [4]. These instances emphasize ML's adaptability across many diverse purposes, underscoring its widespread usage in the world.

Even though ML models are applied in many contexts, the human-led world is inherently marked by bias, and unfortunately, this bias can infiltrate ML models, leading them to act in ways that may not be fair or just. For instance, racial bias has been observed in recidivism prediction models; the COMPAS model was shown to systemically predict that African Americans are more likely to re-offend compared to white individuals who had committed the same crime, revealing profound unfairness in its outcomes [5]. Gender bias has also manifested in financial domains; Apple used an ML model to determine credit card application approvals, which was found to favor male applicants over female applicants, resulting in disproportionately higher rejection rates for female applicants [6]. Even models designed for conversation and information retrieval, such as ChatGPT, are not immune to these biases and have sometimes produced factually incorrect or even sexist, racist, or offensive text [7].

To better understand bias and fairness issues in ML models, adopting a human-in-the-loop perspective is essential, as humans play a key role in the ML development and deployment pipeline, appearing at its various stages. For instance, some people serve as data workers, often compensated for creating or labeling data, thus supplying datasets for ML models [8]. Meanwhile, for many, the contribution is more indirect: the decisions they make, tasks they undertake, or content they post on social media all feed into the evolution of these models [9]. The human-generated data provides training material for ML models, offering

insights into the workings of a human-led world and enabling the models to comprehend the dynamics of human life more effectively. However, the human role in ML extends beyond merely providing data for model training. Once these models are deployed in decision-making scenarios, such as determining loan approvals, they make crucial decisions on people, who then interact with the models in more nuanced and strategic ways [10–12]. For instance, within the field of loan lending services, the pattern of rejections or approvals dictated by an ML model may shape an individual's decision to either continue using that specific service or explore alternative options. These strategic reactions or behavioral data generated by humans can be leveraged by ML models to refine and update their understanding of the world. This, in turn, influences the ML models' subsequent decisions, further impacting people's long-term interactions with them. All of this continuous cycle of interaction and adaptation underscores a relationship where people not only shape but also respond to the technology that increasingly governs various aspects of their lives, revealing a complex interplay between humans and ML.

The different roles that humans play in the development of the ML pipeline significantly affect bias and fairness issues in ML models. This is because biases and fairness concerns of ML models may both originate from and directly impact humans themselves. In addition, humans' strategic reactions to ML models can greatly influence the long-term fairness and biases of these models. Specifically:

- **Humans may provide biased annotated data for training ML models**: Humans play a pivotal role as data annotators, especially on crowdsourcing platforms, such as Amazon Mechanical Turk[1], and contribute their expertise to a variety of tasks, from fact-checking to text translation, simplifying the solution for many large-scale data problems [13]. However, human biases can sneak into data annotations and significantly decrease the data quality procured from annotators [14]. For example, when tasked with fact-checking subjective content, such as political statements, an annotator's personal biases can lead to skewed or erroneous labels [15]. That is, if a political statement aligns with their personal beliefs, they might erroneously label it as true, even if it is a factual error. This potential for bias is not just limited to subjective tasks. Even in more objective

[1]↑https://www.mturk.com/

contexts, errors due to underestimation, overestimation, or flawed reasoning can introduce inaccuracies to the training data [16], leading ML models astray in their problem-solving once trained on these datasets. When annotators introduce biases during data labeling (i.e., stage 1 of Figure 1.1), the resulting skewed training datasets cause ML models to learn and internalize these biases (i.e., stage 2 of Figure 1.1). Consequently, when these models generalize from the biased data, they may reflect the same kind of biases and produce predictions or classifications that systematically favor certain groups or outcomes over others, leading to unfair and often inaccurate results in real-world applications.

- **ML models' decisions on people lead to their strategic reactions**: Once deployed, the model's actions might encourage strategic behavior, such as individuals gaming the system to maximize personal gain [12, 17, 18] (i.e., stage 3 of Figure 1.1). For instance, in a loan-lending context where an ML model is a decision-maker, if one gender consistently receives loan approvals, they might exploit the system to their advantage, while the opposite gender, facing consistent rejection, may choose to use the system less [12].
- **Updating ML models with strategic reactions may escalate long-term fairness issues**: As time progresses, users' strategic responses to model decisions are collected as their behavioral data (i.e., stage 4 of Figure 1.1). This data, in turn, will be then used to update the ML model influencing the model's further decisions (i.e., stage 5 of Figure 1.1). Considering an algorithmic management scenario—such as a gig work platform where an ML model governs task assignments—if the model favors one group over another (e.g., by gender), members of the advantaged group may receive more tasks and remain more engaged, generating behavioral data that predominantly reflects their participation. In contrast, disadvantaged workers may disengage or reduce their effort in response to perceived unfairness, leading to underrepresentation in the data. When the model is subsequently updated using this skewed behavioral data, the system may reinforce and amplify its initial bias, further widening disparities over time [10, 11]. This cyclical process might inadvertently perpetuate biased decisions, increasing their effects over time. As the model updates and potentially becomes more biased, it may further provoke strategic reactions from people, creating a vicious cycle of escalating bias.

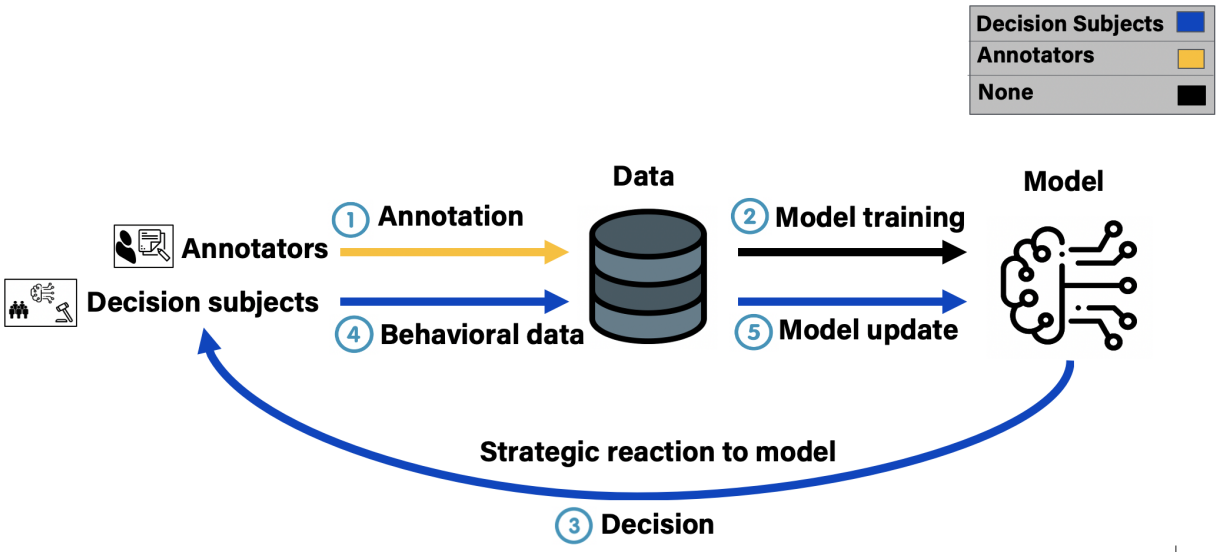


Figure 1.1. From a human-in-the-loop perspective, the ML pipeline illustrates the cyclical nature of bias and fairness issues. Importantly, humans can introduce biases during data generation, especially through annotations. This compromises the fairness of ML models during training. Such compromised fairness can lead decision subjects to react strategically, especially when the model disproportionately favors certain sub-groups. Frequently, the behavioral data from these reactions is incorporated to refine the ML models. However, this strategy can sometimes amplify the model's existing fairness concerns in subsequent updates, perpetuating a vicious cycle. In this illustration, distinct colors signify different groups of people—annotators or decision subjects—engaged in specific stages of the pipeline.

These dynamics underscore the intricate relationship between humans and ML models. While human involvement is indispensable in the ML pipeline, it also introduces many complexities concerning bias and fairness. Consequently, ensuring fairness and mitigating bias in ML models necessitate a vigilant examination of each stage of the ML pipeline where humans play a unique role. Addressing these challenges requires a human-in-the-loop perspective. Therefore, in this dissertation, we adopt this perspective to thoroughly investigate the ML pipeline. We focus on the diverse roles humans have in both introducing and influencing biases, and consequently, in shaping the fairness of the ML models.

1.1 Thesis Statement

The fairness issue in ML models arises from human involvement and significantly impacts humans themselves. Biases introduced by humans in generating data can affect the fairness of models, eliciting strategic reactions from humans, which, in turn, can influence subsequent updates to the ML model. An essential step in addressing the fairness issues in ML models is to comprehend and model human behavior and biases at each of these stages.

1.2 Understanding the Fairness of ML Pipeline from a Human-in-the-loop Perspective

Starting with the first stage of the pipeline, as shown in Figure 1.1, exploration extends to how peoples' cognitive biases infiltrate the labels they generate for data items. In this thesis, we focus on one type of bias that annotators might exhibit, that is their confirmation bias (i.e., the tendency to favor information that confirms existing beliefs and values). People are known to exhibit confirmation bias when fact-checking political statements [19–21]. In light of this, we adopted a experimental setting where annotators are tasked with distinguishing between factual and opinion statements. They were instructed that a statement should be classified as factual if it is presented as a fact, regardless of its correctness, including cases where it is a false factual statement. This approach was utilized to investigate whether annotators' political stances introduce bias into their annotations. Our initial findings indicate that, in tasks focused on differentiating factual from opinion statements, annotators might label statements as factual only if they align with their personal values. Consequently, we developed an algorithm that quantitatively models the confirmation bias and, by accounting for it, aggregates the provided labels to correctly infer ground-truth labels for each data-item. Our findings indicate that directly modeling a cognitive bias and devising a model that accounts for it significantly improves aggregated annotation accuracy. Specifically, evaluations on real-world annotations demonstrate that our proposed bias-aware label aggregation algorithm outperforms seven popularly used baseline algorithms in accurately inferring ground-truth labels of various political statements, especially when annotators re-

flect more confirmation bias in their annotations. We find that our algorithm works best when annotators' own political stances are widely dispersed or even polarized.

We then move on to investigate how the fairness properties of fair and unfair ML models affect decision subjects' (i.e., individuals affected by the models' decisions) repeated interactions with and perceived fairness of the models (i.e., the third stage of the pipeline; see Figure 1.1). While the fair model treats subjects from both groups equally, the unfair model systematically favors one group over another. Previous research on "long-term fairness" [11, 22–26] theoretically explored how people's long-term interactions with ML models influence their behavior over time. Consequently, we incorporated this setting into an empirical study to examine real-world behavior. We designed a task setting that was carefully thought out to mirror real-world loan application scenarios where the loan decisions are made by an ML model with respect to decision subjects' "qualification" (e.g., credit score) for a favorable decision (e.g., loan approval). Using this simulated loan application task, we then investigated how an ML-based loan decision system's favoritism towards a particular group, or its unbiased (fair) treatment of various groups, influences the strategic reactions of those affected by its decisions. The decision subjects' strategic reactions included retention decisions, such as choosing to remain in or leave the system. Additionally, we also assessed the subjects' long-term perceptions of the ML model's fairness after their interactions concluded with the model. The results unveil that when decision subjects repeatedly interact with an ML-based decision system and can respond strategically to it, their fairness perceptions and retention seem to be more influenced by the system's favoring tendency of their own group rather than the system's unbiased treatment of different groups. However, subjects' qualifications for the loans act as a moderator for this effect. For example, we find that under biased treatment, higher-qualified subjects perceive the system as less fair than lower-qualified ones, and when their own group is favored, they show a smaller increase in their fairness perceptions. Additionally, higher-qualified subjects are marginally more likely to leave the system when they belong to a disadvantaged group. Finally, our comprehensive analysis further reveals that subjects' retention in the ML-based decision system is mainly driven by their own prospects of favorable decisions, while their fairness perceptions are more complexly influenced by the system's treatment of people in the other groups.

In our follow-up work, we maintain our focus on stage 3 of the pipeline but consider a more comprehensive set of strategic reactions available to decision subjects. In this work, not only can subjects decide whether to be subject to the ML models' decisions, but they also have the option to improve their qualification for a favorable decision, such as loan approval. We investigate various schemes that make improvement either more difficult or easier for subjects with increasing qualifications, examining how improvement difficulty can change the strategic reactions that subjects display in the simulated loan application task. Furthermore, we introduce another dimension of complexity by investigating the impact of the presence or absence of the ML models' fairness evaluation based on prominent protected attributes, such as gender. Our results overall indicate that, with the existence of varying difficulty of improvement opportunities, subjects' interactions with the ML model change in nuanced ways. Specifically, whether the ML model treats groups without bias, or whether it favors or disfavors a subjects' own group, does not significantly influence the subjects' willingness to continue interacting with the model anymore. Similarly, model's fairness properties also do not affect subjects' motivation to improve their qualifications. However, consistent with our findings where qualification improvement was not incorporated into the task setting, here subjects still perceive the model as less fair if it consistently disfavors their group. The subsequent analysis, which explores the effects of the ML models' fairness evaluation using or not using the subjects' sensitive attributes, does not alter these findings. Overall, we find that the relationship between ML models' fairness properties and both engagement (i.e., subjects' retention and qualification improvement) and fairness perceptions is complex. The factors influencing engagement might differ from those shaping fairness perceptions regarding the ML model.

Building on findings from decision subjects' repeated interactions with ML models that were not updated over time, the final part of this dissertation investigates how fairness evolves over time when models are retrained using behavioral data generated in response to models' previous decisions. Specifically, we simulate and empirically examine the feedback loop that emerges across stages 3 to 5 of the ML pipeline (Figure 1.1), where deployed models influence human behavior, that behavior is captured as data, and the models are updated accordingly. Through a simulation study of ML-based task assignment systems, we model decision sub-

jects who react to perceived fairness by adjusting their task completion behavior—either in terms of retention or work quality—and raters who may introduce biased evaluations of subjects' performance. Over time, we observe how these reactions influence future model updates, revealing that even initially fair models can drift into unfairness when exposed to skewed or biased feedback. Surprisingly, strong reactions—such as subjects consistently leaving early—can sometimes limit the model's ability to update, which may incidentally slow the growth of disparities. Conversely, when disadvantaged subjects remain in the system and improve their work quality, the biased feedback loop can be partially corrected, also contributing to a slower increase in disparity over time. To complement the simulation findings, we also conduct a human-subject experiment in which subjects repeatedly interact with a deployed and evolving ML-based task assignment system. Subjects are placed in treatments where they can skip tasks or report unfairness. The results validate key insights from the simulation: intuitive remedies—such as simply assigning more tasks to compensate for prior unfairness—can slow the growth of disparities, but only to a limited extent if the underlying model continues to learn from biased feedback. Additionally, decision subjects adapt their behavior in response to perceived unfairness—an aspect generally not assumed in the simulation studies modeling human behavior. Taken together, these findings demonstrate that fairness is not a static property but one that evolves through continuous interactions between dynamic human behavior and algorithmic adaptation. In other words, this final part of the dissertation presents a study—combining simulation-based modeling with empirical experimentation—that highlights the value of hybrid approaches for understanding and addressing fairness in algorithmic management systems. It emphasizes the need to design fairness-aware interventions that are robust to feedback loops and the long-term dynamics of human-ML interaction.

1.3 Contributions and Thesis Overview

This dissertation is divided into three major sections of work. The first section focuses on modeling a bias and mitigating it within the inferred ground-truth data annotations, touching upon Figure 1.1, stage 1 **(Chapter 2)**. The second section delves into the decision

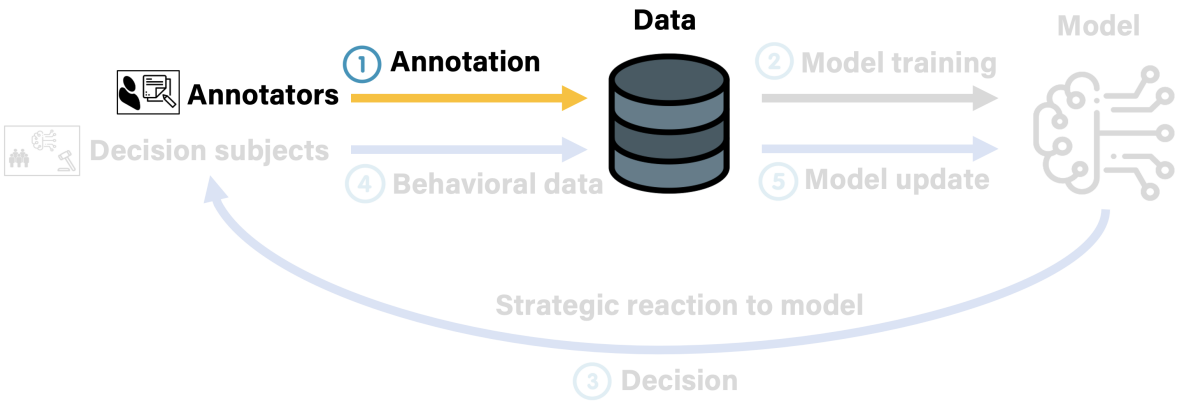
subject's long-term strategic reactions to an ML model's decision fairness, encompassing stage 3 of Figure 1.1 **(Chapters 3 & 4)**. Finally, the third section, encompassing stages 3 to 5 of the ML pipeline as depicted in Figure 1.1 **(Chapter 5)**, examines how the cycle from ML deployment to human reaction and subsequent model update unfolds through a dynamic feedback loop, investigating how ML fairness evolves under repeated interactions and continuous retraining.

My work and contributions related to investigating the fairness and bias within the ML pipeline include:

- Use of probabilistic graphical models to represent the confirmation bias of subjects when providing data annotations. **In Chapter 2**, we employ probabilistic graphical models to represent the confirmation bias exhibited by subjects during data annotation. This approach demonstrates a method of quantitatively modeling human cognitive biases. Building on this, we then develop an algorithm that uses the expectation-maximization technique, specifically designed to mitigate this bias in the final inferred ground-truth annotations.
- The first empirical investigation of decision subjects' repeated interactions with an ML-based decision system. **In Chapter 3**, we reveal insights into how fairness perceptions and retention are influenced by ML-based decision systems. We find that both aspects are significantly affected by the system's tendency to favor the subjects' own group. Specifically, retention is primarily driven by the prospects of favorable decisions, while fairness perceptions are more nuanced. **In Chapter 4**, further examinations introduce the concept of qualification improvement as a strategic reaction. These findings indicate that the ML model's tendency to favor the subjects' own group no longer significantly affects their engagement with the model. However, subjects still perceive the model as less fair if it systematically disfavors their group. Additionally, our results remain consistent, irrespective of the varying difficulty in improvement and whether sensitive attributes are considered in assessing the ML model's fairness.
- A simulation and empirical investigation of how fairness evolves through feedback loops across deployment, human reaction (behavioral data), and model update stages of the ML

pipeline. **In Chapter 5**, we model repeated interactions between decision subjects and ML systems, where subjects react to fairness of the model's task assignments, and the model is periodically updated using subjects' behavioral data. We simulate both fair and unfair initial models and observe how their fairness properties change over time. Our results show that biased feedback can cause initially fair models to degrade, while subjects' reactions can influence the feedback loop, sometimes creating data gaps that slow the growth of disparities or self-correcting feedback loops, and other times reinforcing those disparities even further. We validate these dynamics through a human-subject experiment in which subjects interact with an constantly updating task assignment model, with options to skip tasks or report issues. The results validate key insights from the simulation: intuitive remedies—such as assigning additional tasks to compensate for prior unfairness—can slow the growth of disparities, but only to a limited extent if the underlying model continues to learn from biased feedback.

2. ACCOUNTING FOR COGNITIVE BIASES IN CROWDSOURCED LABEL AGGREGATION



Exploring the "Annotation," stage 1, of the machine learning pipeline. The complete figure is presented in Figure 1.1 of Chapter 1.

We begin with the initial and pivotal stage of the machine learning (ML) pipeline, examining the "Annotation" stage, depicted as stage 1 in Figure 1.1 of Chapter 1. This stage enables ML models to grasp the complexities of the human world. Training data, often sourced from individuals who are compensated for completing specific tasks, encapsulates a myriad of task types—both objective and subjective. As a result, annotations are prone to a wide range of human biases. Consequently, addressing and accounting for these biases is vital in this foundational phase to ensure that ML models do not perpetuate biased decisions when deployed in the real-world. Typically, crowdsourcing is employed for data collection, offering a convenient avenue to gather diverse annotations from a broad spectrum of contributors for vast datasets. However, a long-standing challenge in crowdsourcing is how to control the quality of crowd work [27]. Recently, it has been recognized that an important factor contributing to the limited work quality of individual crowd workers is that they are prone to a wide range of *biases* in their work. For example, workers may be influenced by social biases (e.g., racial bias, gender bias) during their annotation process [28, 29]. The design of crowdsourcing tasks, such as the order in which information is shown to workers, may also have subtle impacts on workers and trigger their cognitive biases, such as the anchoring bias and ambiguity effect [30, 31].

Another common type of cognitive bias that crowd workers are often subject to is their *confirmation bias*, which refers to people's tendency of favoring information that confirms their previously existing beliefs and values [32]. Indeed, researchers have showed that when judging the truthfulness of news statements, crowd workers tend to believe those statements coming from speakers off the same political party that they have recently voted for to be more true [21]. Similarly, Hube *et al.* [20] revealed that when crowd workers are asked to determine whether a statement is neutral or opinionated, they are more likely to label a statement as neutral if its stance aligns with their own opinions.

In practice, to obtain high-quality annotations from crowd workers, a redundancy-based strategy is often deployed. That is, the same task is completed by multiple workers, and numerous label aggregation algorithms have been proposed to infer the ground-truth answer for each task based on the collection of annotations obtained on it [33–37]. While these algorithms adopt various models to characterize worker behavior during the label generation

process, they seldom take worker's cognitive biases, such as their confirmation bias, into account. In so doing, the current crowdsourced label aggregation algorithms might have missed the opportunity to further improve the inference accuracy by explicitly modeling how worker's cognitive biases have influenced their work quality.

Therefore, in this chapter, we focus on worker's confirmation bias and propose a new label aggregation algorithm to account for it. Specifically, we formulate a probabilistic model of the label generation process by assuming that among other factors, worker's label on a task is influenced by both the values of the worker and the values expressed in the task. We then make use of the expectation-maximization algorithm to simultaneously infer the values of each worker, the values of each task, as well as the ground-truth answer for each task.

To examine the effectiveness of the proposed algorithm, we collect annotations from real crowd workers on Amazon Mechanical Turk on the subjective task of differentiating factual statements from opinion statements, for which workers are shown to indeed exhibit a degree of confirmation bias in their annotations. We find that compared to a set of baseline label aggregation algorithms, the proposed bias-aware label aggregation algorithm achieves a higher level of accuracy in uncovering the ground-truth label for each task. We further investigate the robustness of the proposed algorithm through simulations using synthetic datasets. Our simulation results highlight several scenarios that the proposed algorithm shows the largest advantage over baseline algorithms, such as when crowd workers suffer from confirmation bias in their annotations to a larger extent and when the distribution of worker's values is more dispersed or even polarized.

2.1 Related Work

2.1.1 Quality Control in Crowdsourcing.

To solicit high-quality work from inexpert crowd workers, researchers have proposed a variety of strategies such as providing effective incentives to workers [38], training novice workers [39], assigning tasks to workers with relevant skills [40], and enabling communication between workers on the same task [41]. Yet, in practice, the most widely adopted approach for ensuring the quality of crowd work, especially for simple classification tasks, is to assign

a task to multiple workers and then infer its correct answer using all annotations collected on it.

To effectively combine multiple annotations and infer the ground-truth label for a task, researchers have designed various label aggregation algorithms to improve the inference accuracy by explicitly characterizing how worker's quality in a task is affected by multiple factors. For example, Whitehill *et al.* [33] characterized worker's labeling process using a probabilistic graphic model assuming that a worker's label on a task is influenced by the worker's skill level as well as the task difficulty. Welinder *et al.* [34] introduced a more sophisticated model to capture worker's diverse skills on various latent topics underlying a task. More recently, Braylan and Lease [42] extended label aggregation algorithms from simple annotations (e.g., class labels) to complex annotations (e.g., open-ended text) by modeling the distances between annotations. Moreover, Li *et al.* [43] proposed algorithms that ensure the aggregated labels satisfy fairness constraints. For a more complete review of label aggregation algorithms in crowdsourcing, please see [37].

2.1.2 Bias in Crowdsourced Annotations.

Recent studies showed that crowd workers could be influenced by a wide range of biases during their annotation process. Such biases can be triggered by the design of the tasks. For example, it is shown that grouping multiple data items together in a batch for workers to label may lead to the "in-batch annotation bias," that is, a workers judgment on one data item is affected by other data items within the batch [31]. Similarly, workers are also subject to the "sequential bias" in their labeling process such that their annotation on one task might be influenced by the previous task that they see as well as the label they provide on it [44, 45]. Within a single task, the ways that information is presented and the order that questions are asked can also result in worker's cognitive bias which negatively impacts the work quality [30]. In addition, workers may exhibit biases in their annotations as a result of the interaction between the characteristics of the worker and the task. For example, Biswas *et al.* [29] showed that when crowd workers are asked to assess the recidivism risk of criminal

defendants, they tend to slightly favor defendants of their own race, showing some degree of in-group bias.

Another type of bias that crowd workers are prone to, especially in subjective tasks, is confirmation bias. Via experimental studies, it is found that crowd workers tend to label a piece of news as true rather than fake, or a statement as neutral rather than opinionated, if the information expressed in the news or statements align well with the worker's own belief and value [20, 21]. Researchers have also revealed that confirmation bias may largely explain why in the real-world crowdsourcing applications of misinformation flagging on social-media platforms, the news sources flagged by the crowd tend to be the most popular (and largely reliable) ones [46].

2.1.3 Mitigate Confirmation Bias in Crowdsourcing.

Most recently, researchers have explored different approaches to mitigate crowd worker's confirmation bias and reduce the negative impact the bias brings to work quality, which have mixed success. For example, Hube *et al.* [20] showed that raising people's awareness of their own bias can effectively reduce worker's bias in annotations. On the other hand, it is found that enabling workers with different beliefs and values to work on the same task and interact with each other does not help reduce worker's bias [47]. This chapter provides a new approach in "mitigating" confirmation bias— we explicitly model how worker's confirmation bias sneak into their annotations, and then design algorithms based on such model to reduce the bias in the final, aggregated labels.

2.1.4 Bias-aware Label Aggregation

In this section, we outline our algorithmic approach for crowdsourced label aggregation which takes annotators' confirmation biases into account. We consider subjective labeling tasks in which annotators are asked to provide binary labels in each task, and importantly, one of the two candidate labels is generally perceived to be more "*preferable*" (e.g., a piece of news is "true" rather than "fake," a statement is "neutral" rather than "opinionated"). On these tasks, annotators might be subject to confirmation bias—they might favor infor-

mation that confirms their previously existing beliefs or values, hence increase their chance of providing the preferable label in tasks containing the favorable information.

2.1.5 Label Generation Model

Consider the scenario that N annotators are asked to complete M binary labeling tasks. An annotator i's label on task j is denoted as $l_{\mathrm{ij}} \in \{0, 1\}$, with 0 representing the preferable label (e.g., "true news", "neutral statement"). Our goal is to determine the true label, $z_{\mathrm{j}} \in \{0, 1\}$, for each task j using all the labels collected on it. To model annotators' possible confirmation bias during their label generation processes, we assume the observed labels l_{ij} depend on several causal factors: (1) the values implied by the information in the task; (2) the annotator's values; (3) the annotator's degree of bias characterizing how much the annotator is subject to confirmation bias, (4) annotator's inherent tendency to provide the preferable label, and (5) the true label of the task. Under our model, the chance for annotator i to provide the preferable label on task j (i.e., $l_{\mathrm{ij}} = 0$) is characterized as:

$$P(l_{\mathrm{ij}} = 0 \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a) = \frac{1}{\mathrm{e}^{a[(1-p_{\mathrm{i}})(s_{\mathrm{j}}-c_{\mathrm{i}})^2 + p_{\mathrm{i}} z_{\mathrm{j}}]}} \tag{2.1}$$

In Eqn. 2.1, for simplicity, we model the values of both the annotators and the information expressed in the tasks using a single dimensional spectrum—the values of annotator i are captured by the parameter $c_{\mathrm{i}} \in [0, 1]$, while the values of the information contained in task j are captured by the parameter $s_{\mathrm{j}} \in [0, 1]$[1]. For example, when considering the leftright political spectrum, $c_{\mathrm{i}} = 1$ (or $s_{\mathrm{j}} = 1$) could mean the values of annotator i (or the values implied by information in task j) are extremely conservative, while $c_{\mathrm{i}} = 0$ (or $s_{\mathrm{j}} = 0$) means the values of annotator i (or the values implied by information in task j) are extremely liberal. Annotators' confirmation bias is captured via the *distance* between c_{i} and s_{j}—holding all other variables equal, the closer c_{i} and s_{j} are to each other, the more likely annotator i will provide the preferable label in task j (i.e., $P(l_{\mathrm{ij}} = 0)$ is larger).

[1]↑Our model can easily be extended to cases where the values of annotators and tasks are characterized in a multi-dimensional space.

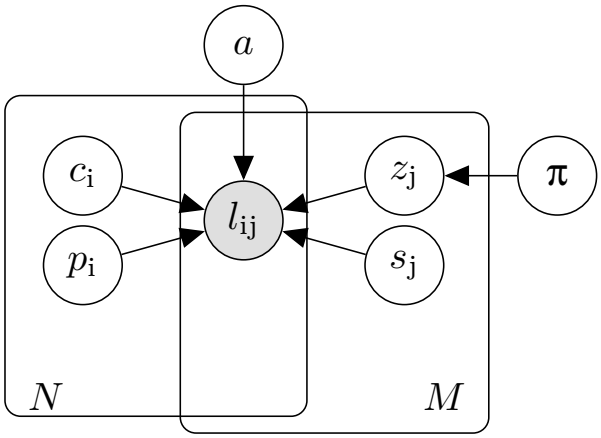


Figure 2.1. The probabilistic graphical model of annotators' label generation process. The shaded node is observed.

We further use the parameter $p_{\mathrm{i}} \in [0, 1]$ to characterize the extent to which annotator i is subject to confirmation bias. Here, $p_{\mathrm{i}} = 0$ means that annotator i is heavily influenced by her confirmation bias, such that she decides her label on tasks (almost) entirely based on how much the information contained in the task aligns with her values. Conversely, when $p_{\mathrm{i}} = 1$, annotator i is not influenced by her confirmation bias at all, such that she decides her label on tasks (almost) entirely based on the ground truth label z_{j} of the task, and $z_{\mathrm{j}} \sim \text{Bernoulli}(1 - \boldsymbol{\pi})$ (i.e., the prior probability for a task to have the preferable label as its ground truth is $\boldsymbol{\pi}$, $P(z_{\mathrm{j}} = 0) = \boldsymbol{\pi}$). When $0 < p_{\mathrm{i}} < 1$, the annotator is influenced by her confirmation bias to some degree, and the smaller p_{i} is, the more she is subject to the confirmation bias.

Finally, we use a global parameter $a \in [0, +\infty)$ to represent annotators' inherent tendency to provide the preferable label on any task, or in other words, annotators' base rate of providing the preferable label in tasks. When $a = 0$, the base rate for annotators to provide the preferable label in tasks is very high, while $a = +\infty$ means the base rate for annotators to provide the preferable label in tasks is very low.

Our entire label generation model for the crowdsourced annotators is shown in Figure 2.1. Given a set of observed labels $\mathbf{L} = \{l_{\mathrm{ij}}\}$, the end goal of our label aggregation algorithm is to infer the most likely ground-truth label $\mathbf{z} = \{z_{\mathrm{j}}\}$ for each task, as well as the values of all hidden parameters (i.e., $\mathbf{s} = \{s_{\mathrm{j}}\}, \mathbf{c} = \{c_{\mathrm{i}}\}, \mathbf{p} = \{p_{\mathrm{i}}\}, a, \boldsymbol{\pi}$).

Table 2.1. Summary of model parameters and their acquisition status.

Parameter	Hidden?	Type	Description / Acquisition
c_{i}	Yes	Annotator-level	Value orientation of annotator i, inferred.
p_{i}	Yes	Annotator-level	Confirmation bias strength of annotator i, inferred.
s_{j}	Yes	Task-level	Value implied by task j, inferred.
z_{j}	Yes	Task-level	Ground-truth label of task j, inferred.
a	Yes	Global	Global tendency to give preferable label, inferred.
$\boldsymbol{\pi}$	Yes	Global	Prior probability that $z_{\mathrm{j}} = 0$, inferred.
l_{ij}	No	Observed	Label from annotator i on task j, observed.
N	No	Fixed	Number of annotators, known.
M	No	Fixed	Number of tasks, known.

2.1.6 Model Inference

We use the Expectation-Maximization (EM) algorithm to estimate the maximum likelihood estimates of the hidden parameters and infer the values of the hidden variables z_{j}.

In particular, in the Expectation step, we compute the posterior probabilities for each hidden variable z_{j} based on the current estimates of parameters and the observed labels:

$$\begin{aligned} p(z_{\mathrm{j}} \,|\, \mathbf{L}, \mathbf{c}, \mathbf{p}, \mathbf{s}, a, \boldsymbol{\pi}) &\propto p(z_{\mathrm{j}} \,|\, \boldsymbol{\pi}) p(\mathbf{L} \,|\, z_{\mathrm{j}}, \mathbf{c}, \mathbf{p}, \mathbf{s}, a) \\ &\propto p(z_{\mathrm{j}} \,|\, \boldsymbol{\pi}) \prod_{\mathrm{i} \in W_{\mathrm{j}}} p(l_{\mathrm{ij}} \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a) \end{aligned}$$

Here, we use W_{j} to denote the set of all annotators who have provided labels on task j. When $l_{\mathrm{ij}} = 0$, $p(l_{\mathrm{ij}} \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a)$ can be computed using Eqn. 2.1; otherwise, $p(l_{\mathrm{ij}} \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a) = 1 - P(l_{\mathrm{ij}} = 0 \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a)$.

For the Maximization step, we search for optimal parameter values to maximize the auxiliary function Q, i.e., the expectation of the complete data log-likelihood:

$$\begin{aligned} Q(\mathbf{c}, \mathbf{p}, \mathbf{s}, a, \boldsymbol{\pi}) &= E[\ln p(\mathbf{L}, \mathbf{z} \,|\, \mathbf{c}, \mathbf{p}, \mathbf{s}, a, \boldsymbol{\pi})] \\ &= E[\ln \prod_{\mathrm{j}} (p(z_{\mathrm{j}} \,|\, \boldsymbol{\pi}) \prod_{\mathrm{i} \in W_{\mathrm{j}}} p(l_{\mathrm{ij}} \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a))] \\ &= \sum_{\mathrm{j}} E[\ln p(z_{\mathrm{j}} \,|\, \boldsymbol{\pi})] + \sum_{l_{\mathrm{ij}} \in \mathbf{L}} E[\ln p(l_{\mathrm{ij}} \,|\, c_{\mathrm{i}}, p_{\mathrm{i}}, s_{\mathrm{j}}, z_{\mathrm{j}}, a)] \end{aligned}$$

The expectation is taken with respect to the posterior distributions of z_j that are obtained from the previous E-step. In each M-step, we use gradient descent to update hidden parameters to the values that locally optimize Q.

2.2 Experiment

In this section, we examine that on real-world subjective labeling tasks where annotators could suffer from confirmation bias, whether the proposed bias-aware label aggregation algorithm can help improve the accuracy of inferred labels.

2.2.1 Data Collection

To empirically evaluate the effectiveness of the proposed algorithm, we first collected a set of annotations generated by real crowd workers on the subjective task of differentiating factual statements from opinion statements. We used this task in our study since previous research showed that U.S. adults were more likely to label both factual and opinion statements as factual when they appealed more to their political side (i.e., "factual" is the preferable label) [48].

Table 2.2. Examples of gun control related statements that we used in our study.

Statement	Label	Values
Gun bans alleviate intimate partner homicide.	Factual	Liberal
Active shooter events in the U.S. are sometimes associated with mental illness.	Factual	Conservative
Easy usage of guns increases firearm-related deaths.	Opinion	Liberal
Most of the problematic shooting events were led by mentally ill people.	Opinion	Conservative

Specifically, from the list of controversial topics in US politics[2], we selected "gun control" as the main topic and created a set of statements related to it. We created these statements by first reviewing gun control related debate transcripts on an online debate platform DEBATE.ORG, and extracted the main talking points (e.g., gun violence, illegal guns) from both the supporters and opponents of gun control. Given a talking point, we extracted

[2]↑ https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues.

factual statements related to it from the latest Wikipedia pages, and rewrote them slightly to remove obvious cues indicating the statements as factual (e.g., statistics). To create opinion statements related to the talking point, we then adopted the Wikipedia neutral point of view (NPOV) criteria[3] to identify those opinionated arguments made by participants on DEBATE.ORG on this point that violate the NPOV criteria. In the end, we obtained a set of 12 statements, and Table 2.2 shows some example of the statements.

Next, we posted a human intelligence task (HIT) on Amazon Mechanical Turk (MTurk) to recruit workers to evaluate this set of statements. Our HIT was open to U.S. workers only. Each worker was asked to review all 12 statements in the HIT. For each statement, the worker was asked to decide whether it is a "factual statement," regardless of whether they think it is accurate or not, or an "opinion statement," regardless of whether they agree with it or not. We also included the "I don't know" (IDK) option in each task, in case workers are not sure about their answer. We further inserted an attention check question in the HIT, in which workers were instructed to select a pre-defined option. Finally, we asked workers to self-report their political stance on a 7-point Likert scale, with 1 representing very liberal, 4 representing neutral, and 7 representing very conservative.

In total, 110 workers completed our HIT and passed the attention check, among whom 57 were leaning liberal, 42 were leaning conservative, and 11 were neutral. Out of $110 \times 12 = 1320$ labels generated by these workers, we obtained 107 IDK labels (i.e., 8.1% of the labels are IDK)[4]. We considered the IDK labels as absent and did not include them in our further analyses.

2.2.2 Understanding Worker's Confirmation Bias

We start by understanding whether workers actually exhibited any confirmation bias when labeling factual and opinion statements in our HIT. To characterize the values that

[3]↑ https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

[4]↑The mean and median number of tasks that a worker selected the IDK label was 0.97 and 0, respectively, and the number of IDK labels selected by individual workers showed a long-tail distribution. We also found that workers who self-reported as neutral tended to select the IDK label more frequently than workers who self-reported as leaning liberal or conservative (the percentage of IDK labels given among all labels generated by the liberal, neutral, and conservative workers were 6.0%, 16.7%, and 8.7%, respectively).

different statements express, we recruited another 47 MTurk workers to review these statements and determine that in a debate about gun control, whether the statement would be more likely used by a person holding liberal views or conservative views as their argument. For each statement, we took the majority answer from MTurk workers as the values of the statement (see Table 2.2 for examples).

Similar as the method used in [20], we focused on analyzing worker's incorrect annotations to quantify the worker's confirmation bias. Specifically, for a worker i, we categorized her mistakes into four types and computed the the misclassification rates correspondingly:

- $\mathbf{ER_L^{fct\to opn}}$(i): among all factual statements with liberal values, the fraction of statements that worker i incorrectly labeled as opinion statements
- $\mathbf{ER_C^{fct\to opn}}$(i): among all factual statements with conservative values, the fraction of statements that worker i incorrectly labeled as opinion statements
- $\mathbf{ER_L^{opn\to fct}}$(i): among all opinion statements with liberal values, the fraction of statements that worker i incorrectly labeled as factual statements
- $\mathbf{ER_C^{opn\to fct}}$(i): among all opinion statements with conservative values, the fraction of statements that worker i incorrectly labeled as factual statements

If workers were indeed influenced by confirmation bias during their annotation process, we expect that for workers holding liberal (conservative) views, they have larger (smaller) $\mathrm{ER_C^{fct\to opn}}$ and $\mathrm{ER_L^{opn\to fct}}$, but smaller (larger) $\mathrm{ER_L^{fct\to opn}}$ and $\mathrm{ER_C^{opn\to fct}}$. Therefore, we define the following metric to represent the bias of worker i:

$$\begin{aligned}\mathrm{bias_i} = \mathrm{zscore}(\mathrm{ER_C^{fct\to opn}}(\mathrm{i})) + \mathrm{zscore}(\mathrm{ER_L^{opn\to fct}}(\mathrm{i}))\\ - \mathrm{zscore}(\mathrm{ER_L^{fct\to opn}}(\mathrm{i})) - \mathrm{zscore}(\mathrm{ER_C^{opn\to fct}}(\mathrm{i}))\end{aligned} \tag{2.2}$$

where $\mathrm{zscore}(\cdot)$ represents the function standardizing the misclassification rates within each category (i.e., $\mathrm{zscore}(x) = \frac{x-\bar{x}}{\sigma}$). Intuitively, the larger $\mathrm{bias_i}$ is, the more worker i favors information with liberal values.

Table 2.3. Considering only the N most difficult tasks (i.e., the top N statements with the lowest average labeling accuracy), the negative correlation between a worker's stance and bias score is significant.

Top N	Correlation coefficient (ρ)	p-value
1	–0.192	0.044
2	–0.243	0.011
3	–0.255	0.007
4	–0.182	0.057
5	–0.221	0.021

To see whether workers indeed showed the tendency to favor information that was consistent with their own values, we look into the relationship between workers' self-reported political stance and the computed bias scores on them. Considering workers' annotations on all 12 statements, the average bias scores for liberal, neutral, and conservative workers are 0.18, -0.47, and -0.12, respectively, and we find a negative, albeit non-significant, correlation between workers' stance and their bias scores (Pearson correlation coefficient $\rho = -0.086; p = 0.374$). This means that compared to neutral and conservative workers, liberal workers indeed favored information with liberal values slightly more, implying some degree of confirmation bias. More interestingly, as shown in Table 2.3, when we restrict our attention to the subset of statements that are most difficult for workers (i.e., worker's average accuracy on the statement was the lowest among all 12 statements), we see significant negative associations between worker's stance and bias score, suggesting that workers might be influenced by their confirmation bias to the largest degree on the difficult tasks.

2.2.3 Evaluating Label Aggregation Performance

We now move on to compare the effectiveness of the proposed algorithm in accurately inferring the ground-truth labels of different tasks against baseline methods. In particular, we consider the following seven baseline methods:

- **Majority vote (MV)**: the ground-truth label of a task is the majority vote over all labels on that task.

- **GLAD**: the algorithm proposed in [33] which assumes a worker's label on a task is affected by both the worker's skill and the task difficulty.
- **Multi**: the algorithm proposed in [34] that models each annotator as a multidimensional entity to capture the worker's diverse skills on various latent topics.
- **VI-BP**: the algorithm proposed in [35] that transforms the label aggregation problem into a standard inference problem in graphical models and solves it via belief propagation.
- **Minimax (MM)**: the algorithm proposed in [49] which assumes a separate probabilistic distribution for each worker-task pair, and uses a minimax entropy method to infer ground-truth labels for each task.
- **ZenCrowd (ZC)**: the algorithm proposed in [36] which iteratively estimates worker reliability, removes unreliable workers, and infers ground-truth labels.
- **CBCC**: the algorithm proposed in [50] which assumes communities exist within workers and those workers belonging to the same community share similar misclassification pattern.

Note that *none* of these baseline algorithms explicitly accounts for worker's confirmation bias when aggregating crowdsourced labels. We implemented these baseline algorithms using the open-sourced code repository provided by Zheng *et al.* [37]. We further implemented the proposed bias-aware label aggregation algorithm, and we terminated the EM-based inference after convergence or 1000 iterations, whichever was reached earlier[5]. In total, we implemented eight different label aggregation algorithms.

[5]↑To account for the impact of parameter initialization on the performance of the algorithm, we deployed an empirically effective heuristic to restart the EM algorithm. We ran EM for three times. For all three runs, we adopted a relatively uninformative initialization for p_{i}, $\boldsymbol{\pi}$, and a ($p_{\mathrm{i}} = 0.5, \boldsymbol{\pi} = 0.5$, and $a = 2$). Then, in the first EM, we initialized $c_{\mathrm{i}} = 0.5$ and initialized all statements' values from one extreme (e.g., $s_{\mathrm{j}} = 1$), hoping that this run of EM would return an accurate ordering of c_{i}. Then, in the second EM, we initialized $s_{\mathrm{j}} = 0.5$ and $c_{\mathrm{i}} = 1$, hoping to get an accurate ordering of s_{j}. In the third EM, we initialized c_{i} (s_{j}) using the final c_{i} (s_{j}) values from the first (second) EM. In the end, we report the inference results from the EM run that gives the highest likelihood of the data.

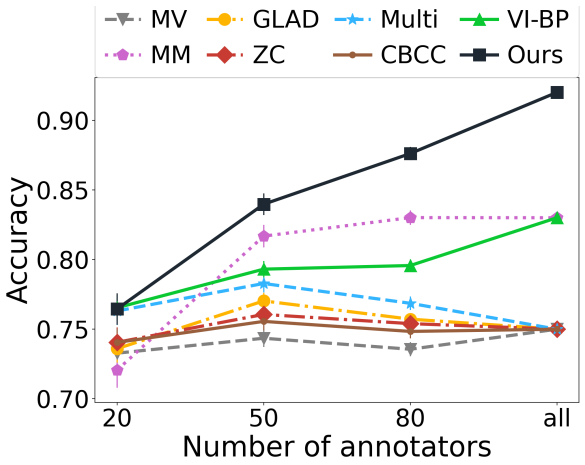


Figure 2.2. Comparing the performance of different algorithms in accurately inferring ground-truth labels on the real-world dataset, as the number of annotators increases. Error bars represent the standard errors of the mean. Note that uncertainty in inference accuracy due to random sampling does not exist when all worker's annotations are used in the inference (i.e., "all" in the x-axis).

We applied all these eight algorithms on the annotations that we collected from MTurk workers for differentiating factual and opinion statements, and inferred the ground-truth label for each statement. Figure 2.2 compares the accuracy of the inferred labels when using different algorithms. In addition to making inference using the entire set of annotations from all 110 workers, to see how the accuracy of the inference varies with the number of annotators, we also randomly sampled annotations from K ($K \in \{20, 50, 80\}$) workers and inferred the ground-truth label for each statement using only the subset of annotations provided by these K workers. For each K, we repeated the random sampling process for 100 times, and the average accuracy of the inferred labels across 100 trials is presented in Figure 2.2 for each algorithm. Clearly, we find that by taking worker's confirmation bias into consideration, our proposed label aggregation algorithm almost always achieves higher inference accuracy than all baseline algorithms, and its advantage over baseline algorithms becomes more salient as the number of annotators increases.

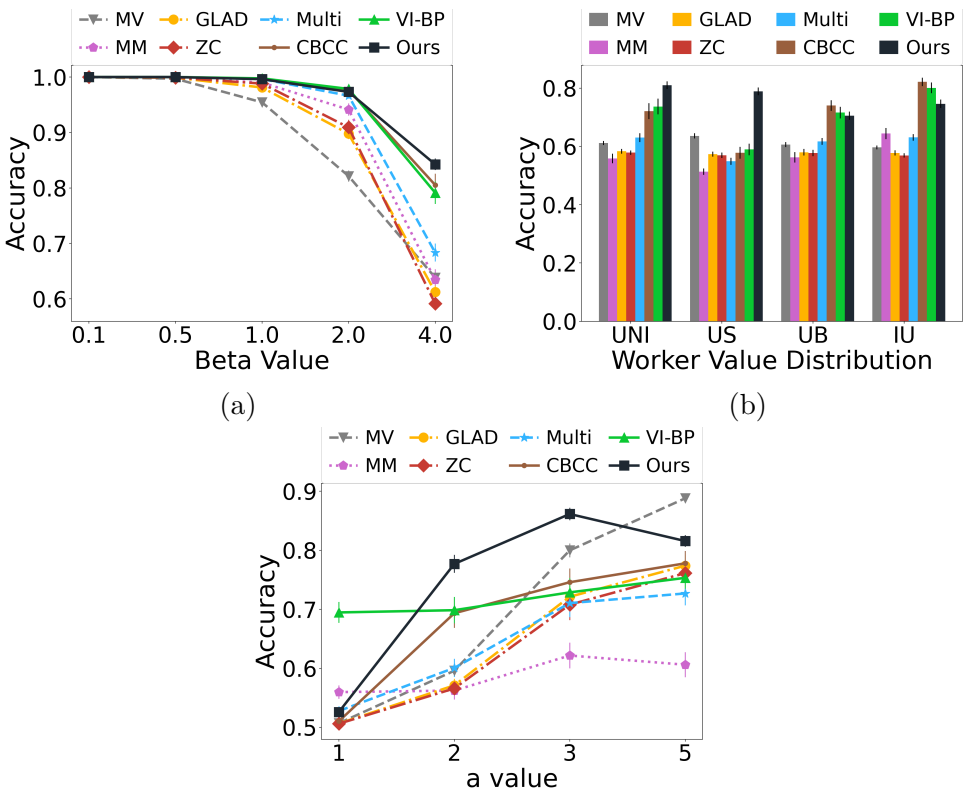


Figure 2.3. Comparing the performance of different algorithms in accurately inferring ground-truth labels on synthetic datasets as the degree that workers suffer from confirmation bias changes (2.3a), the distribution of worker's values changes (2.3b), or the tendency for workers to provide the preferable label changes (2.3c). Error bars represent the standard errors of the mean.

2.3 Simulation

Finally, we conduct simulations on synthetic datasets to explore when the proposed bias-aware label aggregation algorithm shows the largest advantages over baseline algorithms.

2.3.1 The Impact of Confirmation Bias Degree.

First, we examine that compared to baseline algorithms, how the performance of the proposed algorithm changes with the degree to which crowd workers are subject to confirmation bias. To do so, we generated synthetic datasets of worker annotations following the label generation model that we describe in Section 2.1. In particular, for each dataset, we randomly created $M = 100$ tasks. For each task, the values it took was drawn uniformly

randomly between 0 and 1 (i.e., $s_j \sim U[0,1]$), and with 50% chance it had the preferable label (i.e., $z_j \sim$ Bernoulli(0.5)). We then simulated a group of $N = 25$ workers by setting $a = 2$, sampling each worker's values uniformly randomly between 0 and 1 (i.e., $c_i \sim U[0,1]$), and setting $p_i \sim \text{Beta}(1, \beta)$. Intuitively, the larger the value of β is, the more crowd workers suffer from confirmation bias.

To simulate different degrees of confirmation bias, we considered five different values of β: 0.1, 0.5, 1, 2, 4. For each value of β, we generated 50 synthetic datasets by simulating worker's annotation on each task according to Eqn. 2.1. Given a specific dataset, we next used all eight label aggregation algorithms to infer the ground-truth label for each task. Figure 2.3a shows how the inference accuracy of different algorithms, averaged over the 50 datasets, changes with β. It is clear that as crowd workers suffer more from confirmation bias (i.e., β increases), while the inference accuracy of all algorithms decreases, the advantage of our bias-aware algorithm over the baseline algorithms becomes larger. In other words, using the proposed algorithm to aggregate crowd-generated annotations is especially helpful when crowd workers exhibit a higher level of confirmation bias.

2.3.2 The Impact of the Distribution of Worker's Values.

We next explore how the distribution of crowd workers' own values affects the performance comparison between different label aggregation algorithms. We again simulated annotations from $N = 25$ workers on $M = 100$ tasks. For each task, $s_j \sim U[0,1]$ and $z_j \sim$ Bernoulli(0.5), while for each worker, $a = 2$ and $p_i \sim \text{Beta}(1, 5)$. For each worker's values c_i, we considered four types of distributions from which it could be randomly drawn: (1) *Uniform (UNI)*: $c_i \sim \text{Beta}(1, 1)$, reflecting the case that crowd workers' values are uniformly spread over the spectrum; (2) *U-shape (US)*: $c_i \sim \text{Beta}(0.5, 0.5)$, reflecting the case that crowd workers are polarized and tend to hold divergent and extreme values; (3) *Unbalanced (UB)*: $c_i \sim \text{Beta}(1, 2)$, reflecting the case that most workers lean towards one extreme on the spectrum of values; and (4) *Inverse-U shape (IU)*: $c_i \sim \text{Beta}(2, 2)$, reflecting the case that most workers lean towards the middle of the spectrum of values. Again, given a specific values distribution, we simulated 50 synthetic worker annotation datasets, and the

average inference accuracy of different label aggregation algorithms is shown in Figure 2.3b. Here, we observe that the advantage of our bias-aware algorithm is particularly salient when workers' values are widely dispersed or even polarized. When workers' values lean towards one extreme or the middle of the spectrum—that is, when most workers' values are somewhat similar—the performance of the proposed algorithm is on par with the best-performing baseline algorithms (i.e., CBCC and VI-BP).

2.3.3 The Impact of Base Rate of the Preferable Label.

Lastly, we look into how worker's tendency of providing the preferable label (i.e., worker's "positive bias") changes the performance of various algorithms. We simulated 50 datasets, with each dataset containing $N = 25$ workers and $M = 100$ tasks. Further, we set $s_{\mathrm{j}} \sim U[0,1]$, $z_{\mathrm{j}} \sim \text{Bernoulli}(0.5)$, $c_{\mathrm{i}} \sim \text{Beta}(1,1)$, and $p_{\mathrm{i}} \sim \text{Beta}(1,5)$. We then varied $a \in \{1,2,3,5\}$, and Figure 2.3c presents the average inference accuracy of different label aggregation algorithms. Interestingly, we find the proposed algorithm has the largest advantage over baseline algorithms when workers have a moderate level of base rate of providing the preferable label. When workers have very high base rates to provide the preferable label (e.g., $a = 1$), the proposed algorithm performs worse than the VI-BP algorithm. On the other hand, when workers are unlikely to provide the preferable label (e.g., $a = 5$), while the proposed algorithm outperforms many baselines, a simple majority vote can be the most effective aggregation strategy if the distribution of worker's values is balanced.

2.4 Conclusion

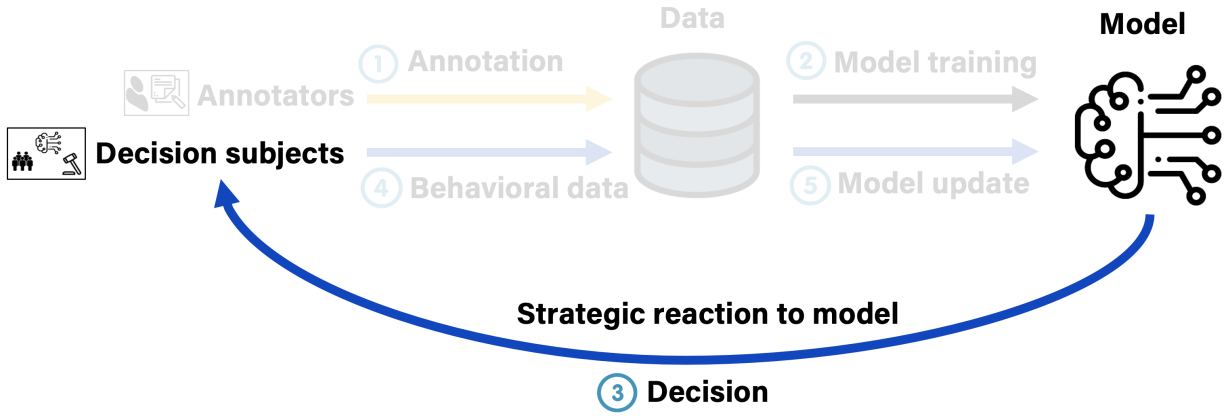
Crowdsourcing has become a prevalent tool for gathering data from humans. As humans are often subject to various types of biases, the challenge of how to carefully process the crowdsourced data to minimize the negative impact that people's biases bring to data quality becomes pressing. In this chapter, we focus on confirmation bias, a particular type of cognitive bias, and propose a new label aggregation algorithm based on a quantitative model which characterizes how crowd workers are influenced by their confirmation bias in their an-

notations. The evaluation results on both real-world data and synthetic data demonstrate the effectiveness of our proposed method.

2.5 Acknowledgments

The work presented in this chapter was conducted with Ming Yin. The entire content of this chapter is published as "Accounting for Confirmation Bias in Crowdsourced Label Aggregation" in the Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI '21) [15]. We sincerely appreciate the anonymous reviewers for their valuable feedback, which helped to improve the quality of this research. We also thank the support of Purdue University and the crowd workers who participated in the study.

3. INVESTIGATING DECISION SUBJECTS' REPEATED INTERACTIONS WITH ML MODELS



Exploring the "Decision," stage 3, of the machine learning pipeline. The complete figure is presented in Figure 1.1 of Chapter 1.

We now turn our attention to "Decision," the third stage of the machine learning (ML) pipeline, as depicted in Figure 1.1 of Chapter 1. Considering the wide array of contexts where ML models are used to automate decision-making in high-stake domains, such as loan lending [51], hiring [52], and immigration policing [53], the importance of fairness becomes evident. Unfortunately, many of the ML-based decision systems that are deployed inherit pre-existing biases from the datasets used in their training. This leads them to treat individuals from different socio-demographic groups unfairly. For example, an ML model reviewing credit card applications was found to exhibit gender bias, granting a male applicant 20 times more credit limit compared to a female applicant with the same qualifications [54]. In another case, an ML model widely used in US hospitals to allocate healthcare to patients was discovered to systematically discriminate against African Americans [55, 56].

The possibility of ML-based decision systems to behave unfairly has sparked great interests among researchers to investigate various methods for ensuring fairness in such systems. While earlier works tackle this challenge mostly by adjusting the training data, processes, and outputs of the ML systems [57–60], more recently, an increasing number of studies start to take a more human-centered view by probing deeper into what does a "fair" ML-based decision system mean to *people*. For example, user interfaces have been designed to elicit diverse subjective fairness notions from different stakeholders [61]. Experimental studies have been conducted to understand people's preferences over multiple fairness definitions that potentially compete with one another [62–64]. Frameworks have also been proposed to learn context-aware fairness notions from humans' situated fairness judgements [65].

As ML-based decision systems bring about real-world consequences to people's lives, another important line of research regarding fairness of these systems is to examine what factors affect the fairness perceptions of *decision subjects* of an ML-based decision system (i.e., those people about whom the decisions are made by the system), and how. To this end, Wang *et al.* [66] show that in the one-shot interaction with an ML-based decision system, decision subjects perceive the system to be fairer both when the system is not biased against any specific group (i.e., the system is "fair" across groups), and when the system's decisions on them are in their favor. However, in practice, decision subjects often can repeatedly interact with an ML-based decision system and strategically respond to the system by, for

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

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

Our first study involved two experimental treatments by varying decision outcomes of the bank's ML system across loan applicants of different groups. Through this study, we found that when a decision subject interacts with an ML-based decision system repeatedly, both her fairness perceptions and her retention in the system is significantly affected by *whether the system is in favor of her own group*, rather than whether the system treats people of different groups in an unbiased way. Decision subjects with higher qualification levels also had significantly higher retention in the system, but their fairness perceptions of the system did not change. More interestingly, we noted that the decision subject's qualification level moderates the impacts on her fairness perceptions and retention in the ML system that

are brought up by the system's decision outcomes. As for the decision subject's sensitivity to fairness, we observed that subjects who value fairness more tended to perceive ML-based decision systems as more unfair and be less willing to participate in these systems, but we did not find any significant moderating effects associated with the subject's fairness sensitivity.

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

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

3.1 Related Work

There is a growing body of work on understanding how humans adopt, interact with, and trust the ML-based decision-making systems [67–77]. Among many other factors, whether the ML system is "fair" is deemed as a critical factor that will affect people's perceptions of and reactions to the system. However, while there are numerous fairness definitions being proposed in the computer science literature [78–84], there is no clear agreement over a particular definition [62], and fairness requirements could be highly context-dependent [85]. Therefore, many empirical studies have been designed to solicit human preferences over different fairness definitions for a variety of decision-making contexts [62–64]. For example, for loan lending decisions, Saxena *et al.* [62] compared three fairness definitions and found that the calibrated fairness definition tends to be preferred by laypeople.

More recently, a more human-centered perspective has been taken to understand the fairness of ML-based decision systems. That is, instead of searching for a single, objective definition of fairness, fairness is increasingly being treated as a subjective concept, and various studies have been carried out to examine the range of factors that may affect people's fairness perceptions of an ML system. For instance, Hannan *et al.* [86] found that in resource allocation scenarios, people's fairness perceptions are influenced by what resource is being allocated, who allocates the resources, and sometimes even how the questions regarding fairness perceptions are asked. Other key influencing factors include whether and how the system's decisions are explained [87–89], the ways that the ML system's decisions are presented and visualized [90], and people's personal experience related to the algorithmic decision making scenario [91]. Researchers have also explored fairness perceptions of an ML-based decision systems from different stakeholders' points of view. For example, by *independently* controlling the ML system's decisions on individuals (i.e., favorable vs. unfavorable) and the system's treatment across groups (i.e., biased vs. unbiased), Wang *et al.* [66] showed that decision subjects' fairness perceptions of an ML system are predominately affected by whether the system makes a decision that is in their favor, although holding all else equal, decision subjects also perceive a system that exhibits unbiased treatment across different groups as fairer. On the other hand, from system developers' perspectives, factors used and processes involved in algorithmic decision making of an ML system are essential for them to judge the system's fairness level [92].

Compared to the prior work, in this chapter, we aim to re-examine decision subjects' fairness perceptions and retention in ML-based decision systems as they interact with these systems *repeatedly*. This perspective of repeated/long-term interactions has been taken in more theoretical examinations of fairness in ML, in which researchers often argue that an ML model that makes one-shot fair decisions by enforcing static fairness constraints may not lead to long-term well-being of those groups it aims to protect. This is because the decisions that an ML model makes on people, in the long run, can also reshape people, including changing the qualification distributions of different groups [10, 22], changing the group representation over time [11], affecting people's willingness to invest in their qualification [93], and causing different levels of precarity to people [94]. Our work complements this theoretical line of work

from two perspectives: First, we extend the discussions from decision subjects' well-being in repeated interactions to their fairness perceptions in repeated interactions, and we suspect these perceptions may also be different from those perceptions in one-shot interactions. For example, the impact of an ML model being unfair across groups on decision subjects' fairness perceptions of the model may either be strengthened in the repeated interactions due to people's repeated exposure to the model's biased treatment, or be weakened as some people see the possibility to "exploit" such biased treatment to optimize their own utility. Second, as many theoretical studies use simulated models to capture the long-term user dynamics [11, 26], our investigation into decision subjects' retention in repeated interactions with an ML model could provide empirical evidence for characterizing the user dynamics more realistically.

3.2 Study 1

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

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

3.2.1 Experimental Design

3.2.1.1 Experimental Tasks.

We recruited human subjects to play a game in our experiment, in which each subject was asked to play as a small business owner, and would interact with a bank repeatedly by applying loans from it to support her business. Subjects were told that this bank uses an ML model to make lending decisions, and they could decide whether to keep applying for loans from this bank of their own volition. The main interface of this game is shown in Figure 3.1. More specifically, upon arrival at the game, each subject was assigned with a loan applicant profile that represents her *throughout* this game (Figure 3.1A), which included 5 features:

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

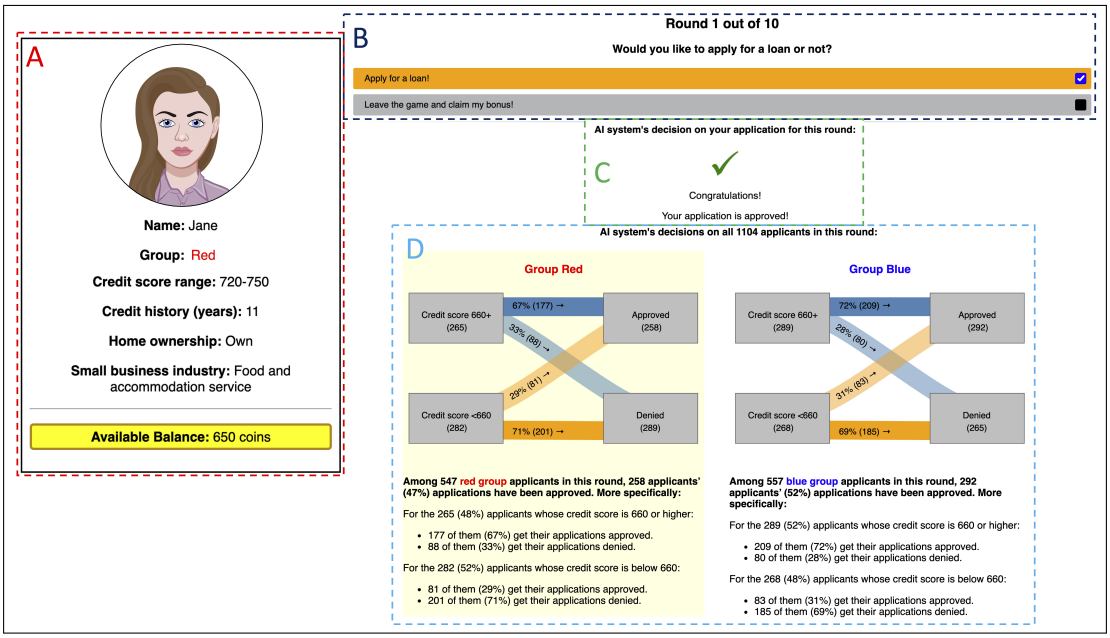


Figure 3.1. The main interface of the game. The loan applicant profile assigned to the subject is presented on the interface (Part A). In each round, the subject needed to decide whether to continue to apply loans from the bank (Part B). If the subject decided to apply for a loan in one round, the ML model's lending decision on the subject would be revealed to her (Part C), and the subject could also get a summary of the ML model's decisions on all applicants in this round (Part D).

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

Beyond the loan applicant profile, the subject was also given an "account" with an initial balance of 600 "coins." The subject then needed to interact with the bank for *at most* 10 rounds. In each round, the subject was asked to decide whether she'd like to continue to apply for a loan from the bank (Figure 3.1B). If yes, 50 coins would be deducted from her

account as the application fee. The bank's lending decision, which was decided by the ML model, would then be revealed to her (Figure 3.1C)—if the ML model approved her loan application, the subject would receive a reward of 100 coins; otherwise the subject would receive nothing. The subject would also be able to view the summary information of the ML model's decisions on *all* applicants in this round—broken down by the applicant's group—before moving on to the next round (Figure 3.1D; see Section 3.2.1.2 for details). However, in one round, if the subject decided not to continue to apply for a loan from the bank, she would immediately leave the game and be re-directed to the end of the experiment.

Overall, this game was designed to closely reflect the real-world scenario that decision subjects can freely decide whether they are willing to "stay in the system" to take part in ML-based decision making (i.e., whether they want to be subject to a particular ML system's decisions) as they repeatedly interact with the system. These participation decisions are often made as the decision subjects—with some knowledge of their own qualification levels—observe the ML system's decisions on themselves and on others over time. Moreover, while the decision to participate is often costly (i.e., the participate decision in the game is associated with a "fee" of 50 coins), decision subjects could benefit from such participation when the ML system grants a favorable decision to them (i.e., a favorable decision is associated with a "reward" of 100 coins in the game). Note that in this game, we assume the ML model is not updated over time, and the model's decisions do not result in changes in the decision subjects' qualification levels that are significant enough to affect the model's future decisions on them. Understanding decision subjects' fairness perceptions and retention in repeated interactions with the ML-based decision systems after relaxing these assumptions will be an interesting future work.

Table 3.1. The decision matrices used by the ML model in different treatments of Study 1 on loan applicants of different groups. Number in each cell represents the probability for the ML model to approve/deny an applicant when the applicant's credit score falls into the range as specified in the corresponding row.

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

(a) Fair model: Red/Blue group

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

(b) Unfair model: Red group

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

(c) Unfair model: Blue group

3.2.1.2 Experimental Treatments.

We created two treatments by varying properties of the bank's ML model:

- **"Fair" model**: In this treatment, the bank's ML model treats applicants from different groups *equally*. Specifically, this ML model makes a stochastic lending decision based on a "decision matrix," as shown in Table 3.1a. According to this matrix, *regardless of the group identity of the applicant*, the chance for this ML model to approve the loan for an applicant with a high credit score (i.e., score ≥ 660) is 70%, while the chance to approve the loan for an applicant with a low credit score (i.e., score < 660) is 30%.
- **"Unfair" model**: In this treatment, the bank's ML model is unfair as it is systematically *in favor of applicants from the red group*. Specifically, for applicants of the red group, the ML model makes its stochastic lending decision based on the decision matrix as shown in Table 3.1b—the probabilities for the ML model to approve the loan for a red group applicant with a high or low credit score are 90% or 40%, respectively. In contrast, for applicants of the blue group, the ML model makes its stochastic lending decision based on

the matrix as shown in Table 3.1c—the probabilities for the ML model to approve the loan for a blue group applicant with a high or low credit score are 50% or 20%, respectively.

Specifically, in one round of the game, if the subject decided to apply for a loan from the bank, the bank's lending decision on her would be made by the ML model of the subject's *assigned treatment*—Given the decision matrices of the ML model, the subject's group identity and credit score range would be used to determine the probability of loan approval, and then the ML model would randomly realize its lending decision on the subject according to this probability. Moreover, to allow the subject to get a sense of the ML model's overall decisions on applicants of different groups, we told the subject that many other people had also applied loans from the bank in the same time period, and we showed the summary information of the ML model's decisions on *all* these applicants to the subject[1]. In particular, in each round, we simulated another N loan applicants where N is an integer uniformly randomly drawn from the interval of $[1000, 1200]$. Again, for each of these N applicants, we randomly generated her profile (i.e., each feature value was uniformly randomly sampled from the set of possible values) and determined the lending decision for her using the ML model of the subject's *assigned treatment*. Finally, we displayed the ML model's decisions on all these N applicants to the subject through flowcharts (Figure 3.1D)[2], with decisions on red group applicants and blue group applicants shown in separate flowcharts. We also provided textual explanations along with the flowcharts to help subjects better interpret information in the flowcharts.

We note that if we consider each loan applicant's qualification level (i.e., their credit score range) as the "ground truth" for the lending decision, the decision matrices in Table 3.1 are effectively the ML models' *confusion matrices*. Since each loan applicant's profile was generated uniformly randomly, when considering the ML model's *overall* performance regardless of the group identity of the decision subjects, the fair ML model and the unfair ML model had exactly the *same* expected performance with respect to a range of metrics

[1]↑In reality, decision subjects may get access to such summary information of the ML model's decisions on decision subjects of different groups due to media coverage or scientific investigation of the ML model, such as [95, 96].

[2]↑We chose to use flowcharts since previous study suggested that flowcharts could best support laypeople's understanding of the performance of algorithmic models [97].

such as accuracy, positive prediction rate (PPR), false positive rate (FPR), and false negative rate (FNR)[3]. However, while the fair ML model treats decision subjects of different groups equally, the unfair model is in favor of decision subjects from the red group according to all these metrics—it had a higher accuracy, a higher PPR, a higher FPR, and a lower FNR, on red group applicants.

3.2.1.3 Experimental Procedure.

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

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

Once qualified, the subject would be *randomly* assigned to one of the two experimental treatments and start to play the game. As explained in Section 3.2.1.1, in each round of the game, the subject decided whether to continue to apply for a loan from the bank. If yes, the bank's lending decision on her, as well as on all applicants in this round, would then be revealed to the subject[5]. The subject's account balance would also be updated based on the lending decision she received. To make sure that the subject at least had some interaction with the ML system, we required each subject to apply for a loan in the *first* round. After

[3]↑We considered the decision of approving the loan as the positive decision. On expectation, both the fair model and the unfair model had an accuracy of 70%, a positive prediction rate of 50%, a FPR of 30%, and a FNR of 30%, across all decision subjects.

[4]↑All of our experiments were approved by the IRB of the authors' institution.

[5]↑On the interface, we used a light yellow background to highlight the ML model's decisions on applicants coming from the subject's *own* group to allow subjects better contrast the model's performance on different groups.

that, the subject could continue to apply for loans for a maximum of 9 more rounds, but she could also decide not to apply for loans anymore in any round, which would immediately redirect the subject to the end of the game.

At the end of the game, the subject was asked to answer a few exit-survey questions. In particular, the subject first reported some demographic information (e.g., gender, age). Then, the subject was asked to indicate how fair she perceived the bank's ML system was—We adapted a set of six fairness perception statements from those used in [66] (e.g., "The bank's ML system is fair to manage loan applications."), and the subject evaluated how much she agreed with each statement on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). The subject could also comment on her perceived fairness of the ML system via free text. To quantify the subject's sensitivity to fairness (i.e., to what extent the subject values fairness), we created another set of 4 statements as below:

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

Finally, we conjectured that a decision subject's fairness perceptions and retention in an ML-based decision system might be influenced by the subject's *risk attitude* (i.e., how much the subject is willing to take risks). We thus included another set of statements created in previous studies [98] to measure the subject's risk attitude (e.g., "I prefer friends who are exciting and unpredictable."). Again, for statements related to both fairness sensitivity and risk attitude, the subject rated how much she agreed with each statement on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

After completing the exit-survey, we would reveal to the subject the amount of bonus payment she received in this game—We converted the amount left in the subject's account to

[6]↑We reversed the rating for negative statements.

her bonus payment using a rate of 500 coins to \$1.5. Thus, together with the base payment of \$1.5 of this HIT, the subject could earn a maximum of \$4.8 from this game[7].

3.2.2 Analysis Methods

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

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

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

[7]↑The median value of time that subjects spent on our HIT was 20 minutes, and the median payment to subjects was \$3.3, leading to an effective hourly wage of \$9.9.

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

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

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

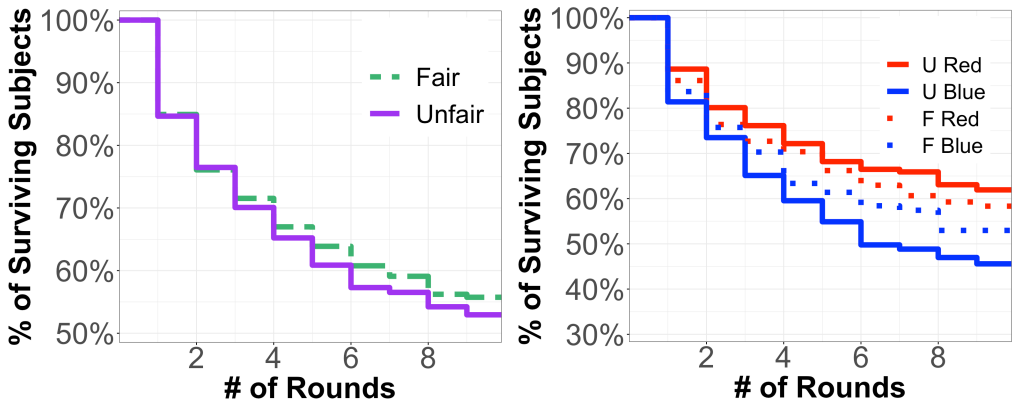
3.2.3 Results

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

3.2.3.1 RQ1: The impacts of the ML system's decision outcomes

We start by examining whether the ML system's fairness level across groups influences decision subjects' fairness perceptions and retention. Results are shown in Table 3.2 (Models 1 and 3). We find that whether the ML system treats decision subjects of different groups equally does *not* significantly affect either decision subjects' perceived levels of fairness of the ML system (Model 1) or their willingness to participate in the system (Model 3). Indeed, as shown in Figure 3.2a, regardless of whether the ML system treats decision subjects of

different groups in a biased way or not, overall, subjects in the two treatments departed from the ML system at a similar rate.



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

Figure 3.2. Survival curves showing the fraction of subjects who continued to apply for a loan from the bank after the X-th round in the two experimental treatments. In Figure 3.2b, "U" ("F") represents the treatment with unfair (fair) model.

In contrast, whether the ML system is in favor of the group that the decision subject belongs to affects both the subject's fairness perception and retention (Models 2 and 4). When a decision subject's group is favored by the ML system in receiving the preferable decision, the decision subject seems to perceive the model as fairer (Model 2, though not significant), and has a marginally higher level of willingness to stay in the ML-based decision system (Model 4, $p = 0.075$)—As shown in Figure 3.2b, subjects who were assigned to the treatment with the unfair ML model and the red group (i.e., the group being favored by the unfair model) tended to apply for loans from the bank for more rounds. However, when a decision subject's group is disfavored by the ML system, the decision subject significantly decreases her perceived fairness level of the system (Model 2, $p = 0.039$), as well as her retention in the system (Model 4, $p = 0.023$; also see the blue solid line in Figure 3.2b). Notably, we also find that decision subjects' risk attitude is significantly correlated with their fairness perceptions and retention in the ML system—the more risk-seeking the decision subject is, the fairer she perceives the ML system to be and the more she is willing to stay in the system.

Table 3.2. Regression models predicting decision subjects' perceived levels of fairness and retention in the ML system, based on properties of the ML system's decision outcomes. Coefficients and standard errors are reported. †, *, and *** represent significance levels of 0.1, 0.05, and 0.001, respectively.

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

3.2.3.2 RQ2: The role of decision subjects' qualification levels.

To first see how a decision subject's qualification level, *by itself*, correlates with her fairness perceptions and retention in an ML-based decision system, we simply add the qualification level as a covariate into each of the four regression models that we have constructed for **RQ1**. Results are reported as Models 1, 3, 5, 7 in Table 3.3, which consistently indicate that a decision subject with a higher qualification level is significantly *more* likely to participate in the ML-based decision system ($p < 0.001$ for both Models 5 and 7), but her perceived fairness level of the ML system is not significantly different.

Next, we explore whether a decision subject's qualification level *moderates* the impacts of the ML system's decision outcomes on the subject's fairness perceptions and retention. We do so by including the interaction term(s) between the subject's qualification level and the independent variable(s) representing properties of the ML system's decision outcomes into the regression models. Results are reported as Models 2, 4, 6, 8 in Table 3.3. For example, consider the impacts of the ML system's fairness level across groups on subjects—In Model 2, we find that while decision subjects with lower qualification levels perceive the unfair ML system to be marginally fairer than the fair ML system ($p = 0.099$ for "Biased treatment"),

decision subjects with higher qualification levels significantly decrease their perceived fairness level of the ML system when the ML system is unfair across groups (Model 2, $p = 0.027$ for the interaction). However, such decrease does not result in any significant different change in highly-qualified subjects' retention in the ML system as compared to low-qualified subjects (Model 6). When it comes to the impacts of the ML system's tendency to favor a subject's own group on the subject, we detect a significantly negative interaction (Model 4, $p = 0.014$) between the qualification level and the independent variable of "advantaged" when examining subjects' fairness perceptions of the ML system. This means that highly-qualified subjects increase their perceived fairness level of the ML system to a *smaller* degree compared to low-qualified subjects when the ML system favors the group that they belong to. Finally, we detect a marginally negative interaction between the qualification level and independent variable of "disadvantaged" in influencing subjects' retention in the ML system (Model 8, $p = 0.099$), suggesting that highly-qualified subjects decrease their retention in the system to a slightly *larger* extent compared to low-qualified subjects when their own group is placed at the disadvantaged position by the system.

Table 3.3. Regression models predicting decision subjects' perceived levels of fairness and retention in the ML system, based on properties of the ML system's decision outcomes and subjects' qualification levels. Coefficients and standard errors are reported. †, *, **, and *** represent significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.

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

3.2.3.3 RQ3: The role of decision subjects' sensitivity to fairness.

Finally, we examine the role that a decision subject's sensitivity to fairness plays in influencing her fairness perceptions and retention in the ML-based decision system. Similar as our analyses in Section 3.2.3.2, we first include the subject's fairness sensitivity score as a covariate into our regression models, and results are shown in Table 3.4. In all these models, we consistently find that the decision subject's sensitivity to fairness is significantly negatively correlated with the subject's fairness perceptions and retention in the ML system ($p < 0.05$). In other words, the more the decision subject values fairness, the more unfair she perceives the ML system to be, and the less she is willing to participate in the system. We next add the interaction term(s) between the subject's fairness sensitivity score and the independent variable(s) into each of the regression models, but we do not detect any significant interactions in all these models, suggesting that subjects' sensitivity to fairness do not seem to moderate the impacts of the ML system's decision outcomes on subjects' fairness perceptions and retention.

Table 3.4. Regression models predicting decision subjects' perceived levels of fairness and retention in the ML system, based on properties of the ML system's decision outcomes and decision subjects' sensitivity to fairness. Coefficients and standard errors are reported. †, *, and *** represent significance levels of 0.1, 0.05, and 0.001, respectively.

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

3.3 Study 2

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

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

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

3.3.1 Experimental Design

We again recruited human subjects to play the same game of loan application as that in Study 1, with only one key difference—In this experiment, *all subjects were assigned to the red group*. We then created three experimental treatments by varying how the ML system treats the red group applicants in relative to blue group applicants, while *controlling the prospects of the favorable decision for red group applicants*. Specifically, in all three treatments, the ML system makes its stochastic lending decisions to applicants of the red group based on the same decision matrix as shown in Table 3.5a (i.e., approve the loan for a red group applicant with a high or low credit score with a probability of 70% or 30%, respectively). However, the ML system makes lending decisions to the blue group applicants in different treatments based on different decision matrices:

- **Unbiased**: In this treatment, the bank's ML system makes lending decisions on blue group applicants based on the same decision matrix as that for the red group (i.e., Table 3.5a). Thus, the bank places applicants in neither group at the advantaged position.
- **Red advantaged**: In this treatment, the bank's ML system uses the decision matrix as shown in Table 3.5b to make its lending decisions on blue group applicants (i.e., approve the loan for a blue group applicant with a high or low credit score with a probability of 50% or 20%, respectively). Thus, the bank places applicants from the red group at the advantaged position.
- **Red disadvantaged**: In this treatment, the bank's ML system uses the decision matrix as shown in Table 3.5c to make its lending decisions on blue group applicants (i.e., approve the loan for a blue group applicant with a high or low credit score with a probability of 90% or 40%, respectively). Thus, the bank systematically discriminates against applicants from the red group.

Table 3.5. The decision matrices used by the ML model in different treatments of Study 2 on loan applicants of different groups. Number in each cell represents the probability for the ML model to approve/deny an applicant when the applicant's credit score falls into the range as specified in the corresponding row.

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

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

(b) Blue group ("Red advantaged")

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

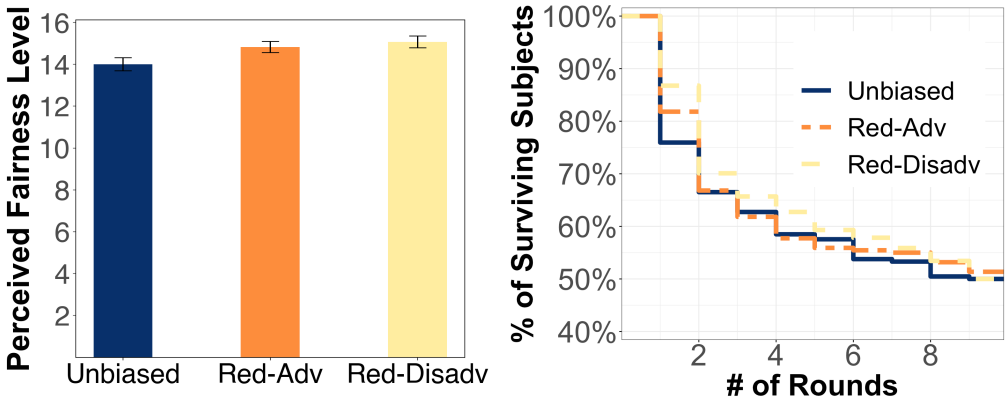
(c) Blue group ("Red disadvantaged")

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

This design may allow us to determine that when the ML system is in favor of a subject's group, whether the subject increases her fairness perceptions and retention in the system simply due to the higher prospect of receiving the favorable decision from the system—If

yes, we expect to see minimal differences across the three treatments on subjects' fairness perceptions and retention; otherwise, we expect to see some differences across treatments.

Again, we posted this experiment as a HIT on MTurk to U.S. workers only, where it had an identical procedure as the experiment in Study 1 (see Section 3.2.1.3) except for the following differences: (1) Workers who had participated in the experiment in Study 1 were *not* allowed to participate in this experiment; (2) workers were randomly assigned to one of the three treatments as defined above.



(a) Perceived Fairness (b) Retention

Figure 3.3. The fairness perceptions and retention for subjects in Study 2. (3.3a): Subject's average perceived levels of fairness of the ML system; error bars represent the standard errors of the mean. (3.3b): Survival curves showing the fraction of subjects who continued to apply for a loan from the bank after the X-th round in the three treatments of Study 2.

3.3.2 Experimental Results

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

group still has any impact on the subject's fairness perceptions and retention, after fixing the subject's prospect of receiving the favorable decision.

Figure 3.3a compares subjects' perceived level of fairness of the ML system across the three treatments, and Figure 3.3b shows the subjects' survival curves in the three treatments. In Figure 3.3b, we observe minimal differences across the three treatments regarding subjects' willingness to stay in the ML system. This seems to suggest that changes in subjects' retention in an ML system is mainly driven by subjects' own prospects of receiving the favorable decision, rather than the relative advantage or disadvantage for subjects' group to receive the favorable decision over the other group. However, in Figure 3.3a, we notice a surprising trend in subjects' perceived fairness perceptions of the ML system—According to our regression results, compared to subjects in the "Unbiased" treatment, subjects in both the "Red advantaged" treatment and the "Red disadvantaged" treatment increased their perceived level of fairness of the ML system, either marginally or significantly (e.g., red advantaged vs. unbiased: $p = 0.066$, red disadvantaged vs. unbiased: $p = 0.011$). This clearly indicates that subjects' fairness perceptions of the ML system are *not* solely determined by their own prospects of receiving the favorable decision. Moreover, the fact that subjects reported the highest perceived level of fairness to the ML system which actually places them at a disadvantaged position in receiving the favorable decision also seems to suggest that there are more factors beyond the relative advantage/disadvantage over the other group that substantially influence subjects' fairness perceptions.

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

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

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

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

Table 3.6. Regression models predicting decision subjects' perceived levels of fairness and retention in the ML system, based on properties of the ML model's decision outcomes, after combining data from Studies 1 and 2. Coefficients and standard errors are reported. †, *, **, and *** represent significance levels of 0.1, 0.05, 0.01, and 0.001, respectively.

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

Results of our regression models are reported in Table 3.6. Regarding subjects' perceived fairness level of the ML system (Models 1–3), we consistently find that an ML system that grants more favorable decisions is perceived as fairer by subjects[10]. Moreover, the relative advantage of a subject's group over the other group in receiving the favorable outcome is also a driver of the subject's increases in her perceived fairness level of an ML system when the system is in favor of her group. In contrast, we find that the subject's prospect in receiving the favorable decision seems to be the sole driver for the changes in her retention in an ML system (Models 4–6).

3.4 Conclusions and Discussions

In this chapter, via two experiments, we examined how decision subjects' fairness perception and retention in an ML-based decision system might be influenced by various factors,

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

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

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

3.4.1 On the impact of the ML system's biased treatment across groups on decision subject's fairness perceptions.

Our finding of that the ML system's biased/unbiased treatment across group does not seem to affect subjects' average level of fairness perceptions of the system in repeated interactions is different from the results reported in Wang *et al.* [66], when decision subjects can only engage in a one-shot interaction with the ML system. One possible reason for this discrepancy is that in our experiments, we did not associate the group identity of loan applicants with specific socio-demographic features like race and gender, which may come with their unique social and historical contexts that can heighten people's sensitivity to inequality across groups. On the other hand, these results could also reflect the differences in decision subjects' fairness perceptions in repeated vs. one-shot interactions with ML systems. In particular, when decision subjects repeatedly interact with a biased ML system, the ones who are placed at the advantaged position by the system may realize that they could "materialize" the advantages by keeping interacting with the system, while those who are placed at the disadvantaged position can actively choose to "boycott" the system. This possibility for decision subjects to strategically interact with the ML-based decision system in the long run may shift the focus of their fairness perceptions to their own utility, rather than the equality across groups. Our investigation in Section 3.2.3.2 on the moderating role

of decision subjects' qualification levels provides further nuanced results—It turns out that highly-qualified subjects will still significantly decrease their perceived fairness levels when the ML system exhibits biased treatment across groups. However, the low-qualified subjects actually perceive the biased ML system to be slightly fairer, and this is mainly caused by those low-qualified subjects who belong to the group that the ML system is in favor of. In other words, it is those decision subjects who do not deserve a favorable decision yet still be favored by the ML system, who substantially increase their fairness perception of the system, despite it being biased.

3.4.2 On the complexity of fairness perceptions.

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

perceptions of ML, and rigorously examine how they, together, influence people's overall fairness perceptions of ML.

3.4.3 Group retention in repeated interactions and implications.

While it is often believed that people's fairness perceptions of an ML system will influence their adoption of it, our study results suggest that the relationship between fairness perceptions and usage of the ML system is not linear, at least for decision subjects. For example, as shown in Table 3.3, while highly-qualified subjects were shown to significantly decrease their fairness perceptions of a biased ML system, they did not significantly decrease their retention in the ML system accordingly. In fact, as shown in Table 3.6, in our studies, subjects' retention in the ML system seems to be mainly driven by their prospects of receiving the favorable decision. This implies that in the long run, the group of decision subjects who have lower prospects of receiving the favorable decision might become increasingly under-represented over time, which can further influence the ML system's performance on the under-represented group as it continues to update its training data. This is true even if ML system is trained with fairness constraints but the factor equalized across groups by the constraints is not the positive prediction rate [11].

3.4.4 Limitations and future work.

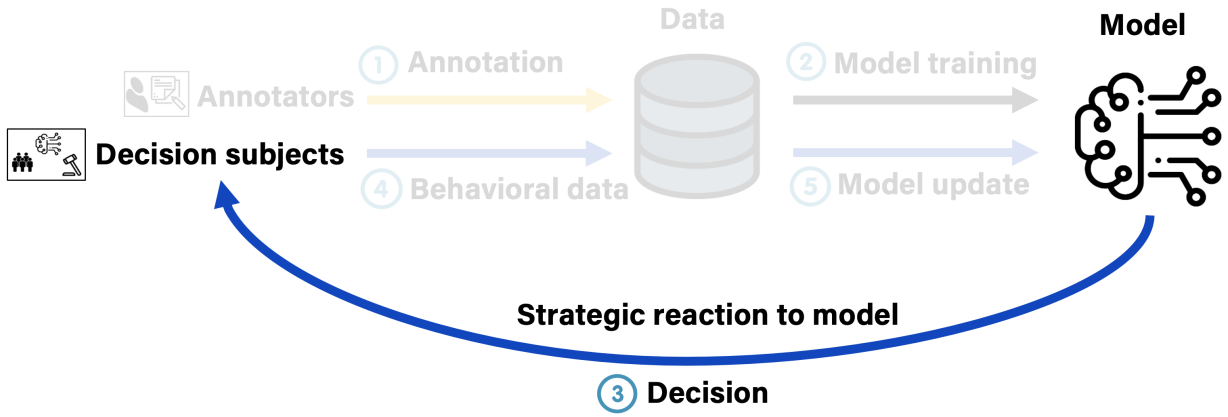
Our study was conducted in the context of ML-based loan lending systems, and we used a specific set of "parameter" values when designing the game in our experiments (e.g., the cost associated with the participation in the ML system and the reward brought up by a favorable decision). Cautions should be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

It's therefore interesting to explore in these cases, how decision subjects' partial knowledge of the ML system's performance affect their fairness perceptions and retention.

3.5 Acknowledgments

The work presented in this chapter was conducted with Ming Yin. The entire content of this chapter is published as "Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems" in the Proceedings of the 5th AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), Oxford, UK [99]. This work is supported in part by the NSF FML program in collaboration with Amazon under grant IIS-2040800. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

4. THE EFFECT OF QUALIFICATION IMPROVEMENT ON DECISION SUBJECTS' REPEATED INTERACTIONS WITH ML MODELS



Continuing to explore the "Decision," stage 3, of the machine learning pipeline. The complete figure is presented in Figure 1.1 of Chapter 1.

As we transition to the fourth chapter, our exploration of the machine learning (ML) pipeline continues on the "Decision" stage—the third stage—as highlighted in Figure 1.1 of Chapter 1. We now shift the focus to scenarios where *decision subjects*—individuals directly impacted by decisions made by ML models—exhibit more pronounced strategic reactions during their interactions with these models. Unlike the previous chapter, where decision subjects' reactions were limited to strategic retention decisions and their "qualifications" for favorable decisions remained static, we now explore contexts in which these subjects actively strive to improve their "qualifications" to obtain more favorable decisions, thereby introducing an additional layer of strategic behavior. We begin by introducing the problem setting and then proceed to elaborate on the unique contributions and perspectives this chapter offers.

The rapid development of ML technologies has made it possible to automate decision making in many domains. However, it has been discovered that ML models often acquire pre-existing biases in the dataset used for their training, resulting in the unfair treatment to individuals from different demographic backgrounds in various applications including criminal justice [5] and recruitment [100]. This increased awareness of fairness issues of ML has led to many recent studies in understanding people's fairness perceptions of and reactions to ML models [12, 17, 63, 101]. These studies look into the perspectives of different stakeholders, among which a key stakeholder is decision subjects. For example, Wang *et al.* [17] found that when decision subjects only interacted with an ML model once, both the model's unbiased treatment across different groups of decision subjects and the model's favorable treatment to them resulted in an increase in the perceived fairness of the ML model.

Meanwhile, a recent line of literature on the long-term dynamics and implications of fairness in ML [10, 11, 22, 23, 25, 26, 93] has drawn the community's attention to the fact that in reality, decision subjects could often interact with an ML model *repeatedly*. During these *long-term interactions* with an ML model, decision subjects are no longer passively accepting the ML model's decisions on them as is. Rather, the ML model's treatment to decision subjects may shape how they ***actively and strategically respond to the ML model***, which may further impact their perceptions of the model in the long run. For example, one strategic response decision subjects could take in their long-term interactions

with ML is to decide whether to continue interacting with the ML model and be subject to its decisions [11, 12]—decision subjects have the freedom to quit using an ML model if they wish so. As another example, decision subjects could also respond to ML decisions on them by investing in effort to improve themselves, hoping for an increased chance of receiving the favorable decision in the future [10, 93]—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). Thus, to better understand people's fairness perceptions of and reactions to ML models in these more realistic repeated interactions, one may naturally ask that when decision subjects can strategically respond to ML decisions:

- **RQ1**: How will the ML model's fairness properties affect decision subjects' engagement with the model (e.g., willingness to improve themselves and to be subject to ML decisions)?
- **RQ2**: How will the ML 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 the gap between their "desired" qualification level and their current qualification level, and is not affected by the ML model's decision fairness at all. However, one may also speculate that how the ML model treats a subject's group could substantially affect their drive to improve themselves. For example, if the ML model consistently places a subject's group at a disadvantage position, 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 of the impact of the ML model's decision fairness on decision subjects' willingness to improve their qualification, predicting how their willingness to keep interacting with the ML model or their perceived fairness of the ML model

are affected by the ML model's fairness properties also becomes difficult. This is because both retention and fairness perceptions can be highly related to the final qualification level that decision subjects could reach.

To make things more complicated, answers to these questions may even 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 ML 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 ML fairness on decision subjects is larger if ML 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 respond to ML decisions:

- **RQ3**: Do answers to RQ1–RQ2 change if the difficulty for decision subjects to improve their qualification vary with the subject's current qualification level in different ways?
- **RQ4**: Do answers to RQ1–RQ2 change when the ML 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 human-subject experiments on Amazon Mechanical Turk. Subjects in our experiments completed a simulated loan application task that was carefully designed to mirror real-world loan application scenarios where the loan decisions are made by an ML model. Specifically, in Study 1 ($N = 368$), subjects started the experiment with a randomly-assigned persona containing information such as their group identity ("red" or "blue") and their initial credit score (serving as their initial qualification level). After receiving an initial loan approval decision from the bank's ML model, subjects could then interact with the ML model for up to 9 rounds. In each round, they could choose to: (1) apply for a loan without attempting to improve their own qualification (reflecting that in the real world, the action of improving one's qualification is voluntary), (2) trigger additional effort/cost to make a qualification improvement attempt (such attempt will succeed at a fixed rate), and then apply for a loan, or (3) stop being subject to the ML model's

decisions by leaving the bank (reflecting that the action of continue interacting with a model is voluntary). The bank's ML model was designed in a way such that subjects with higher qualification levels generally had higher chance of getting the loan approval. In addition, we created two treatments to reflect varying levels of fairness in the ML model's decisions—in the "fair ML" treatment, the ML model treated subjects in red and blue groups equally, while in the "unfair ML" treatment, the ML model systematically favored subjects from the red group over those from the blue group in granting loans. Therefore, Study 1 enabled us to examine the impacts of ML fairness on decision subjects' engagement with and perceived fairness of the ML model as they strategically respond to the ML model's decisions on them, when the difficulty of qualification improvement does not vary with one's current qualification level and the fairness of the ML model is not examined with respect to salient protected attributes. To understand the generalizability of our findings, we conducted two more replication studies. The only difference between Study 2 ($N = 713$) and Study 1 was that in Study 2, we varied the difficulty of qualification improvement to either increase or decrease with the subject's current qualification level. Moreover, Study 3 ($N = 416$) was the same as Study 1, except for subjects' group identity was set to be their self-reported gender rather than a randomly assigned value, and correspondingly, the fairness properties of the ML model was examined with respect to gender.

Our experimental results show that when decision subjects could strategically respond to the ML model's decisions on them in their long-term interactions with the model, in general, their willingness to improve their qualification and willingness to keep interacting with the ML model are *not* influenced by the fairness properties of the ML model's decisions. This holds true both when the difficulty of qualification improvement changes with one's current qualification level in different ways, and when the ML 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 ML model as less fair if the model biases against the subjects' group by placing them at a disadvantaged position in receiving the favorable decision. The influence of the ML model's decision fairness on decision subjects' perceived fairness of the model is also observed to be larger when it becomes more difficult for people with low qualification to improve their chance of getting the favorable

decisions. We conclude by providing possible explanations for our findings and discussing the implications, limitations, and future work.

4.1 Related Work

An increasing line of work has focused on defining, understanding, and enforcing fairness in ML in recent years. For example, many studies proposed different ways of defining fairness and formalized these definitions as algorithmic constraints [78–84]. However, it has been recognized that there is no universal definition of fairness that can be applicable for all kinds of contexts and satisfies diverse expectations and requirements [64, 102, 103]. This complexity of the fairness notion has inspired many studies on understanding whether and when do humans perceive an ML model as "fair" in decision making [65, 104]. It was found that while some laypeople believe that making an important decision on the basis of past data is inherently unfair [105], people who have high mistrust in human systems tend to believe ML-based decision systems to be as fair as the human decisions [106]. However, when laypeople get educated about algorithmic biases, they relate this to broader issues of racial injustice and economic inequality, profoundly affecting their trust in companies and products using these algorithms [107]. Additional studies have investigated into humans' fairness preferences over various fairness definitions for different decision-making contexts [63, 64, 108], and identified various factors that may affect people's fairness perceptions of an ML-based decision making system, including the explanations used for the ML system's decisions [109–111], the presentation of the ML's decisions [112], features that the ML system utilizes to make its decisions [113, 114], and people's own characteristics like their gender and past education [90].

Of particular relevance to our study in this chapter are a few recent works on understanding *decision subjects*' fairness perceptions of an ML 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 ML model once. For example, Yurrita *et al.* [18] studied the influence of explanations, human oversight, and contestability on fairness perceptions of the decision subjects in a one-shot interaction with the ML 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 ML model, decision subjects perceive an ML model as fairer both if the ML model makes a favorable decision on them and if the ML model is not biased towards or against any particular group [17].

More recently, there is a line of literature on "long-term fairness" [11, 22–26] emphasizing that in the real world, decision subjects often engage in *long-term* interactions with the ML model, and the dynamics between the ML model's decisions on subjects and subjects' strategic reactions to those decisions could create feedback loops. One real-world domain that is frequently studied in the long-term fairness literature is loan lending—for a loan applicant characterized by a profile $\mathbf{x}$, the bank may use its ML-based loan approval system to make loan lending decisions on the applicant, while the applicant may strategically respond to these decisions by staying or leaving the system and/or changing their profile $\mathbf{x}$, which may further change the ML model's training data and impact the model's decision-making policy in the future. This shift from one-shot interaction to long-term interaction has inspired some studies looking into how do decision subjects react to and perceive ML models in their long-term interactions with ML. For example, it was found that when decision subjects could repeatedly interact with an ML-based decision system and had the choice to leave the system at any time, their willingness to stay in the system and their perceived fairness of the system are not significantly affected by the ML system's unbiased treatment across groups, but mostly determined by whether the system is in favor of the subject's own group [12].

Compared to earlier works, this study considers a more realistic set of strategic actions that decision subjects could adopt in their long-term interactions with ML models, i.e., subjects had the full freedom to determine both whether they'd like to continue interacting with the ML model, and whether to improve themselves to increase their chance of getting the favorable decision from the ML model in the future. In the real world, the latter 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 ML model to change their situation to be favorably treated by the model [115]. Previously, a few theoretical works have analyzed cases where an ML model's past decisions

on decision subjects or its overall performance like true/false positive rate can affect the transition of decision subjects' qualification or decision subjects' willingness to invest in their qualification [10, 93]. These studies have found that ensuring fairness constraints in an ML model in a single round of interaction can either promote equality or exacerbate disparity in terms of the qualification rates of different groups in the long run. However, these studies often make simplified assumptions about how decision subjects would respond to the ML model's decisions in determining their qualification transitions. In our study, we take an empirical perspective instead and are interested in examining, in practice, how the ML model's fairness properties affect decision subjects' qualification improvement, retention, and fairness perceptions, in decision subjects' long-term, strategic interactions with the ML model.

4.2 Study 1

To understand decision subjects' long-term interactions with ML models when they could strategically respond to the ML model's decisions on them by voluntarily determining whether to improve their qualification for the favorable decision and whether to continue interacting with the ML model, we conducted a series of human-subject experiments[1]. In particular, in Study 1, we aim to first answer **RQ1**–**RQ2** in an environment where (1) the ML 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.

4.2.1 Experimental Design

4.2.1.1 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 simulated studies of the long-term dynamics of ML fairness [10, 22, 26] as well as experimental studies of fairness perceptions of ML [12, 18]. Specifically, each subject

[1]↑All of our experiments were approved by the IRB of the authors' institution.

was assigned with a randomly-generated persona of a small business owner containing 5 attributes; this persona can be viewed as the subject's loan application "profile". We highlight two key attributes in each subject's persona:

- **Group identity**: The subject's group membership, which was set to be either "red" or "blue".
- **Initial credit score level**: The subject's credit score level at the beginning of the experiment, which could be taken from one of the 9 possible ranges in the set of {300–350, $\cdots$, 500–550, $\cdots$, 700–750}; the subject could later decide to "improve" their credit score with some cost (described in detail below). We told subjects that their credit score level largely reflects 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.

The design choice of setting a subject's group identity as one of the two arbitrary values (red vs. blue) was made to reflect that in Study 1, the ML model's fairness properties across groups are *not* defined with respect to salient protected attributes (in Study 3, we will examine the generalizability of our results in settings where the ML model's fairness properties are defined with respect to salient protected attributes). 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 receiving their assigned profile, the subject was asked to use it to apply for loans from their "local bank" to support their business, and they were told that the bank utilized an ML model to approve or deny loan applications. 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 ML model's decision fairness. After that, they could interact with the ML

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[2], and then apply for a loan (with a cost of 50 coins). The credit score improvement attempt was *not* guaranteed to be successful (see below for more details)—if successful, subjects would be notified, and the ML model's loan approval decision on the subject would be based on the updated credit score level[3].
- **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 from the bank 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 ML 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%; we will vary how the success rate changes with subjects' current qualification level in Study 2 to examine the generalizability of our results). 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[4]. Once the subject successfully progressed to the next credit level, they would at least maintain that level,

[2]↑ 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).

[3]↑The highest credit score level a subject could reach was 800–850. Once a subject reached this level, they would not be able to further improve their credit score.

[4]↑Subjects were not explicitly told about this success rate but could experience it through their actual improvement attempts.

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 chose to apply for a loan in that round (i.e., choose to either "improve and apply" or "apply without improvement"), the ML model's decision on the subject would be revealed to them. A flowchart summary of the ML model's decisions in this round on applicants of different groups, especially for those with similar credit scores as the subject, would also be provided to the subject (see Figure 4.1 for an example flowchart; more details are discussed in Section 4.2.1.2).

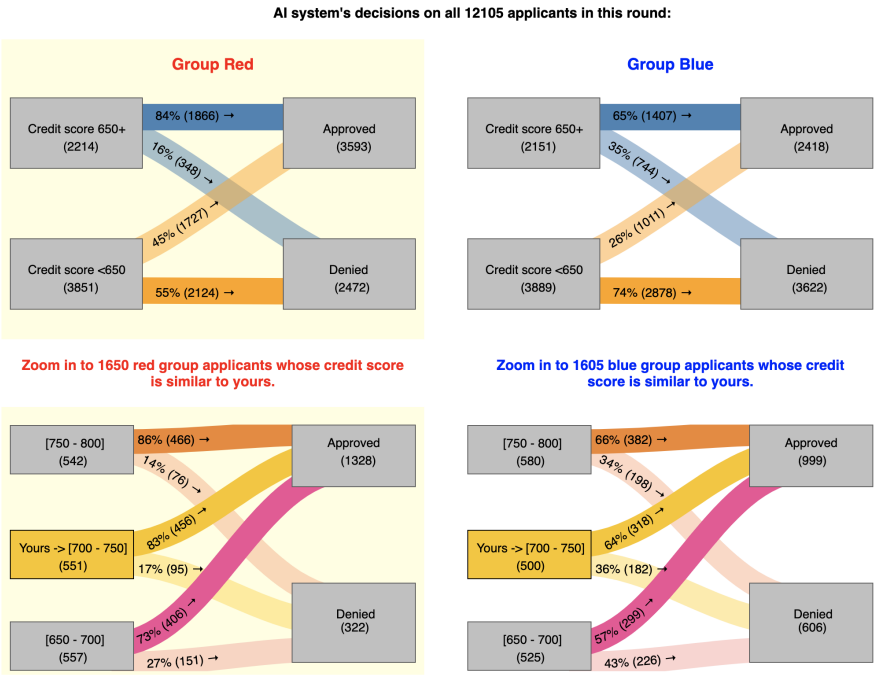


Figure 4.1. An example of the flowchart that subjects in the unfair ML treatment saw in the experiment, which summarizes the ML model's decisions on different groups of applicants in the past round. Subjects could see the frequency of the ML model approving/denying loans both for applicants with-/without "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).

Figure 4.2 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 ML model: (1) participating in the decision making process

(thus triggering the usage of the ML model) is costly; (2) receiving a favorable decision from the ML model is rewarding; (3) the improvement of qualification is costly and has uncertainty; and (4) decision subjects have the freedom to respond to the ML model's decisions by deciding whether to improve their qualification and/or whether to continue being subject to the ML model's decision. In other words, our experimental task abstracts the long-term interactions between decision subjects and an ML model in a real-world loan lending context.

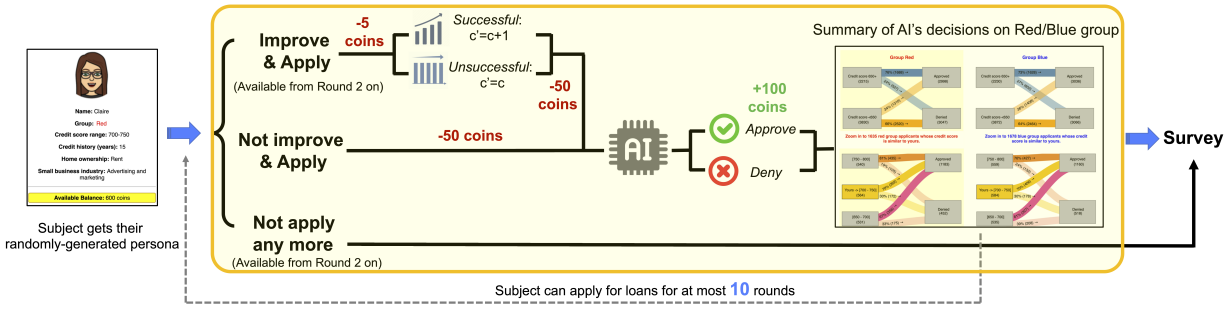


Figure 4.2. An illustration of the process of the loan application task. Here, c represents the subject's current qualification level, while c' denotes the qualification level after an improvement attempt, which can either remain the same or advance to the next level.

4.2.1.2 Treatments.

By varying the fairness level of the bank's ML model across groups, i.e., whether the ML model exhibited systematic biases towards applicants from a certain group, we created two treatments:

- **Fair ML model**: In this treatment, the bank's ML model was fair towards subjects in both the red group and the blue group. In particular, as shown in Table 4.1a, *regardless of the subject's group identity*, the ML 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 ML model**: In this treatment, the bank's ML model was unfair and *systematically biased against subjects in the blue group*. Table 4.1b shows this model's approval

rate for subjects in the red group, while Table 4.1c shows this model's approval rate for subjects in the blue group. As shown in the tables, for every credit level, the ML model's approval rate for red group subjects with this credit level was always 20% higher than that for blue group subjects[5]. Same as that in the previous treatment, the ML model's increment in approval rate for each increased credit level was kept at 7% for both red and blue groups.

Table 4.1. The ML model's probability of approving/rejecting loan applications in different treatments. The probability for approving the loan for a particular applicant is decided by the applicant's group identity and their current credit score level.

(a) Fair model: Red/Blue group

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

(b) Unfair model: Red group

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

(c) Unfair model: Blue group

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

So, if the subject decided to apply for a loan in one round, the ML model's loan approval decision on them would be stochastically decided by the approval rate for their most updated credit level, given both the subject's treatment assignment and group identity. At the end of each round where the subject applied for a loan, we also presented to the subject a

[5]↑Compared to that in the "fair ML" treatment, given any credit level, in the "unfair ML" treatment, the approval rate for subjects of the red group increased by 10% while the approval rate decreased by 10% for subjects of the blue group. Since the subject's group identity was sampled uniformly randomly from the two candidate values (i.e., red vs. blue), for each credit level, the expected chance for a subject getting their loans approved in the "unfair ML" treatment was still the same as a subject with the same credit level in the "fair ML" treatment.

summary of the ML 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 who applied for loans from the bank in this round. For each of these N applicants, we first randomly generated their persona, and then simulated the ML model's decision on them using the approval rates defined by *the subject's assigned treatment*. We then visualized the ML model's decisions on all N applicants using two sets of flowcharts—the first set showed the ML model's approve/deny decisions for applicants with/without "high" credit scores (i.e., a score of at least 650), while the second set zoomed in to applicants with similar credit scores as the subject and showed the ML 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 the flowcharts in Figure 4.1). Within each set, the ML model's decisions on red/blue group applicants were shown separately.

4.2.1.3 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.

In our experiment, the subject first created a nickname and selected an avatar for their persona. Next, we provided instructions for the experiment, including an interactive tutorial explaining each attribute's meaning on the subject's randomly-assigned persona profile. Additionally, this tutorial also 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 ML model's decisions displayed in the flowcharts. To assess subjects' understanding of the tutorial, we then administered a quiz consisting of 5 questions. Subjects were only allowed to advance to the actual experiment after correctly answering all 5 questions.

As the actual experiment started, the subject was first randomly assigned to one of the two treatments as described in Section 4.2.1.2. 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[6]. The survey included questions about the subject's demographics (e.g., race, age, education), perceived fairness of the ML model, sensitivity to fairness, risk attitude, and level of empathy. 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 series of statements from 1 (strongly disagree) to 5 (strongly agree). Specifically, statements regarding subjects' perceived fairness of the ML model (e.g., "The banks ML system is fair to manage loan applications.") were adapted from [17]. Statements on subjects' sensitivity to fairness (i.e., how much the subject holds fairness as a core value; an example statement was "I would stop using an ML system if it is unfair, even if it tends to be in favor of me") were adapted from [12]. Finally, statements related to the subject's risk attitude (e.g., "I like to do frightening things.") and level of empathy[7] (e.g., "I get a strong urge to help when I see someone who is upset.") were taken from [98] and [116], respectively. A detailed list of the survey questions can be found in the appendix A.

Finally, upon survey completion, the subject received their bonus that was proportional to their remaining account balance (i.e., we converted every 500 coins to $2.00). The base payment of our experiment was $2, and the maximum bonus a subject could earn from our experiment 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.

4.2.2 Analysis Methods

We had three dependent variables in our analysis:

[6]↑The entire set of survey questions are provided in appendix A.
[7]↑We included the level of empathy that an individual exhibits as a covariate in our analysis, since highly empathetic people may be more concerned about the fairness of an ML model, even if the ML model is not biased against them.

- **Improvement**: The number of times that the subject made a qualification improvement attempt during the loan application task, with a higher value indicating a higher level of willingness for the subject to improve themselves;

- **Retention**: The number of times that the subject applied for a loan in our experiment, with a higher value indicating a higher level of willingness for the subject to continue interacting with the ML model and be subject to the ML model's decision;

- **Fairness perceptions of the ML model**: This was obtained through summing up the subject's rating to statements regarding the ML model's fairness in the exit-survey; the higher the rating, the fairer the subject found the ML model to be.

We fit the data we collected from our experiment into two sets of regression models to predict subjects' engagement (**RQ1**) and fairness perceptions (**RQ2**). In the first set of models, we coded the ML model's fairness level across group as a single binary independent variable (i.e., "*Fair ML*"; subjects in the fair ML model treatment had the value of this variable set to 1, otherwise 0); this allowed us to examine how the ML model's fairness level across groups affects all dependent variables (i.e., improvement, retention, and fairness perceptions). For the second set of models, we coded the ML model's treatment to the subject's own group using two binary independent variables—"*advantaged*" and "*disadvantaged*", which represented whether the ML model systematically favored or disfavored the subject's group compared to the other group. For subjects in the fair ML model treatment, both "advantaged" and "disadvantaged" were set to 0; for subjects in the unfair ML model treatment, red group subjects had advantaged=1 and disadvantaged=0 while blue group subjects had advantaged=0 and disadvantaged=1. This enabled us to understand how the ML model's treatment to the specific group that the decision subject belongs to affects all dependent variables. In all of our regressions, we controlled the subject's initial credit level, their sensitivity to fairness, risk attitude, and empathy level as covariates.

4.2.3 Experimental Results

A total of 368 subjects participated in our experiment (see the appendix A for the subject demographics information). We analyzed the full dataset collected from these 368 subjects to answer our research questions.

Table 4.2. Regression models predicting decision subjects' improvement, retention, and perceived fairness based on the ML model's decision fairness for Study 1. Our results indicate that subjects who are being placed by the ML model at the disadvantaged position for receiving the favorable decision rate the model as less fair (Model 6). Coefficients and standard errors are reported. *, **, and *** represent statistical significance levels of 0.05, 0.01, and 0.001 respectively. Significant coefficients on independent variables of interests are bolded.

	Improvement		**Retention**		**Perceived Fairness**	
	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*
Fair ML	-0.03		-0.42		0.59	
	(0.35)		(0.40)		(0.52)	
Advantaged		0.49		0.77		0.27
		(0.43)		(0.49)		(0.64)
Disadvantaged		-0.40		0.08		**-1.41***
		(0.42)		(0.48)		**(0.63)**
Risk attitude	0.05	0.04	0.07	0.06	0.36***	0.34***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.07)	(0.07)
Fairness Sensitivity	-0.00	-0.01	-0.03	-0.03	-0.11	-0.13
	(0.08)	(0.08)	(0.09)	(0.09)	(0.11)	(0.11)
Empathy	0.13*	0.13*	0.12$^{\cdot}$	0.12$^{\cdot}$	-0.11	-0.11
	(0.06)	(0.06)	(0.07)	(0.07)	(0.09)	(0.10)
Initial Credit Score	-0.13*	-0.13	0.22**	0.22**	0.19*	0.20*
	(0.07)	(0.06)	(0.07)	(0.07)	(0.09)	(0.10)
Constant	2.49*	2.54*	3.76**	3.42**	11.51***	12.27***
	(1.12)	(1.12)	(1.28)	(1.28)	(1.68)	(1.67)

4.2.3.1 RQ1: Impacts of the ML model's decision fairness on engagement.

First, we look into that in decision subjects' long-term, strategic interactions with an ML model, how the ML model's decision fairness affects decision subjects' engagement with the

model, including their willingness to improve their qualification and their willingness to be subject to the ML model's decisions. Table 4.2 (left panel) reports the regression results for decision subjects' willingness to improve their qualification. Interestingly, we find that when decision subjects could strategically respond to ML decisions in their repeated interactions with an ML model, their average level of willingness to improve their qualification is *not* significantly affected by either the ML model's fairness level across different groups (Model 1), or whether the ML model is biased towards or against the subject's group (Model 2). Similar observations can also be made for decision subjects' retention (Table 4.2, middle panel, Models 3 and 4)—once decision subjects get the opportunity to strategically react to the ML model's decisions on them, on average, they are equally willing to keep interacting with the ML model regardless of the model's decision fairness, both across groups and specifically towards their group.

4.2.3.2 RQ2: Impacts of the ML model's decision fairness on fairness perceptions.

However, when examining the impacts of the ML model's decision fairness on decision subjects' average level of fairness perceptions of the model (Table 4.2, right panel, Models 5 and 6), we have different findings. While the ML model's fairness level across different groups still does not appear to significantly influence decision subjects' average level of perceived fairness of the model (Model 5), we do find that when the ML model systematically disfavors the subject's group, the subject generally perceives the model as significantly less fair (i.e., p-value for the variable "disadvantaged" in Model 6: $p = 0.026$).

4.3 Study 2: When the Qualification Improvement Difficulty Varies with Current Qualification Levels

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 aim to 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**.

4.3.1 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. In each of the two sub-experiments of Study 2, we considered that the success rate of a credit improvement can vary with the subject's current credit level in a specific way:

- **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). 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.
- **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.

We used the same experimental procedure as that in Study 1 by posting the experiment as a HIT on MTurk exclusively for U.S. workers. However, there were two key differences: (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.

4.3.2 Experimental Results

A total of 713 subjects participated in our experiment ("Easy to hard" sub-experiment: 328 subjects, "Hard to easy" sub-experiment: 385 subjects). Below, we analyzed the full

dataset collected from these participants to answer **RQ3** by re-examining **RQ1**–**RQ2** on this dataset using the same analysis methods as in Study 1.

Table 4.3. Regression models predicting decision subjects' improvement, retention, and perceived fairness based on the ML model's decision fairness in the two sub-experiments of Study 2. Our results indicate that subjects' fairness perceptions of the ML model are significantly affected by its fairness properties when it is more difficult for subjects with low qualification to improve their chance of the favorable decision (Models 11–12). Coefficients and standard errors are reported. *, **, and *** represent significance levels of 0.05, 0.01, and 0.001, respectively. Significant coefficients on independent variables of interests are bolded.

	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 ML	0.18		-0.00		0.29		0.01		0.34		**1.45****	
	(0.39)		(0.31)		(0.40)		(0.37)		(0.51)		(0.49)	
Advantaged		0.11		-0.30		0.18		-0.18		-0.06		-0.58
		(0.47)		(0.39)		(0.49)		(0.48)		(0.63)		(0.62)
Disadvantaged		-0.46		0.24		-0.79		0.12		-0.63		**-2.14*****
		(0.48)		(0.36)		(0.50)		(0.44)		(0.64)		(0.57)
Risk attitude	0.13**	0.14**	0.08*	0.08*	0.15**	0.16**	0.18***	0.18***	0.34***	0.35***	0.31***	0.32***
	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)	(0.07)	(0.07)	(0.07)
Fairness Sensitivity	-0.10	-0.10	-0.03	-0.03	-0.13	-0.14	-0.04	-0.04	0.25*	0.25*	0.10	0.11
	(0.08)	(0.08)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)	(0.08)	(0.11)	(0.11)	(0.10)	(0.10)
Empathy	0.13*	0.12	0.10	0.10	0.16*	0.14*	0.12	0.12	-0.14	-0.15	-0.20*	-0.20*
	(0.06)	(0.06)	(0.05)	(0.05)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.09)	(0.09)
Initial Credit Score	-0.11	-0.11	-0.24***	-0.24***	0.16*	0.16*	0.25***	0.25***	0.13	0.13	0.24*	0.23*
	(0.08)	(0.08)	(0.06)	(0.06)	(0.08)	(0.08)	(0.07)	(0.07)	(0.10)	(0.10)	(0.09)	(0.09)
Constant	3.48**	3.79**	2.95**	2.94**	4.19**	4.71***	2.78*	2.78*	8.64***	9.12***	10.23***	11.71***
	(1.32)	(1.33)	(1.00)	(1.00)	(1.37)	(1.38)	(1.23)	(1.22)	(1.74)	(1.76)	(1.60)	(1.58)

4.3.2.1 Re-examination of RQ1: Impacts of the ML model's decision fairness on engagement.

We repeat the analysis on how the ML model's decision fairness affects decision subjects' willingness to improve and willingness to be subject to the model's decisions within each sub-experiment separately, and results are reported in Table 4.3 (left and middle panels, Models 1–8). Consistent with our findings in Study 1, it is found that, on average, decision subjects' willingness to improve themselves and willingness to keep interacting with the ML model are not significantly affected by the ML model's fairness level across groups or its treatment

towards the subject's group, regardless of how the difficulty of qualification improvement changes with the subject's current qualification level.

4.3.2.2 Re-examination of RQ2: Impacts of the ML model's decision fairness on fairness perceptions.

Table 4.3 (right panel, Models 9–12) shows how the ML model's decision fairness affects the average level of decision subjects' perceived fairness of the ML model in the two sub-experiments of Study 2. Interestingly, we find that when it is easier for lowly-qualified subjects to improve their qualification level than highly-qualified subjects (i.e., in the "Easy to hard" sub-experiment), subjects' perceived fairness of the ML model is *not* significantly impacted by either the ML model's group-level fairness or the ML model's treatment towards the subject's own group. In contrast, in the "Hard to easy" sub-experiment, we find that subjects' perceived fairness of the ML model is significantly affected by the ML model's fairness properties. In particular, in this sub-experiment, decision subjects' average perceived fairness of the ML model had a significant increase when the ML model made fair decisions across different groups (i.e., p-value for the variable "fair ML" in Model 11: $p = 0.003$). In addition, those subjects who have been placed at the disadvantaged position by the ML model also perceived the model to be significantly less fair (i.e., p-value for the variable "disadvantaged" in Model 12, $p < 0.001$).

4.4 Study 3: When AI Fairness is Examined On Protected Attributes

Finally, to answer **RQ4**, we conduct a third study, in which the ML model's fairness properties is examined with respect to a salient protected attribute, i.e., gender, and subjects' group identities in the experiment were determined by their self-reported gender in the real world.

4.4.1 Experimental Design

In Study 3, we adopted the experimental design from Study 1 and made only 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 ML" treatment, the ML model used treated male and female subjects equally in granting loans; the approval rate of this ML model to both male and female was determined by Table 4.1a. In contrast, in the "unfair ML" treatment, the ML model exhibited gender bias and systematically favored male over female subjects in granting loans; this ML model's approval rate for male and female was determined by Table 4.1b and Table 4.1c, respectively (i.e., males correspond to the red group in Study 1, and females correspond to the blue group in Study 1). This unfair ML model was designed to mirror the ML model's gender biases against females in the real world that were increasingly revealed by researchers [96, 100, 117, 118]. Finally, we also modified the flowcharts that were shown to subjects at the end of each round, so that they summarized the ML 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 ML 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 avoid actually paying female subjects systematically less in our experiment. In particular, at the beginning of the experiment, subjects were told that beyond the $2 base payment, the bonus payment that they would receive from 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. This deception was used to make subjects believe that their final earnings were directly influenced by the ML model's decisions based on their real-world gender, thus enabling us to capture subjects' genuine reactions towards an ML model that may or may not exhibit gender bias. Upon completing the experiment, subjects were immediately debriefed about the deception they had experienced. The median time subjects spent on the experiment was 19 minutes, which resulted in an hourly rate of approximately $16.70.

4.4.2 Experimental Results

A total of 416 subjects participated in Study 3. In the following, we used the full dataset from these subjects to examine whether our results of Study 1 still hold true when the ML model's fairness properties was examined with respect to gender and subjects' group identity in the experiment was decided by their real-world gender.

Table 4.4. Regression models predicting decision subjects' improvement, retention and perceived fairness based on the ML model's decision fairness for Study 3. Consistent with earlier studies, our findings also indicate here that subjects disadvantaged by the ML model rate the model as less fair (Model 6). Coefficients and standard errors are reported. *, **, and *** represent significance levels of 0.05, 0.01, and 0.001, respectively. Significant coefficients on independent variables of interests are bolded.

	Improvement		**Retention**		**Perceived Fairness**	
	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*	*Model 6*
Fair ML	0.12		-0.45		0.45	
	(0.32)		(0.37)		(0.50)	
Advantaged		-0.04		0.64		0.23
		(0.38)		(0.43)		(0.58)
Disadvantaged		-0.22		0.18		**-1.38***
		(0.42)		(0.47)		**(0.65)**
Risk attitude	0.09*	0.09*	0.13**	0.13**	0.41***	0.41***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.06)	(0.06)
Fairness Sensitivity	0.04	0.04	0.02	0.02	-0.14	-0.13
	(0.07)	(0.07)	(0.08)	(0.08)	(0.10)	(0.10)
Empathy	0.05	0.05	0.01	0.01	-0.01	-0.01
	(0.05)	(0.05)	(0.06)	(0.06)	(0.08)	(0.08)
Initial Credit Score	-0.23***	-0.23***	0.11	0.12	0.36***	0.37***
	(0.06)	(0.06)	(0.07)	(0.07)	(0.10)	(0.10)
Constant	3.14**	3.23**	5.04***	4.53***	9.21***	9.44***
	(1.06)	(1.07)	(1.21)	(1.23)	(1.66)	(1.67)

4.4.2.1 Re-examination of RQ1: Impacts of the ML model's decision fairness on engagement.

We first revisited the analysis to determine the impact of the ML model's fairness on decision subjects' willingness to improve and willingness to be subject to the model's decisions, when the ML model's fairness properties are discussed with respect to decision subjects' real-world gender. The regression results are presented in Table 4.4 (left and middle panels, Models 1–4). Again, we find that, on average, decision subjects' willingness to improve themselves and willingness to keep interacting with the ML model are not significantly affected by the ML model's fairness level across groups or its treatment towards the subjects' group, even when the ML model's fairness is examined with respect to salient protected attributes like gender.

4.4.2.2 Re-examination of RQ2: Impacts of the ML model's decision fairness on fairness perceptions.

We then investigated into the influence of the ML model's decision fairness on the average level of decision subjects' perceived fairness of the ML model. Our regression results are presented in Table 4.4 (right panel, Models 5 and 6). Consistent with our findings from Study 1, we again found that female subjects perceived the ML model as less fair if the ML model systematically biased against females, as evidenced by the significance of the disadvantaged variable in Model 6 (i.e., p-value for the variable "disadvantaged" in Model 6: $p = 0.034$). In contrast, male subjects who were being placed at an advantaged position by the unfair ML model did not show any decrease in their perceived fairness of the model, despite the ML model exhibited a clear gender bias.

4.5 Conclusions and Discussions

In this chapter, we conducted randomized human-subject experiments to investigate how decision subjects engage with and perceive the fairness of an ML model in their long-term interactions with the model when they could strategically respond to the ML model's decisions on them. Our findings suggest that in general, the ML model's fairness properties,

in terms of both whether the ML model treats different groups in a similar way and whether the ML model favors/disfavors the specific group that the subject belongs to, have limited impact on decision subjects' willingness to improve their qualification or keep interacting with the ML model. This is true even when the ML model's fairness properties are examined with respect to salient protected attributes like gender. However, the ML model's fairness properties, especially its biased treatment against a decision subject's own group, still tend to result in a decrease in the subject's perceived fairness of the model. The impact of the decision fairness of ML on decision subjects' perceived fairness of ML also appears to be larger in an environment where improving one's chance of receiving the favorable decision is particularly difficult for those who have low qualification to begin with.

In this section, we reflect on our findings, and address the limitations and future work of our study.

4.5.1 Similarity and difference of our results with earlier findings

Consistent with findings in earlier studies [12, 17], we find in this study that when decision subjects can strategically respond to ML decisions in their long-term interactions with the ML model—especially as they have the opportunities to invest in themselves to improve their chance of receiving the favorable decision—their perceived fairness of the ML model is still often affected by the model's unfavorable treatment to their own group. This indicates that social comparison across groups is still a key factor that contributes to decision subjects' fairness perceptions of an ML model, despite the possibility to improve one's qualification may have shifted some of their attention towards self comparison.

On the other hand, different from prior work, which showed that decision subjects' retention in an ML-based decision system is also significantly affected by the ML model's favorable/unfavorable treatment to the subject's own group [12], we find that this is not the case when decision subjects can strategically decide whether and when to improve their qualification. We speculate that this is because the qualification improvement opportunities provide decision subjects the possibility to utilize these opportunities to "gain something" rather than leaving the system with nothing. Indeed, the ML model's favorable/unfavorable

treatment to the subject's own group is shown to not significantly affect decision subjects' average level of willingness to *improve* their qualification. In particular, for those decision subjects whose group is placed at the disadvantaged position by the ML model, continuing to improve themselves and interact with the ML model for roughly the same number of times but receiving less favorable decisions is certainly not ideal—as reflected in their perceived fairness of the ML model—but from a pure utility point of view, it may be better than simply "boycotting" the ML model. In this sense, similar as the previous work, our findings again demonstrate that decision subjects' engagement with an ML model and their fairness perceptions of the model are likely driven by different factors.

4.5.2 The influences of the qualification improvement difficulty on decision subjects' fairness perceptions

In Study 2, as we change how the difficulty for one to improve their qualification varies with their current qualification level, we generally find the ML's decision fairness still does not have significant impacts on decision subjects' willingness to improve themselves or continue interacting with the model. However, we note some deviations in decision subjects' fairness perceptions of the model. The ML model's unequal treatment towards people in different groups or its biased treatment against a decision subject's own group significantly decreases the perceived fairness of the ML model for subjects in the "Hard to easy" sub-experiment, but not for subjects in the "Easy to hard" sub-experiment. Here, we provide some conjectures on why we have these observations.

We conjecture that subjects' perception of an ML model's fairness is largely affected by two main factors: the model's treatment to their present-self and the model's anticipated treatment to their ideal future-self after qualification improvements. Subjects who begin with relatively high qualification may focus more on their present-self (since the room for qualification improvement is small). However, for subjects who begin with relatively low qualification, the emphasis they put on these two selves might differ, depending on their belief in how successful they would be in reaching their ideal future-self. This conjecture may offer an explanation for our findings in Study 2. Specifically, in the "Easy to hard" sub-experiment, those with relatively low levels of qualification, who arguably value the

improvement opportunities the most, witness high success rates of improvement and possibly form a more optimistic view of their future-selves. Even as the ML model's treatment to their higher-qualified future-self turns out to still exhibit systematic bias, decision subjects may perceive the model as relatively more fair due to the achieved increase in their chance of getting the favorable decisions, and their belief in their ability to continue this increase. Conversely, in the "Hard to easy" sub-experiment, subjects with lower qualification are much less successful in improving their qualification, hence they might focus more on their present-self and are more likely to notice the model's existing biases against them.

4.5.3 Understanding the findings when ML's decision fairness is examined with respect to protected attributes

In Study 3, even when we incorporated subjects' real-world gender as their group identities in the experiment and examined the ML model's fairness properties with respect to groups defined by gender, our findings were still largely consistent with those of Studies 1 and 2. In particular, while female subjects considered the ML model as less fair when being placed at the disadvantaged position in receiving the favorable decision by the model, they showed similar levels of willingness to engage with the model as male subjects. We provide a few possible explanations for this observation. First, prior research has pointed out the frequent challenges women encounter due to societal biases in the real world [100, 119–121]. Given this context, it is possible that female subjects in our study might find the potential gender biases exhibited by the ML model in our experiment to be "familiar", and were disappointed by it. This "disappointment" coupled with their past experiences of dealing with real-world bias might have encouraged them to adapt to the model, resulting in a similar level of engagement with the model as other subjects. A counter conjecture to this, simply from a utility standpoint, is that for female subjects, trying to keep improving their qualifications may still have been a better option than deciding to "boycott" the system and leave. This is because keeping engaging with the ML model, despite its potential biases, still provides an opportunity for female subjects to profit from the system compared to leaving and getting nothing in return. Considering the low-stakes nature of our experimental task, the discrimination that female subjects endured might not have fully convinced them to react

too strongly to the ML model's biased decisions. Although they still rated the ML model as less fair, they did not strongly show this disappointment in their engagement with the ML model, to the extent of leaving and boycotting the system. Thus, we urge the readers to not generalize these findings to other settings where higher stakes are involved.

4.5.4 Implications of our findings

Findings of our study hold important implications for the real-world deployment of ML-based decision systems. First, our results reveal that when decision subjects could strategically decide whether to improve their qualification for the favorable decision beyond determining whether to keep interacting with the model, their overall engagement with the ML model is similar no matter whether the ML model is biased towards or against their own group. This is observed despite that subjects generally consider an ML model as less fair if the model is biased against their group. This discrepancy highlights the importance of going beyond user engagement when assessing the fairness of an ML model. In other words, when decision subjects have the opportunities to improve their qualification, the equality in engagement across different groups of decision subjects should *not* be used as a proxy indicator of a fair ML model. Instead, responsible ML practitioners should delve deeper into truly understanding decision subjects' perceptions of and satisfaction with the ML model to develop fair ML-based decision systems.

On the other hand, compared to previous findings [12, 17], 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 ML model exhibits a degree of unfairness across groups, especially for subjects whose group is placed at the disadvantaged position by the ML model. This means that when institutions that employ ML-based decision systems (e.g., banks) identify fairness concerns of their ML models, 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 institutions some time to address the fairness concerns of the ML 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 ML model and truly improve decision subjects' satisfactions with the model—especially in a real-world environment with competitions (e.g., multiple banks providing loan lending services), the user satisfaction may play a crucial role in shaping long-term user retention.

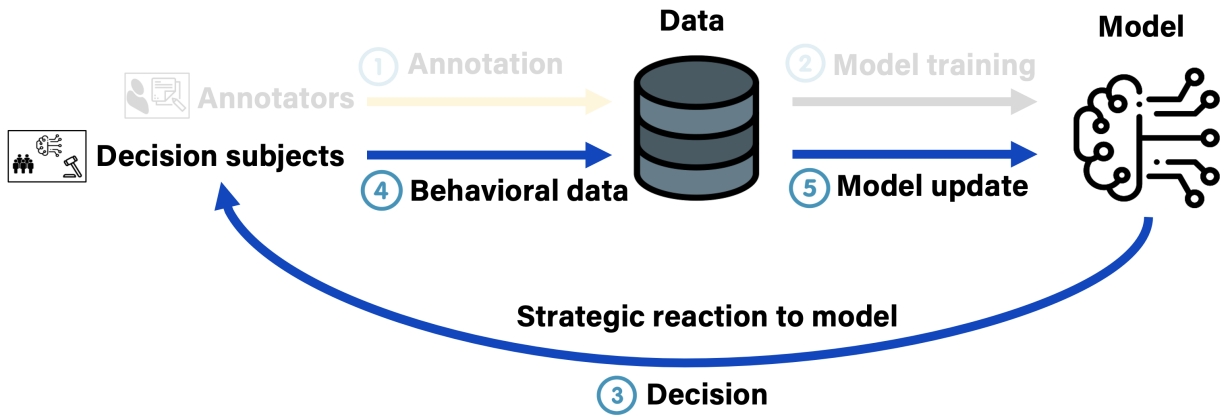
4.5.5 Limitations and future work

Our study was based on a specific context (i.e., a simulated ML-based loan application task). Cautions should be used when generalizing our findings to other settings, especially for high-stakes scenarios. Particularly, our results can be constrained by the specific setup of our experimental task, including the cost we define for each qualification improvement attempt, the cost for each loan application, the reward for each favorable decision, and the linear qualification improvement schemes we adopt in our studies. Future research could systematically vary these "parameters" and investigate how they affect decision subjects' behavior and perceptions. Another important limitation to note is that the ML 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 ML 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 ML fairness.

4.6 Acknowledgments

This work is currently in preparation. I would like to thank my advisor, Professor Ming Yin, for her continued guidance and feedback throughout this project. I also acknowledge the support of the Department of Computer Science at Purdue University, as well as the crowd workers who participated in the study.

5. AN INVESTIGATION OF DECISION SUBJECTS' INTERACTION WITH PERIODICALLY UPDATED ML-BASED TASK ALLOCATIONS



Exploring the "Decision," "Behavioral data," and "Model update," i.e., stages 3–5, of the machine learning pipeline. The complete figure is presented in Figure 1.1 of Chapter 1.

In the previous chapters, we investigated various parts of the machine learning (ML) pipeline, examining not only how human involvement and biases affect ML models but also how these models, in turn, impact the humans who interact with them. As the ML models began to absorb and reflect human biases in their decisions, a complex and reciprocal relationship emerged. However, the analysis in earlier chapters did not account for model updates or the iterative nature of real-world ML applications. In practice, it is essential to consider how behavioral data—generated by humans in response to ML decisions—influences subsequent model updates. Thus, in this chapter, we delve into this unexplored aspect of the human-in-the-loop ML pipeline by explicitly examining the feedback loop formed by the stages of "Decision," "Behavioral Data," and "Model Update," as initially illustrated in Figure 1.1 of Chapter 1, specifically stages 3–5. By simulating human-ML interactions and conducting an empirical study, we investigate how initially fair ML models evolve over time, comparing their trajectories against one another as they get updated with different kinds of data. Thus, this chapter offers the first empirical exploration into how iterative feedback loops between human reactions and ML model updates shape long-term fairness outcomes, significantly expanding upon the non-updating ML model perspectives considered in Chapters 3 and 4. In doing so, we also provide practical insights that complement and extend existing theoretical work on long-term fairness in algorithmic systems [10, 22, 26, 93].

ML systems are increasingly used to make important decisions in domains ranging from employment and education to finance and transportation. Yet, extensive research has shown that these systems often replicate and even amplify existing societal biases when trained on historical data [5, 100]. As a result, algorithmic fairness has emerged as a central concern across disciplines. A growing body of work has explored how different stakeholders—especially decision subjects, the individuals affected by ML decisions—perceive and respond to algorithmic decisions [12, 17, 63]. These studies show that fairness perceptions depend not only on the accuracy or transparency of the ML model but also on how decisions align with group identities, expectations, and social norms.

Importantly, most fairness studies assume a one-shot interaction between a model and a decision subject. However, in real-world applications such as gig work platforms, online education, or credit scoring systems, individuals often interact with ML models repeatedly over

time. This opens up complex, cyclical dynamics where decision subjects no longer passively accept model outputs but instead respond strategically—by changing how they engage with the system, investing in improvement, or withdrawing from participation altogether [10, 11, 93]. Their behavioral responses, in turn, generate new data that may be used to retrain the model, potentially introducing or reinforcing feedback loops. Recent theoretical work on long-term fairness in ML has highlighted the importance of modeling these feedback mechanisms [22, 25, 26], yet few empirical or simulation-based studies have operationalized and tested them in interactive settings, where humans are involved.

In the previous chapters of this dissertation, we studied how human bias enters different stages of the ML pipeline and how people respond to ML decisions in non-updating ML model contexts. However, we did not investigate what happens when the ML model is iteratively updated using data that originates from people's behavioral responses to the model's own decisions. This chapter addresses that gap. We now shift to modeling and empirically testing the feedback loop described in Chapter 1—namely, how decisions made by ML models influence subject behavior, how that behavior is encoded as data, and how it affects the next round of model updates. This exploration provides a more complete view of fairness in human-in-the-loop systems.

To do so, we focus on a specific type of decision subjects—workers who are assigned tasks on a crowdsourcing platform by an ML model—and conduct a two-part study combining simulations with human-subject experiments. This scenario is selected to reflect real-world algorithmic management contexts. First, we design a simulation that models interactions between workers and ML models on a gig platform managed by the models themselves, where the ML models assign tasks to workers. Workers freely choose how many tasks to complete and with what quality, while some simulated raters rate their performance—sometimes unfairly. The ML model is updated over time using these ratings. Crucially, we compare two models: one updated using fair (i.e., true) ratings and another using biased (i.e., systematically unfair) ratings. Despite starting out identical and fair, the models diverge as feedback loops emerge, with the unfairly updated model developing increasing disparities in task assignment between demographic groups. We also simulate worker responses, such

as lowering/increasing retention or reduced/increased effort, and show how these amplify or mitigate the fairness degradation depending on the form and strength of the response.

To validate and extend our findings, we pair the simulation with an empirical experiment in which human subjects interact with a prototype platform that assigns tasks to workers using an ML model. Subjects complete a real-world annotation task (e.g., genuine smile detection) and are randomly assigned to one of three treatments: a treatment where they may skip tasks based on perceived competence—to be cautious about their work quality and thus protect their rating; a remediation treatment where workers who report fairness issues and are verified as belonging to a disadvantaged group receive extra task assignments to offset biased ratings; and a control treatment where neither mechanism is present. Our goal is to observe whether worker reactions and model evolutions mirror those observed in the simulation, and whether giving options to workers such as skipping or assigning more tasks under biased ratings, lead to fairness improvements or further harms.

Together, our simulation and empirical findings highlight a critical insight: feedback loops can emerge even in initially fair systems, and while interventions such as assigning more tasks to disadvantaged groups may slow the growth of disparities, they offer only limited relief if the underlying evaluation mechanisms remain biased—underscoring the importance of addressing root causes of unfairness rather than applying surface-level fixes. By moving beyond one-shot fairness assessments and incorporating repeated interaction, model retraining, and user reactions, this chapter advances the understanding of fairness as a dynamic and evolving property in ML systems. It contributes both theoretical and empirical evidence to the growing literature on long-term algorithmic fairness [10, 11, 25, 93], and extends the human-in-the-loop perspective developed in earlier chapters.

5.1 Related Work

Algorithmic management (AM) systems are becoming central to coordinating labor on gig economy platforms [122–125]. A growing body of literature has explored how these systems influence perceptions of fairness [122, 126–128], organizational commitment [125, 127, 129], productivity [130, 131], and inequality [132–135]. Much of this research focuses on

the interaction between AM and feedback loops—cyclical processes in which user responses (e.g., biased ratings, strategic actions) shape future algorithmic decisions—often reinforcing existing disparities [129, 131, 134, 136, 137].

Starting with the first stage of the feedback loops, worker behavioral data is greatly influenced by worker perceptions and early work by Lee et al. [122] shows that while algorithmic decisions in mechanical tasks may be perceived as efficient and fair, they tend to evoke negative emotions and are perceived as less fair in tasks involving subjective judgment. These perceptions are shaped not only by the decision outcomes but also by the opacity and rigidity of algorithmic logic [129, 138]. Several other studies have categorized practices that can promote higher fairness perceptions in AM by enhancing transparency, offering meaningful choice, and allowing recourse for unfair decisions [123, 126]. Rzepka and Berger [126], for instance, propose mechanisms such as evidence-based redress and delegated dispute resolution. These practices may help reduce worker dissatisfaction and perceptions of injustice/unfairness, which are often amplified by unexplainable outcomes in opaque systems [129]. In gig settings where human supervisors are replaced by algorithms, perceived organizational support (POS) may be lower; however, when mechanisms are provided to enhance this perception, it can lead to increased fairness perceptions in AM [125]. Already some components of the AM lead to better outcomes, for example, several papers find that personal management style and autonomy in AM can lead to increased POS, affective commitment and better worker well-being [125, 127]. Factors such as information quality and system usability have been shown to directly influence both fairness perceptions and emotional attachment to platforms like Uber [125, 127].

A core ethical concern in AM has always been the presence of biased feedback loops. Numerous studies have documented how biased input—such as customer ratings or reviews influenced by race or gender—feeds into AM systems and creates algorithmic outcomes that treat marginalized groups unfairly [128, 132–136]. These outcomes are particularly problematic given that platforms like Uber and Fiverr use such biased metrics in reputation scores and task assignment [131–133]. Bokányi and Hannák [131] specifically illustrate how ride-sharing algorithms unintentionally generate income inequality through self-reinforcing matching dynamics. Similarly, Hannák et al. [134] show how freelancers from marginalized

backgrounds receive fewer opportunities due to biased rankings, which form part of a feedback loop embedded in algorithmic management systems. Workers have developed strategies to resist and adapt to AM. These include gaming platform logic (e.g., logging off to trigger surge pricing), selectively accepting tasks, and collaborating with others to interpret algorithmic behavior [139–141]. Zhang et al. [137] develop a game-theoretic model showing how revealing labor demand signals via algorithmic assignment can unintentionally decrease worker engagement and harm platform revenue—another manifestation of a feedback loop.

To counterbalance top-down AM systems, recent research advocates for bottom-up, stakeholder-driven approaches to mitigate fairness concerns. Liu et al. [142] and Zhang et al. [143] propose intelligent assistants and co-designed tools that provide workers with insights into bias in task allocation and enable collective data aggregation to enhance transparency and fairness. Tools like Reputation Agent [144] promote fairer feedback from customers, helping to mitigate bias reinforcement in algorithmic assessments. Complementing these technical tools, workers themselves have engaged in grassroots resistance strategies—such as organizing, discursive framing, and legal mobilization—to contest algorithmic control and advocate for fairer treatment [145]. Other human-centered mechanisms have also been proposed: algorithmic recourse mechanisms offer individuals actionable steps to challenge and overturn unfavorable algorithmic decisions [146]; protective optimization technologies (POTs) enable communities to externally counteract harmful algorithmic impacts [147]; and community-driven reporting initiatives like CRASH incentivize the identification and correction of algorithmic harms [148].

In sum, the literature reveals that AM systems are not neutral; they mediate power and fairness through complex, often opaque processes [122, 123, 129]. Feedback loops based on biased inputs (such as discriminatory ratings [132–134]) can reinforce disparities over time, while strategic worker responses [137, 141] can further undermine the platform.

While existing work has shown that biased reputation systems can perpetuate inequality, and that workers adapt their behavior in response to platform dynamics, prior studies have largely examined these effects in static or observational settings. Our work builds on this foundation by combining a behavioral simulation with a complementary empirical experiment to directly model and measure how fairness perceptions influence worker behavior over time,

and how this behavior, in turn, influences model evolution. This study contributes to the understanding of how AM evolves under different models of human behavior and how AM systems can be designed with various mitigation methods to eliminate these feedback loops. In other words, by capturing both algorithmic evolution and user reactions under fair and unfair feedback data, our work offers a dynamic, simulation- and empirical-based view of how feedback loops form and influence ML-driven labor systems.

5.2 Simulation

We conducted simulations to determine how an ML model that is updated periodically and started of fair in task assignments could evolve over time as it incorporates feedback data based on people's reactions to the model's decisions. These simulations are intended to reflect what might happen on real-world work platforms that utilize algorithmic management, where ML models estimate workers' capabilities and assign tasks accordingly.

We simulated two types of people that would be contributing to the model updates—the workers and the raters. The workers are the people who complete the tasks that are assigned to them by the ML model. The raters are the people who rate the quality of the tasks that are completed by the workers. In this scenario, workers complete tasks of their own volition and also determine the quality of their work—either good or poor. We call the number of tasks a worker completed to be the worker's "retention" and the fraction of completed tasks that are of good quality to be the worker's "work quality" (or "quality" in short).

Completed tasks are passed down to raters to judge out of all the tasks that are completed by a worker how many of them are of good quality. So, raters are effectively determining a worker's "quality". However, given the subjectivity in raters' judgments, true quality may not always be reflected in their ratings. Specifically, we assume that workers are coming from two demographic groups, one of which is a minority group that typically raters are biased against. In other words, for minority group workers, a rating can be underestimated while for others, it can be overestimated. These ratings are then used to update the ML model and its future assignments/decisions. In other words, we are influencing the ML model's future

decisions based on rater-determined, potentially biased ratings, which in return influences workers' future reactions to the ML model's decisions.

By conducting this simulation, we aim to compare how the ML model evolves over time under different models of worker reaction behavior and rater behavior. Specifically, we seek to uncover how these reactions, as a feedback mechanism, influence changes in the ML model's task assignment decisions.

5.2.1 Simulation Setup

The core steps of the simulation are outlined in Algorithm 1, and the number of sections where they are explained in detail are indicated in parentheses.

The algorithm receives a few input parameters that are used throughout the simulation: S_r, the workers' reaction strength to perceived unfairness with respect to retention (an ordinal variable taking values *none*, *mild*, or *strong*); S_q, the workers' reaction strength to perceived unfairness with respect to work quality (an ordinal variable taking values *none*, *mild*, or *strong*); T, the total number of simulation periods (a positive integer); W, the number of workers simulated per period (a positive integer); f, the rater fairness (a categorical variable taking values *fair* or *unfair*); and N, the maximum number of tasks that can be assigned to a worker in a single time period (a positive integer).

The simulation workflow begins at $t = 1$ and proceeds through successive periods until reaching the total number of simulation periods (T), with W workers interacting with the ML model in each time period. An ML model is updated at the end of each period t, once all workers' interaction data has been collected. Therefore, simulation periods also determine the number of updates to the ML model. We simulate two ML models, which are initially identical and fair in their treatment to worker groups. The model workers will interact with is determined by the f parameter, which is given as an input parameter and determines how the ML model is updated. Specifically, f specifies whether the raters provide fair (i.e., unbiased) or unfair (i.e., biased) ratings, which are then incorporated into the model's updates. In other words, one model is updated using the true quality of workers, while the other systematically underestimates the true quality for the disadvantaged group and

overestimates it for the advantaged group. As an example, in this simulation, we assume the disadvantaged group refers to females and the advantaged group refers to males.

Algorithm 1: Simulation Workflow(S_r, S_q, T, W, f, N)

% **Input Parameters:** S_r, S_q, T, W, f, N

% **Updating Parameters:** α^t_{gl}, β^t_{gl} (success rate estimates), ar^t_g, br^t_g (retention parameters), aq^t_g, bq^t_g (quality parameters); $g \in \{M, F\}$, $l \in \{L, H\}$

ParameterInitialization(W, N) ; // (5.2.1.1)

for $t = 1$ **to** T **do**

- `TaskAssigned` $\leftarrow \{\}$;
- `Performance` $\leftarrow \{\}$;
- **for** $w = 1$ **to** W **do**
  - $[g_w, l_w] \leftarrow$ InitializeWorker() ; // (5.2.1.2)
  - $N_w \leftarrow$ AssignTask($g_w, l_w, \alpha^t_{MH}, \beta^t_{MH}, \alpha^t_{FH}, \beta^t_{FH}, \alpha^t_{ML}, \beta^t_{ML}, \alpha^t_{FL}, \beta^t_{FL}, N$);
  - `TaskAssigned` $\leftarrow$ `TaskAssigned` $\cup\{(N_w, g_w)\}$; // (5.2.1.3)
  - $R_w \leftarrow$ DetermineRetention($N_w, g_w, ar^t_M, br^t_M, ar^t_F, br^t_F$) ; // (5.2.1.4)
  - $Q_w \leftarrow$ DetermineQuality($R_w, g_w, l_w, aq^t_M, bq^t_M, aq^t_F, bq^t_F$) ; // (5.2.1.5)
  - $rating_w \leftarrow$ DetermineRating(Q_w, R_w, g_w, f) ; // (5.2.1.6)
  - `Performance` $\leftarrow$ `Performance` $\cup\{(rating_w, R_w)\}$;
- $[\alpha^{t+1}_{MH}, \beta^{t+1}_{MH}, \alpha^{t+1}_{FH}, \beta^{t+1}_{FH}, \alpha^{t+1}_{ML}, \beta^{t+1}_{ML}, \alpha^{t+1}_{FL}, \beta^{t+1}_{FL}]$ $\leftarrow$ UpdateModel($\alpha^t_{MH}, \beta^t_{MH}, \alpha^t_{FH}, \beta^t_{FH}, \alpha^t_{ML}, \beta^t_{ML}, \alpha^t_{FL}, \beta^t_{FL}$, `Performance`) ; // (5.2.1.7)
- $ar^{t+1}_M, br^{t+1}_M, ar^{t+1}_F, br^{t+1}_F \leftarrow$ UpdateRetentionParam($ar^t_M, br^t_M, ar^t_F, br^t_F$, `TaskAssigned`, S_r) ; // (5.2.1.8)
- $aq^{t+1}_M, bq^{t+1}_M, aq^{t+1}_F, bq^{t+1}_F \leftarrow$ UpdateQualityParam($aq^t_M, bq^t_M, aq^t_F, bq^t_F$, `TaskAssigned`, S_q) ; // (5.2.1.8)

For each time period t, we initialize each worker w with their attributes—that is, their group identity g_w (e.g., male/M vs. female/F) and skill level l_w (e.g., low/L vs. high/H).

Then, the ML model decides the number of tasks assigned to worker w (N_w) based on its estimation of their likelihood of successfully completing tasks, which is determined by a Beta distribution parameterized by α^t_{gl} and β^t_{gl} ($g \in \{M, F\}, l \in \{L, H\}$). Worker w then determines their retention and quality—that is, how many of the assigned tasks to complete (R_w) and with what quality to complete those tasks (Q_w). They make these decisions based on the observed difference in task assignment rates between the two worker groups. Specifically, we maintained Beta distributions to simulate worker reactions for both retention (parameterized by ar^t_g and br^t_g, $g \in \{M, F\}$) and quality (parameterized by aq^t_g and bq^t_g, $g \in \{M, F\}$). Finally, based on the ML model the worker has been interacting with, we gather raters' evaluation of the worker's work quality ($rating_w$), which can be either fair or unfair and is determined by f. Once time period t concludes, we update ML model using the collected ratings, and the model adjusts its task assignment decisions for each group for the next time period $t+1$ (i.e., updates α^{t+1}_{gl} and β^{t+1}_{gl}). Moreover, we assume that the summary statistics of the ML model's task assignment to workers of the two groups in period t would be provided to workers in the next period $t+1$. Thus, to reflect how workers would react to the ML model's fairness across the two groups in determining their retention and quality in time $t+1$, upon calculating the task assignment difference between the two groups in time t defined by Z_t, we update the respective Beta distributions that model each group's reactions to the ML model for the next period—that is, the parameters governing their retention and work quality (retention: ar^{t+1}_g, br^{t+1}_g; quality: aq^{t+1}_g, bq^{t+1}_g) are updated accordingly.

5.2.1.1 Parameter Initialization

We initialize the Beta distributions governing both the ML model and the workers. The following sets of parameters are used:

- $\alpha^t_{MH}, \beta^t_{MH}, \alpha^t_{FH}, \beta^t_{FH}$ (Beta distribution parameters determining the ML model's estimation of highly-skilled male/female's success rate by the start of time period t);
- $\alpha^t_{ML}, \beta^t_{ML}, \alpha^t_{FL}, \beta^t_{FL}$ (Beta distribution parameters determining the ML model's estimation of low-skilled male/female's success rate by the start of time period t);

- $ar_M^t, br_M^t, ar_F^t, br_F^t$ (Beta distribution parameters determining a male/female worker's retention in time period t);

- $aq_M^t, bq_M^t, aq_F^t, bq_F^t$ (Beta distribution parameters determining a male/female worker's work quality in time period t)

The initial Beta distribution parameters for the ML model's task success rates, worker retention, and worker quality are set as follows:

- $\alpha_{MH}^1 = 0.8 \times \frac{W \times N}{4}, \quad \beta_{MH}^1 = 0.2 \times \frac{W \times N}{4}, \quad \alpha_{FH}^1 = 0.8 \times \frac{W \times N}{4}, \quad \beta_{FH}^1 = 0.2 \times \frac{W \times N}{4}$
- $\alpha_{ML}^1 = 0.6 \times \frac{W \times N}{4}, \quad \beta_{ML}^1 = 0.4 \times \frac{W \times N}{4}, \quad \alpha_{FL}^1 = 0.6 \times \frac{W \times N}{4}, \quad \beta_{FL}^1 = 0.4 \times \frac{W \times N}{4}$
- $ar_M^1 = 1, \quad br_M^1 = 1, \quad ar_F^1 = 1, \quad br_F^1 = 1$
- $aq_M^1 = 1, \quad bq_M^1 = 1, \quad aq_F^1 = 1, \quad bq_F^1 = 1$

The simulation initializes the parameters for the ML model (α_{gl}^1 and β_{gl}^1, $g \in \{M, F\}, l \in \{L, H\}$) by assuming that the ML model has gathered the worker behavior data from a single time period. Specifically, in our simulation, W workers interact with the ML model in a single period, with each worker is eligible to complete up to N tasks. This results in a maximum of $W \times N$ task observations in a single time period.

To distribute these observations, we assume:

- Gender and skill level are assigned uniformly at random.

- Consequently, each combination of skill level and gender (i.e., highly-skilled and low-skilled workers across the two groups) receives an equal share of the initial observations: $\frac{W \times N}{4}$ observations per combination.

We assume that the true difference in task success rate are driven by skill level rather than group membership. Thus, the Beta distribution parameters are initialized identically across groups, as follows:

- For highly-skilled workers, the initial task success rate is set at 0.8, yielding:

$$\alpha_{(g)H}^1 = 0.8 \times \frac{W \times N}{4}, \quad \beta_{(g)H}^1 = 0.2 \times \frac{W \times N}{4}$$

- For low-skilled workers, the initial task success rate is set at 0.6, yielding:

$$\alpha^1_{(g)L} = 0.6 \times \frac{W \times N}{4}, \quad \beta^1_{(g)L} = 0.4 \times \frac{W \times N}{4}$$

We make these initialization choices to ensure a controlled and unbiased starting point for the simulation. Workers were distributed uniformly across gender and skill levels to eliminate any initial group imbalances. Differences in initial task success rates were determined solely by skill level rather than demographic group membership, ensuring that any disparities arising over time are attributable to the interaction between model updates and human reactions, rather than pre-existing differences. The initial Beta distribution parameters for success rates reflected reasonable competence levels for highly-skilled (80%) and low-skilled (60%) workers.

Moreover, workers' retention and work quality parameters were initialized with uniform Beta$(1, 1)$ distributions, implying no prior assumptions about their reaction. This initialization enables a fair and interpretable baseline from which to observe the endogenous emergence of disparities through the feedback loops between task assignment, worker reactions, ratings, and model retraining.

5.2.1.2 Worker Initialization

After determining the total number of simulation periods (T) and the number of workers to be simulated (W) in one period, we initialize each worker w with the following two features:

$$w \begin{cases} g_w & \in \{\mathrm{M}, \mathrm{F}\}, \\ l_w & \in \{\mathrm{L}, \mathrm{H}\} \end{cases}$$

Each worker is assigned a demographic; in this simulation, we use gender to represent demographic attributes. Workers are randomly assigned to either `M/Male` or `F/Female` for their gender.

In real-world, individuals differ in their skill levels when performing tasks. To capture this, each worker is also assigned a skill level that reflects their task competence: highly-

skilled (H) workers are more likely to complete tasks correctly, while low-skilled (L) workers are less likely to do so.

5.2.1.3 ML-based Task Assignment

Once workers are initialized, the number of tasks assigned to each worker is sampled from a Binomial distribution. Specifically, the number of tasks assigned N_w is drawn as:

$$N_w \sim \text{Binomial}(N, p), \quad \text{where} \quad p = \frac{\alpha^t_{g_w l_w}}{\alpha^t_{g_w l_w} + \beta^t_{g_w l_w}}$$

Here, N is the maximum number of possible tasks that can be assigned, and p is the model's estimated task completion success rate (the model's current estimate of the worker's probability of successfully completing a task), based on the worker w's group identity g_w and skill level l_w. The Binomial sampling captures the idea that each task assignment is an independent Bernoulli trial with success rate p.

5.2.1.4 Worker Retention Determination

Each worker chooses to complete only a fraction of the tasks assigned to them. We assume that this fraction is determined by a Beta distribution, which reflects the worker's retention behavior at each time period.

Specifically, the fraction of tasks completed, denoted as fractTasks, is sampled from:

$$\text{fractTasks} \sim \text{Beta}\left(ar^t_{g_w}, br^t_{g_w}\right)$$

where ar^t_g and br^t_g are parameters associated with the worker's group identity g_w. These Beta parameters are updated over time as workers perceive the fairness of the ML model's task assignment decisions, by observing how tasks are distributed between different groups.

The number of tasks a worker ultimately completes, R_w, is then calculated as:

$$R_w = \text{fractTasks} \times N_w$$

Further details about how the Beta distribution parameters are updated based on perceived fairness are provided in Section 5.2.1.8.

5.2.1.5 Workers' Work Quality Determination

Each worker's work quality is defined as the number of tasks correctly completed out of the total number of tasks they chose to complete. We assume that the proportion of tasks completed correctly is determined by a Beta distribution, reflecting the worker's quality behavior at each time period.

Specifically, the fraction of correctly completed tasks, denoted as fractCorrect, is sampled from:

$$\text{fractCorrect} \sim \text{Beta}\left(aq_{g_w}^t, bq_{g_w}^t\right)$$

where $aq_{g_w}^t$ and $bq_{g_w}^t$ are parameters associated with the worker's group identity g_w. These Beta parameters are updated over time as workers respond to perceived fairness in task assignment.

To incorporate the influence of skill level, if a worker is low-skilled ($l_w = L$), we subtract 0.2 from the sampled fractCorrect. If the result is negative, it is rounded up to zero.

The final number of correctly completed tasks, Q_w, is then calculated as:

$$Q_w = \text{fractCorrect} \times R_w$$

Further details about how the Beta distribution parameters are updated based on perceived fairness are provided in Section 5.2.1.8.

5.2.1.6 Rating of the Completed Tasks Determination

As mentioned earlier, we have two ML models that we select for the simulation (f): one fairly and one unfairly updating. Ideally, a worker's rating should accurately reflect the proportion of tasks they completed correctly out of the total tasks they chose to complete.

The fair updating ML model adheres to this principle, and the true rating of a worker is determined using the following formula:

$$rating_w = \left(\frac{Q_w}{R_w}\right) \times \text{max\_rating} \quad (5)$$

Assuming the maximum possible rating for evaluating a worker's performance is 5, the unbiased (actual) rating is calculated using the formula above. In the simulation, raters under the fair updating ML model will assign this true rating.

However, under the unfair updating ML model, a biased rating is applied. In this case, simulated raters first calculate the true rating, then sample a bias value b from a Beta distribution:

$$b \sim \text{Beta}(2, 1)$$

For male workers, the sampled value b is added to the true rating—resulting in an overestimation. For female workers, b is subtracted from the true rating—resulting in an underestimation. In both cases, the final rating is clipped to stay within the valid range of $[0, 5]$.

5.2.1.7 ML Model Updates

For model updates, we employ a Bayesian updating mechanism to refine the success rates, which determine the probability of assigning a task to a worker. As mentioned in Section 5.2.1.1, our prior beliefs about success rates are modeled using the Beta distribution. Specifically, we represent these priors as a Beta distribution parameterized by the shape parameters α_{gl}^t and β_{gl}^t, such that the mean of the distribution corresponds to the success rate.

As new data becomes available in the form of ratings reflecting workers' fraction of successful task completion, we use a binomial likelihood to capture the number of successful and failed task completions made by workers for each skill level and gender. A key property of the Beta-binomial conjugate prior is that it allows for simple and consistent updates: the

posterior remains a Beta distribution. Specifically, if we observe s successes and f failures, the updated (posterior) distribution becomes:

$$\text{Beta}(\alpha_{gl}^t + s, \beta_{gl}^t + f)$$

This Bayesian updating process is applied iteratively to each α and β parameter defined by skill level l and group identity g, thus systematically refining our estimates of success rates.

For the current simulation period, we compute each worker's contribution to model updates using the following formulas:

$$\alpha_w^t = \frac{\text{rating}_w}{\text{max\_rating } (5)} \times R_w$$
$$\beta_w^t = R_w - \alpha_w^t$$

The rating reflects the number of tasks a worker is considered to have successfully completed and is used to update the model. Next, we aggregate all α_w^t and β_w^t values across all relevant workers $\mathcal{W}_{gl}$—where $\mathcal{W}_{gl}$ denotes the set of workers who belong to group g and skill level l—and add them to α_{gl}^t and β_{gl}^t, respectively, yielding the updated parameters α_{gl}^{t+1} and β_{gl}^{t+1}.

$$\alpha_{gl}^{t+1} = \alpha_{gl}^t + \sum_{w \in \mathcal{W}_{gl}} \alpha_w^t$$
$$\beta_{gl}^{t+1} = \beta_{gl}^t + \sum_{w \in \mathcal{W}_{gl}} \beta_w^t$$

5.2.1.8 Updating Worker Reaction Towards Unfairness

As described earlier, each worker's retention behavior (fraction of assigned tasks they choose to complete) and work quality (fraction of completed tasks completed correctly) are determined by sampling from Beta distributions. Over time, as workers observe how the ML model assigns tasks to different groups, their perceptions of fairness may change,

leading them to adjust their behavior. We model these evolving reactions by updating the corresponding Beta distribution parameters at the end of each simulation period.

At the end of each time period t, we first calculate a standardized measure of perceived (un)fairness based on task assignment disparities between groups. Specifically, we compute the mean (μ_M, μ_F) and standard deviation (σ_M, σ_F) of task assignments for each group and then calculate a Z-statistic:

$$Z^t = \frac{\mu_M - \mu_F}{\sqrt{\sigma_M^2 + \sigma_F^2}}$$

A positive Z^t indicates that males (Group M) are advantaged over females (Group F), receiving more tasks on average; a negative Z^t indicates that females (Group F) are advantaged over males (Group M).

We define two types of worker reaction models to capture different possible real-world responses to perceived fairness or unfairness:

- **Appreciate-Protest Model**: When the advantaged group perceives advantage, their retention and quality increase; when the disadvantaged group perceives disadvantage, their retention and quality decrease. This model captures real-world behavior where individuals who benefit from favorable treatment may reciprocate by engaging more (higher retention and quality), while those who perceive unfair disadvantage may protest by disengaging (lower retention and quality).

- **Slack-Strive Model**: When the advantaged group perceives advantage, their retention and quality decrease (slack off), while the disadvantaged group perceives disadvantage and responds by increasing retention and quality (strive harder). This model captures real-world behavior where individuals who experience privilege may become complacent, leading to reduced effort, while those facing disadvantage may strive harder to overcome perceived barriers.

Under each model, the Beta distribution parameters for retention (ar_g^t, br_g^t) and quality (aq_g^t, bq_g^t) are updated based on the magnitude of the observed $|Z^t|$ and the pre-specified reaction strengths S_r (for retention) and S_q (for quality), where $S_r, S_q \in \{0$ (none), 1 (mild),

2 (strong)}. Specifically, we first check the sign of Z^t to determine which group is advantaged: if Z^t is positive, the advantaged group is M (i.e., males); otherwise, it is F (i.e., females).

In all cases, updates are performed by adding a positive quantity to the appropriate parameter (either a or b), depending on whether the respective group is advantaged or disadvantaged. For example, assuming in time period t, Z^t is positive (i.e., M is the advantaged group and F is the disadvantaged group), the updates are as follows:

- **Appreciate-Protest Model**:
  - Advantaged group (M):

$$ar_M^{t+1} = ar_M^t + S_r \cdot |Z^t|, \quad br_M^{t+1} = br_M^t$$

$$aq_M^{t+1} = aq_M^t + S_q \cdot |Z^t|, \quad bq_M^{t+1} = bq_M^t$$

  - Disadvantaged group (F):

$$ar_F^{t+1} = ar_F^t, \quad br_F^{t+1} = br_F^t + S_r \cdot |Z^t|$$

$$aq_F^{t+1} = aq_F^t, \quad bq_F^{t+1} = bq_F^t + S_q \cdot |Z^t|$$

- **Slack-Strive Model**:
  - Advantaged group (M):

$$ar_M^{t+1} = ar_M^t, \quad br_M^{t+1} = br_M^t + S_r \cdot |Z^t|$$

$$aq_M^{t+1} = aq_M^t, \quad bq_M^{t+1} = bq_M^t + S_q \cdot |Z^t|$$

  - Disadvantaged group (F):

$$ar_F^{t+1} = ar_F^t + S_r \cdot |Z^t|, \quad br_F^{t+1} = br_F^t$$

$$aq_F^{t+1} = aq_F^t + S_q \cdot |Z^t|, \quad bq_F^{t+1} = bq_F^t$$

Changes to the a parameters push the Beta distributions toward higher values, leading to increases in retention or quality, while changes to the b parameters push the distributions toward lower values, resulting in decreases in retention or quality. If Z^t is negative, the advantaged group becomes F (females) and the disadvantaged group becomes M (males), and their parameters are updated accordingly. In either case, the worker models capture distinct behavioral dynamics in response to perceived fairness or unfairness over time.

5.2.2 Simulation Results

We run the simulation with $W = 10{,}000$ workers over the course of $T = 50$ periods. Each worker can be assigned up to $N = 200$ tasks in each period. We simulate both scenarios of worker behavior (i.e., Appreciate-Protest and Slack-Strive Model). In each scenario, we simulate both levels of rater fairness f (i.e., raters' ratings are either unbiased or biased), and we also vary the reaction strengths S_r (retention) and S_q (quality) across three levels (0 = none, 1 = mild, 2 = strong) to generate multiple behavioral response curves reflecting how workers with varying behaviors react to models with different fairness properties, and how these behaviors further affect the model fairness as it evolves. Unless otherwise specified, the plotted figures show the average results of 100 independent simulation runs.

Each simulation result figure is organized as a 3×3 grid. Each column corresponds to a key outcome variable: retention (Column 1), accuracy (Column 2), and number of tasks assigned (Column 3). Each row corresponds to a different type of metric: Row 1 shows the outcome variable difference between groups (male – female), Row 2 shows average values for the disadvantaged group (females), and Row 3 shows average values for the advantaged group (males). Within each plot, blue lines correspond to simulations where raters provide unfair ratings (f = unfair), and orange lines correspond to simulations where raters provide fair ratings (f = fair). Different line styles represent different reaction strengths (S_r and S_q): dotted lines for no reaction (0), dashed lines for mild reaction (1), and solid lines for strong reaction (2).

5.2.2.1 Simulation Results under the Appreciate-Protest Model

Under this worker behavior model, workers from the advantaged group respond to perceived advantage by *appreciating* (increasing their retention and/or quality), while disadvantaged workers respond to perceived disadvantage by *protesting* (decreasing their retention and/or quality). Below, we explore worker reactions through both retention and quality, only through quality, and only through retention, across varying reaction strengths.

Scenario 1: Workers React to Model Unfairness via both Retention and Quality

In this scenario, workers react to unfairness through both retention and work quality. Specifically, S_r and S_q are varied across three levels (none, mild, strong) together to model workers' behavioral sensitivity to perceived fairness differences.

In the first row of Figure 5.1, the difference plots show that under unfair ratings (blue curves), retention, accuracy, and task assignment disparities grow dramatically over time, especially when reaction strength is strong (solid lines). In contrast, under fair ratings (orange curves), group differences remain relatively small even at high sensitivity levels.

These differences are directly driven by changes in group-specific averages, as shown in Rows 2 and 3. Under unfair ratings, disadvantaged workers (females) experience sharp declines in both retention and quality over time, while advantaged workers (males) initially show rapid increases before stabilizing at elevated levels. For example, the widening retention difference between males and females arises because male retention increases sharply and remains high, whereas female retention decreases substantially. Similarly, the growing accuracy gap is fueled by declining female quality paired with improving male quality. As these patterns accumulate, the task assignment system increasingly favors males, given that ratings are biased in favor of males, males are contributing more data, and that data is of higher quality, while the data provided by females shrinks in quantity and quality—amplifying assignment disparities across periods.

Interestingly, under unfair ratings with very strong worker reactions (solid blue lines across first-row plots), task assignment disparities initially grow rapidly but later plateau compared to the mild reaction case (dashed blue lines). This stagnation stems from strongly

reacting disadvantaged workers severely reducing their engagement (low retention), thereby limiting the data available for the ML model to reinforce biases. In contrast, mildly reacting workers (dashed lines) continue interacting with the system for longer, allowing disparities to keep growing, albeit more gradually.

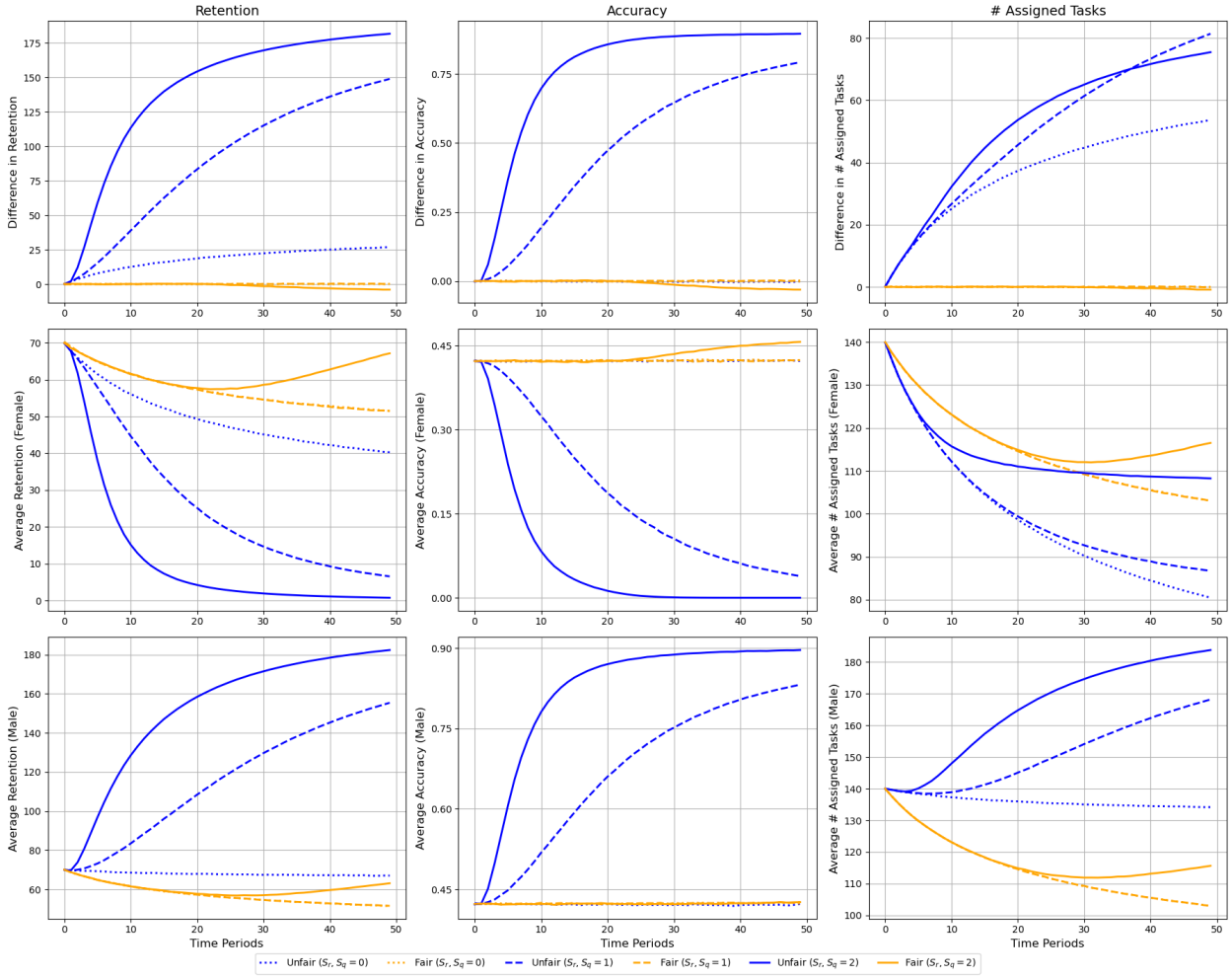


Figure 5.1. Simulation results for the Appreciate-Protest model where workers react via both retention and quality. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages.

Key Insight: Under the Appreciate-Protest model, when workers react to perceived unfairness by adjusting both their retention and quality, group disparities escalate fastest under unfair conditions. However, if reactions are very strong, reduced participation among disad-

vantaged workers can eventually slow or halt further disparity growth by starving the ML model of updating signals. Which means, in the real-world scenarios assigning more tasks to disadvantaged workers when such dynamics are in place, might not necessarily be the right course of action to take.

Scenario 2: Workers React to Model Unfairness via Quality Only

In this scenario, workers react to unfairness only through changes in quality (S_q varied), while their retention remains unaffected ($S_r = 0$).

As shown in Figure 5.2, under unfair ratings (blue curves), quality differences widen steadily over time, especially when the reaction strength is stronger (solid lines). These growing disparities can be directly traced to the group-specific averages shown in Rows 2 and 3. Disadvantaged workers (females) experience a consistent decline in their average work quality, while advantaged workers (males) maintain or improve their quality over time. For fair ratings (orange curves), the quality difference between groups remains minimal throughout. Notably, under strong reactions in the fair condition, the average quality for females slightly increases over time. This trend is not driven by fairness concerns, but rather by small fluctuations in task assignments. Because workers are modeled as highly sensitive to perceived disparities, even minor variations—despite an overall fair environment—can lead them to believe they are advantaged (or disadvantaged), prompting increases (or decreases) in their work quality.

Under unfair ratings, as reaction strength increases and workers adjust their work quality more strongly in response to perceived unfairness, task assignment disparities also grow, as evolving differences in quality influence rater evaluations and, subsequently, the ML model's feedback mechanism. Since retention levels remain stable for both groups—meaning the fraction of assigned tasks they choose to complete stays consistent regardless of perceived fairness—sufficient data is generated for both groups, although the quality of male workers' data is higher. Combined with biased ratings favoring males, the ML model increasingly favors male workers over time, leading to cumulative assignment advantages.

Unlike scenarios where workers react by reducing their retention, here workers remain engaged in the system while reacting only through quality adjustments. This continuous

data flow enables the ML model to keep evolving, steadily embedding and amplifying fairness gaps across periods. In contrast, under fair ratings, quality differences stay small and task assignment remains relatively balanced over time.

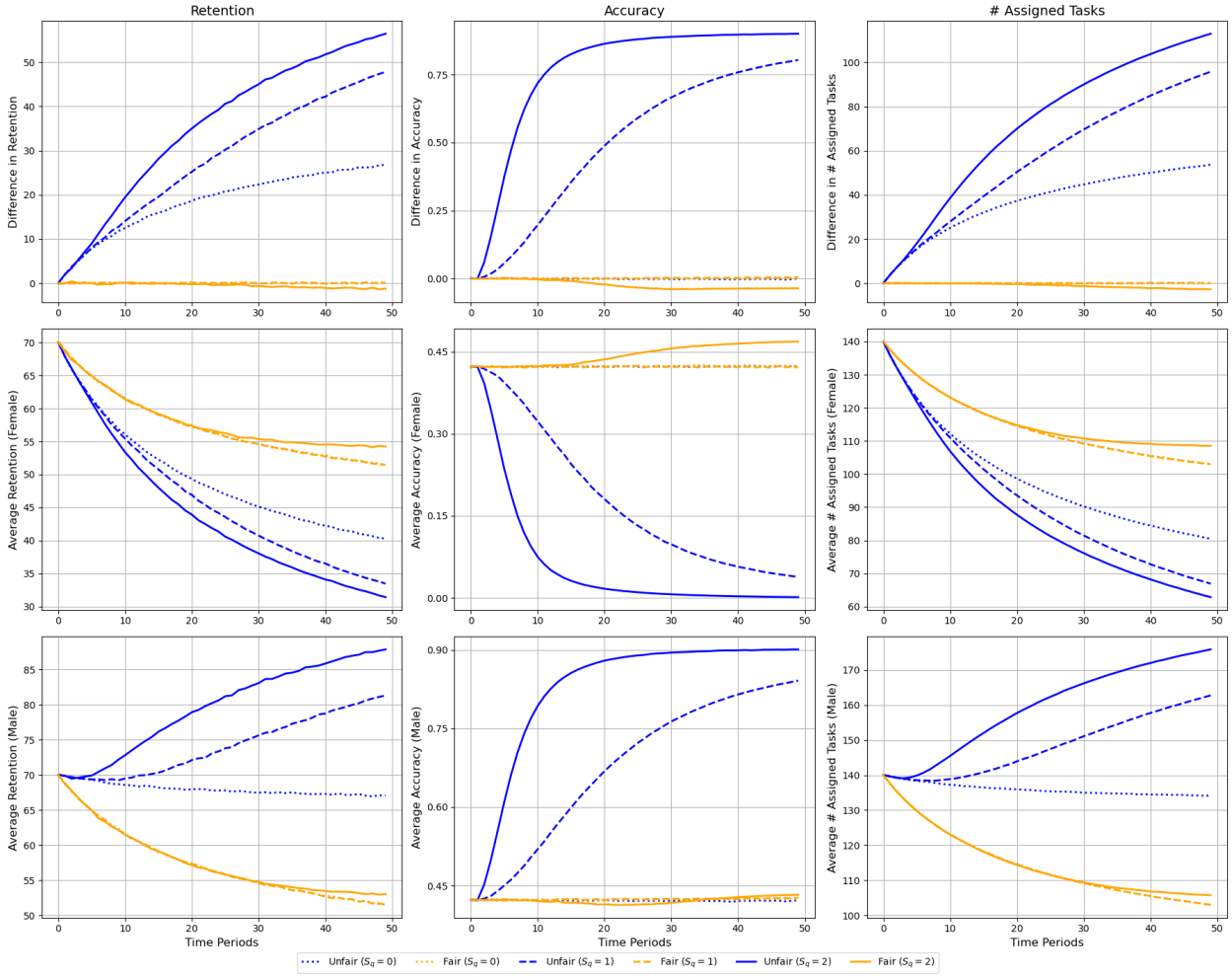


Figure 5.2. Simulation results for the Appreciate-Protest model where workers react via quality only. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages.

Key Insight: Under the Appreciate-Protest model, when workers react to unfairness only through changes in quality while remaining in the system, unfairness exacerbates over time under unfair ratings, especially as reaction strength increases. Sustained engagement en-

ables the ML model to reinforce and magnify perceived fairness disparities over repeated interactions.

Scenario 3: Workers React to Model Unfairness via Retention Only

In this scenario, workers react to perceived unfairness only through changes in retention (S_r varied), while their work quality remains unaffected ($S_q = 0$).

As shown in Figure 5.3, under unfair ratings (blue curves), retention differences between groups widen dramatically over time, especially for stronger reactions (solid lines). In contrast, under fair ratings (orange curves), retention differences remain relatively minor and stable across time. Since quality behavior is not reactive in this setting, accuracy differences stay negligible under both fair and unfair ratings. Task assignment disparities under unfair ratings initially increase alongside the growing retention differences, but eventually plateau as fewer disadvantaged workers (females) remain active in the system to drive further updates. Under fair ratings, task assignment disparities remain consistently small.

These patterns are further explained by the group-specific averages shown in Rows 2 and 3. Under unfair conditions, disadvantaged workers (females) exhibit sharp declines in their retention levels over time, particularly at stronger reaction strengths. Meanwhile, advantaged workers (males) maintain higher and more stable retention. Because of this asymmetry, males continue receiving task assignments while the participation of females drops, explaining the divergence in task assignment rates initially. However, as female participation diminishes severely over time, there is not enough updated data to sustain further assignment bias, causing the task assignment gap to stabilize.

When comparing across reaction strengths, we observe that stronger reactions cause the highest retention disparities over time. However, because disadvantaged workers exit earlier, the resulting data loss slows down the divergence in task assignments over time, leading it to plateau. In contrast, milder reactions allow disadvantaged workers to stay in the system longer, resulting in slower but more prolonged divergence in task assignments.

Compared to scenarios where workers react via quality (or via both quality and retention), the evolution of fairness gaps here is more self-limiting. Once disadvantaged workers leave

the system, the ML model can no longer update unfairly against them. Specifically, this highlights an important real-world implication: while strong protest through non-participation (i.e., retention reduction) can protect disadvantaged workers from prolonged exploitation, it also freezes inequalities that have already formed. Since the underlying evaluation mechanism remains biased, continued participation may only reinforce further unfair treatment. Thus, interventions such as granting additional tasks (e.g., in response to appealing unfairness) may inadvertently backfire—by feeding more data into a biased model, they risk amplifying disparities rather than resolving them.

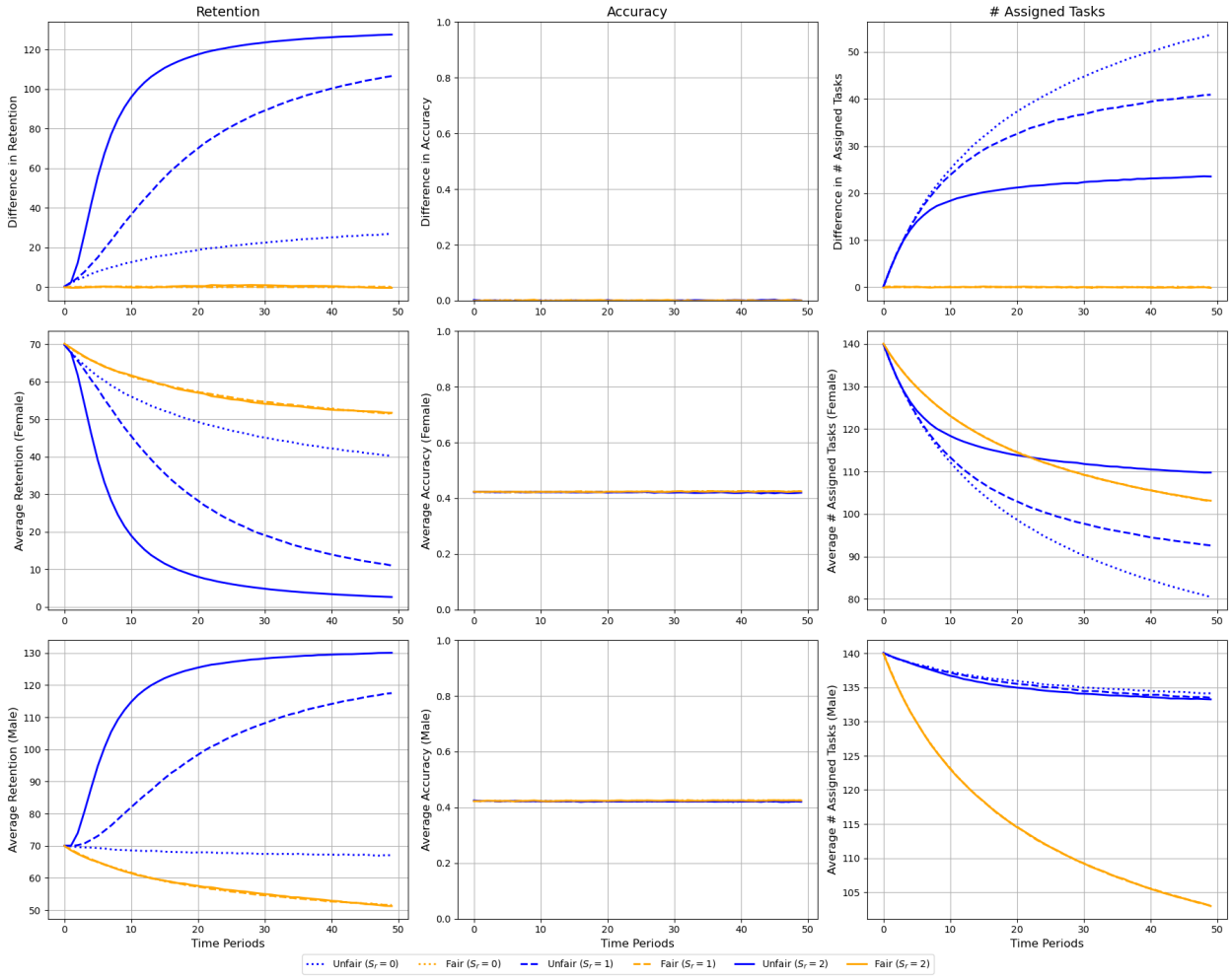


Figure 5.3. Simulation results for the Appreciate-Protest model where workers react via retention only. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages.

Key Insight: Under the Appreciate-Protest model, when workers react to perceived unfairness solely through changes in retention, fairness gaps initially grow but then plateau over time. Stronger reactions lead to stabilization, as reduced participation among disadvantaged groups ultimately halts model evolution, "locking in" disparities at biased levels.

5.2.2.2 Simulation Results under the Slack-Strive Model

Under this worker behavior model, workers from the advantaged group respond to perceived advantage by *slacking off* (reducing their retention and/or quality), while disadvantaged workers respond to perceived disadvantage by *striving harder* (increasing their retention and/or quality). As before, we explore worker reactions through both retention and quality, only through quality, and only through retention, across varying reaction strengths.

Scenario 1: Workers React to Model Unfairness via both Retention and Quality

In this scenario, workers react to perceived unfairness by adjusting both their retention and quality, following the Slack-Strive behavioral model. Specifically, S_r and S_q are varied across three levels (none, mild, strong) to model workers' behavioral sensitivity to perceived fairness differences.

In the first row of Figure 5.4, under unfair ratings (blue curves), retention, accuracy, and task assignment differences initially grow as reaction strength increases. However, unlike the Appreciate-Protest model, the growth slows down significantly over time, and in some cases even reverses for task assignment disparities under strong reactions. In contrast, under fair ratings (orange curves), group differences stay consistently small or moderately increase across all periods, regardless of the reaction strength.

Specifically, examining the group-specific averages in Rows 2 and 3 reveals that, under unfair ratings, disadvantaged workers (females) gradually increase both their retention and quality over time, while advantaged workers (males) gradually decrease theirs. The task assignment difference initially widens due to biased ratings that favor male workers, leading

the model to assign them more tasks. However, as female workers steadily improve their participation and performance—while male workers' engagement and quality decline—the gap begins to slow and eventually narrows in later periods, as disadvantaged workers start to significantly outperform advantaged ones. This self-correcting pattern is distinctive to the Slack-Strive model and is not observed under the Appreciate-Protest dynamics.

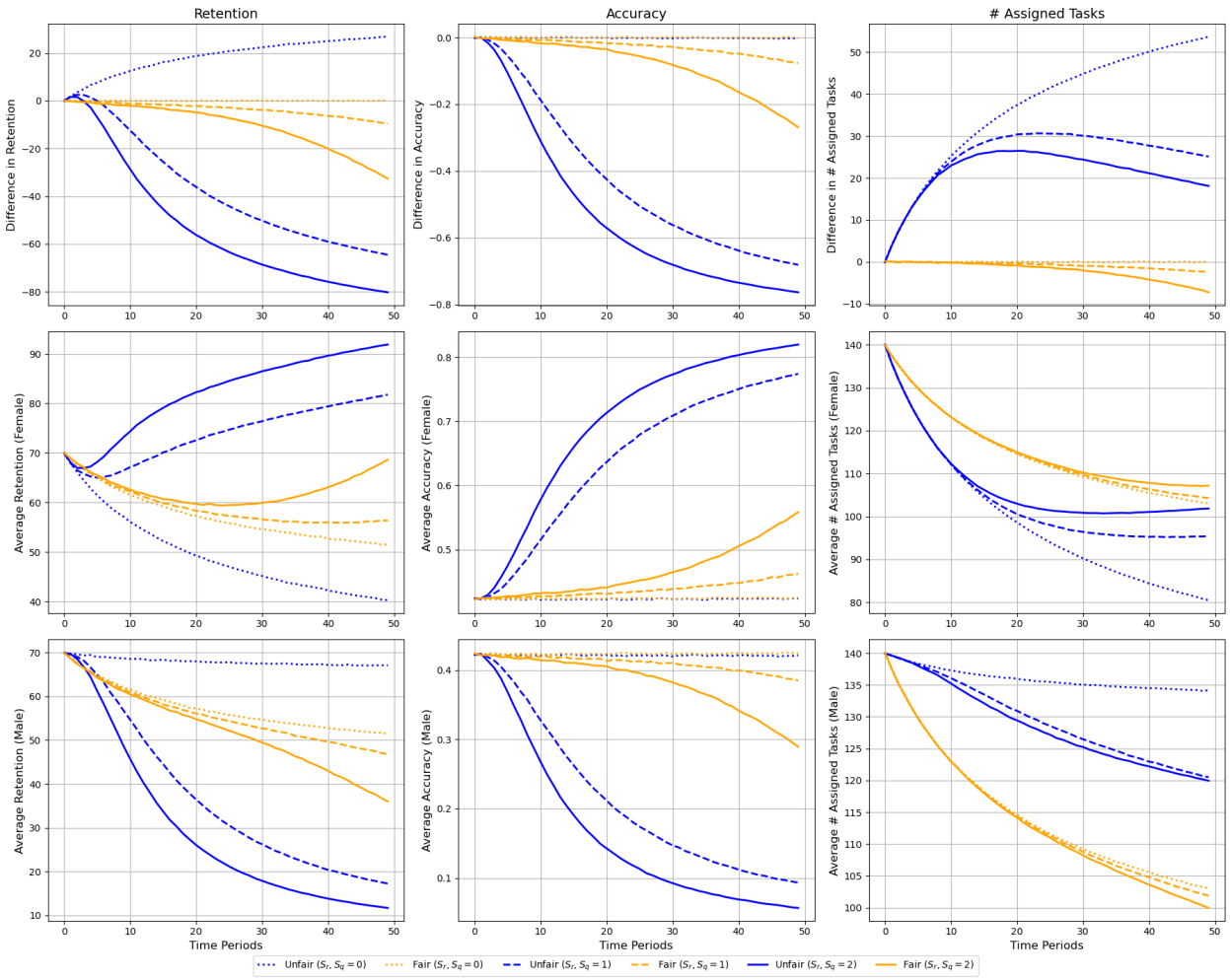


Figure 5.4. Simulation results for the Slack-Strive model where workers react via both retention and quality. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages.

When considering the role of reaction strength, stronger reactions (solid lines) not only accelerate the initial widening of differences in retention and quality but also hasten the subsequent reversal in task assignments. Specifically, under strong Slack-Strive reactions,

disadvantaged workers strive harder more quickly, while advantaged workers slack off more rapidly, leading to an earlier and more pronounced narrowing of task assignment gaps.

Compared to scenarios under the Appreciate-Protest model, where disparities tended to escalate unchecked or plateau, the Slack-Strive behavioral dynamic shows a self-correcting mechanism: disadvantaged workers' strategic efforts to strive help mitigate unfairness amplification over time.

Key Insight: Under the Slack-Strive model, when workers react to perceived unfairness by adjusting both retention and quality, fairness disparities under unfair ratings initially grow but later slow down or even reverse. Strong striving by disadvantaged workers, combined with slacking by advantaged workers, creates a self-correcting feedback loop that mitigates long-term amplification of unfairness. However, this dynamic depends critically on continued engagement from disadvantaged workers and their persistent efforts to outperform despite unfair initial conditions.

Scenario 2: Workers React to Model Unfairness via Quality Only

In this scenario, workers react to perceived unfairness only through changes in quality (S_q varied), while their retention remains unaffected ($S_r = 0$).

As shown in Figure 5.5, under unfair ratings (blue curves), accuracy differences between groups widen substantially over time, especially when reaction strength is higher (solid lines). While the initial widening of task assignment disparities is primarily driven by biased ratings that favor male workers, emerging quality gaps—where female workers begin to outperform males—gradually counteract this bias, especially as reaction strength increases. Over time, these performance differences help slow down and partially correct the disparity in task allocations. Retention differences observed are primarily a byproduct of evolving disparities in task assignments—since the fraction of assigned tasks they choose to complete stays consistent regardless of perceived fairness. In contrast, under fair ratings (orange curves), task assignment differences remain relatively small and stable throughout the simulation. However, for quality, some highly sensitive workers increase their quality as they perceive

small differences between the task assignments of both groups as unfairness and increase their quality, which creates the accuracy differences across groups.

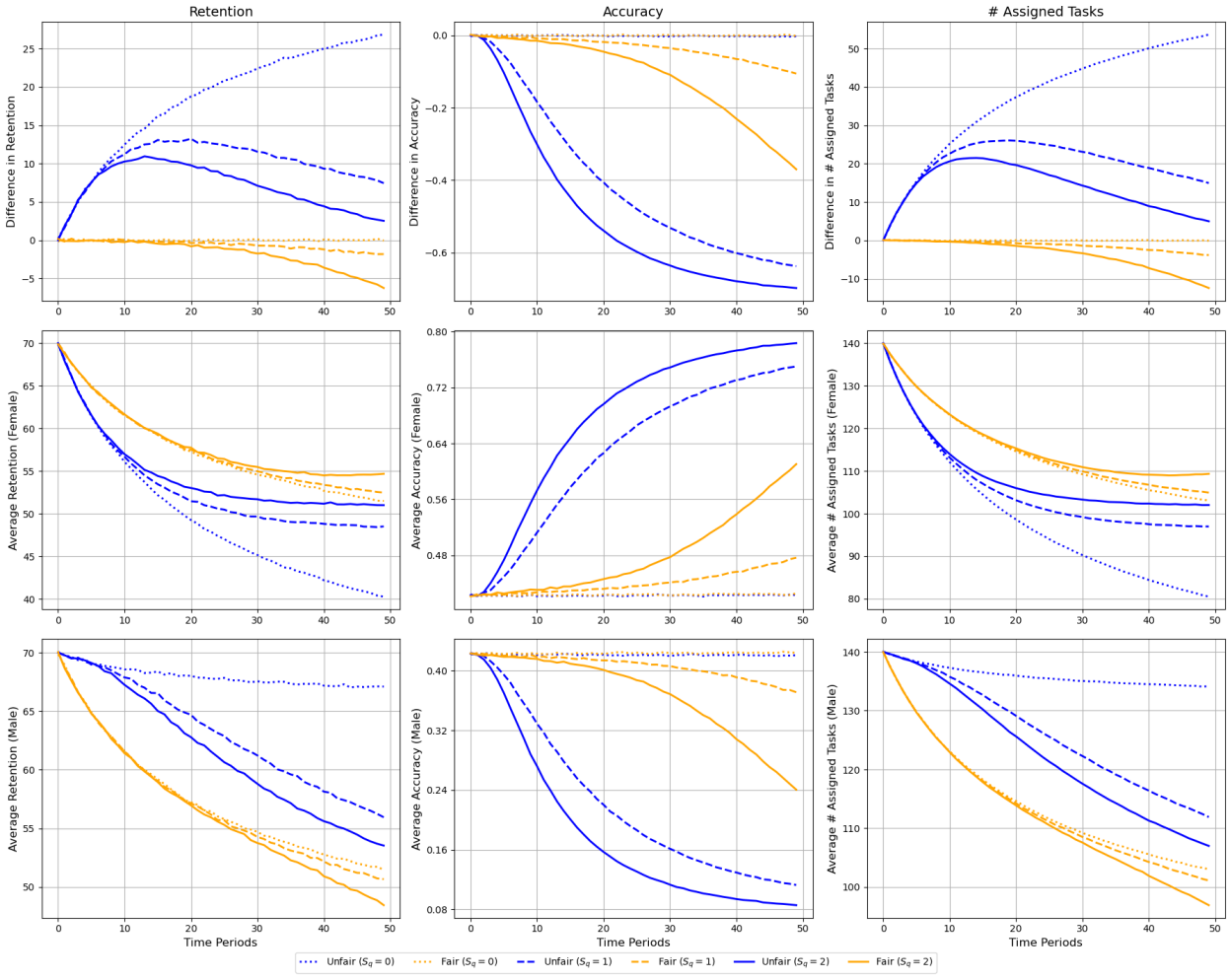


Figure 5.5. Simulation results for the Slack-Strive model where workers react via quality only. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages over time.

Examining the group-specific averages (Rows 2 and 3) reveals that under unfair conditions, disadvantaged workers (females) gradually improve their quality over time, while advantaged workers (males) experience a steady decline. This dynamic allows disadvantaged workers to slowly offset the initial bias embedded in task assignment decisions. As a result, although task assignment disparities initially widen due to biased ratings, they eventually stabilize or even begin to narrow as the disadvantaged group starts to outperform

the advantaged group in quality. Specifically, because workers remain engaged and continue generating data, the feedback loop shifts: male workers contribute increasingly lower-quality data while female workers provide increasingly higher-quality data, leading to a corrective adjustment in the model's task assignment decisions over time. This gradual reversal dynamic is distinct from the Appreciate-Protest model, where disparities tend to escalate indefinitely without self-correction. Reaction strength plays an important role in shaping this trajectory. Stronger reactions (solid lines) accelerate later reversals. That is, stronger Slack-Strive reactions enable disadvantaged workers to more rapidly improve their performance relative to advantaged workers, helping to mitigate assignment disparities earlier in the simulation.

These findings highlight an important real-world implication: in systems where disadvantaged participants remain engaged and respond to unfairness by improving their quality, while advantaged participants gradually slack off, initial unfairness can be corrected over time through sustained striving. However, the speed and extent of this correction are heavily influenced by the strength of participants' reactions, with stronger striving accelerating the reversal of disparities.

Key Insight: Under the Slack-Strive model, when workers react to perceived unfairness only through changes in quality while remaining engaged in the system, disadvantaged workers' striving quality and advantaged workers' slacking quality can gradually mitigate task assignment disparities over time, even in initially biased environments. Stronger quality reactions accelerate this corrective dynamic, enabling disparities to be addressed more quickly.

Scenario 3: Workers React to Model Unfairness via Retention Only

In this scenario, workers react to perceived unfairness only through changes in retention (S_r varied), while their work quality remains unaffected ($S_q = 0$).

As shown in Figure 5.6, under unfair ratings (blue curves), retention differences initially grow, especially as reaction strength increases (solid lines). In contrast, under fair ratings (orange curves), retention differences between groups remain limited over time. Task assignment disparities under unfair ratings still steadily increase over time, particularly when reactions are strong. Meanwhile, for the fair ratings all task assignment differences remain

at similar levels for all conditions. Examining the group-specific averages (Rows 2 and 3), we observe that under unfair conditions, disadvantaged workers (females) maintain relatively higher retention rates over time compared to advantaged workers (males), who exhibit sharper declines in participation. This striving-versus-slacking pattern creates an imbalance: although disadvantaged workers stay more engaged, the biased rating system and the resulting retention differences progressively lead to larger assignment gaps over time.

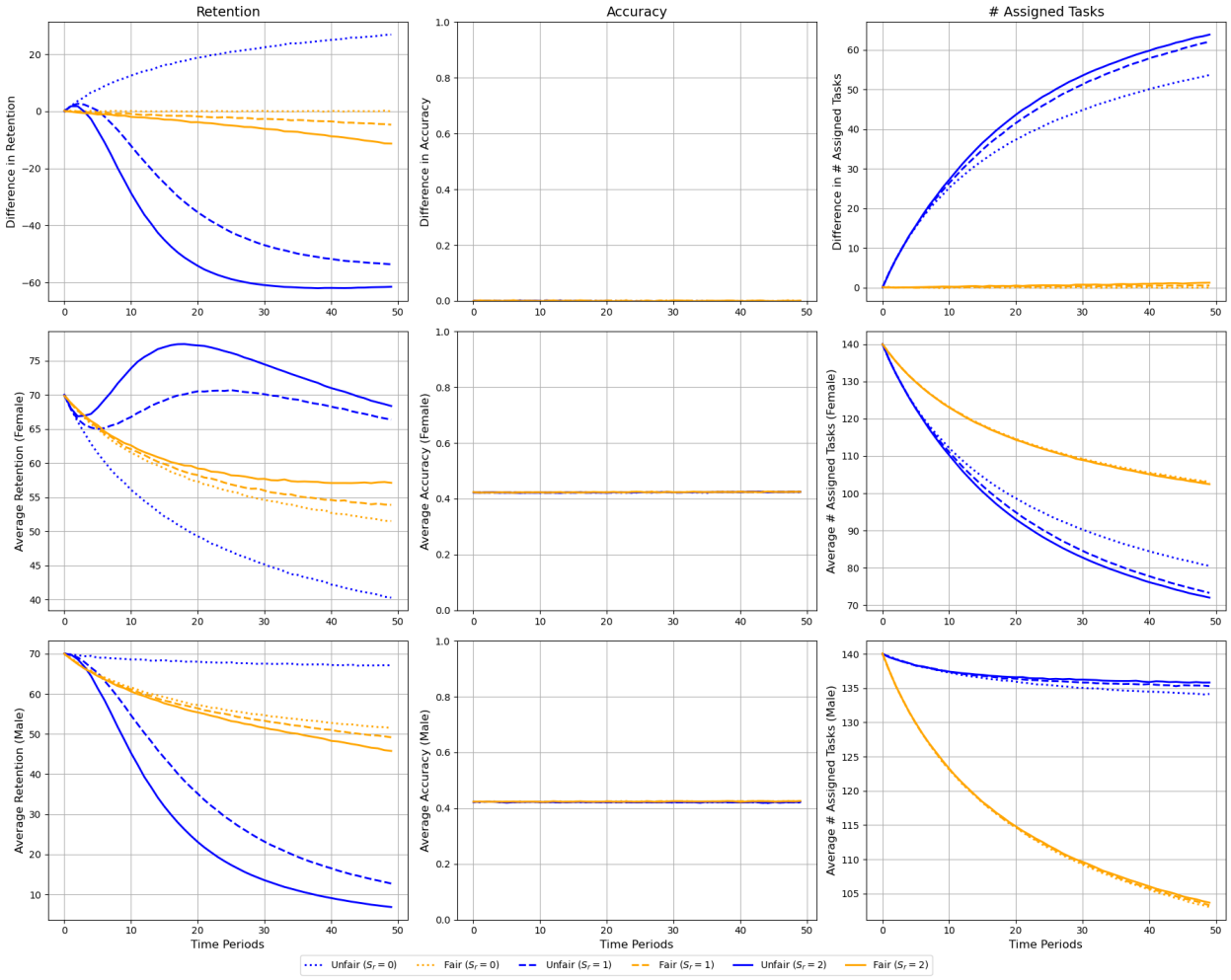


Figure 5.6. Simulation results for the Slack-Strive model where workers react only via retention. Retention difference, accuracy difference, and assigned task difference between males and females ($Males - Females$) are shown, along with group-specific averages.

When considering the role of reaction strength, we observe that in the no-reaction case under unfair ratings (blue dotted line), retention differences between groups are primarily

driven by disparities in task assignments, rather than by active worker responses to unfairness. In contrast, stronger reactions (solid lines) accelerate the early widening of retention gaps, which indirectly amplifies task assignment disparities across repeated periods. Specifically, although disadvantaged workers strive to remain engaged, the system—shaped by biased ratings and the declining participation of advantaged workers—continues to disfavor females in task assignments through a reinforcing feedback loop. Critically, because disadvantaged workers stay without simultaneously improving their work quality, there is no counteracting high-quality signal to disrupt the biased updates, allowing fairness gaps to persist and even widen over time.

Key Insight: Under the Slack-Strive model, when workers react to perceived unfairness only through changes in retention while remaining engaged, disadvantaged workers' higher participation alone is insufficient to correct bias. Instead, it can unintentionally exacerbate assignment disparities under unfair ratings, particularly when reaction strength is high.

5.2.2.3 Simulation Results for Appealing Interventions

To explore how a specific intervention—allowing workers to appeal for their perceived unfairness in task assignment and assigning extra tasks to disadvantaged workers—affects group disparities under biased model update dynamics, we simulate a setting where disadvantaged workers can "appeal" fairness issues and potentially receive additional tasks as compensation. We set f = unfair, meaning biased ratings are used for ML model updates. The appealing intervention mechanism activates when the observed task assignment disparity from the previous time period, $|\mu^{t-1}_{male} - \mu^{t-1}_{female}|$, which also includes extra tasks granted to the disadvantaged group, exceeds a predefined threshold (e.g., 10). Disadvantaged workers may appeal perceived unfairness with probability $p \in \{0, 0.1, 0.3, 0.5, 0.8\}$, and, if triggered, receive additional tasks drawn from a uniform distribution Extra $\sim \mathcal{U}[\tau - v, \tau + v]$, with $\tau = |\mu^{t-1}_{male} - \mu^{t-1}_{female}|, v = 10$.

We model two types of behavioral responses: retention-based ($S_r = 1, S_q = 0$) and quality-based ($S_r = 0, S_q = 1$). Each behavioral response is analyzed under both the Appreciate-Protest and Slack-Strive worker models to examine how different responses, com-

bined with the appealing interventions, influence retention, accuracy (quality), ML-assigned tasks, total task assignments (ML + Extra), and retention normalized by total task assignments.

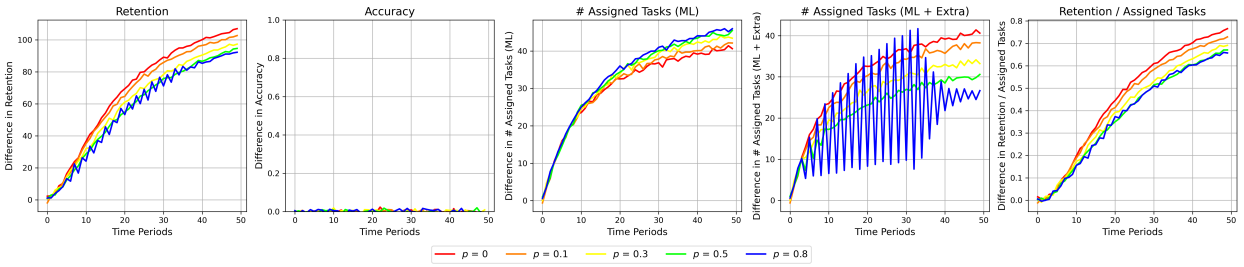


Figure 5.7. Appreciate-Protest model with retention-based reactions ($S_r = 1, S_q = 0$). Group differences ($M - F$) are tracked over five appealing intervention probabilities. Oscillations in the ML + Extra task assignment difference show the limitations of the intervention, providing a non-consistent relief.

It is important to note that the results presented here are from a single representative simulation run rather than an average across multiple runs, in contrast to previous simulations. Averaging across multiple runs obscures the dynamic feedback mechanisms of the system. In particular, in conditions where the appealing probability is high (e.g., $p = 0.8$), oscillations in assignment disparity (ML + Extra) arises due to the conditional nature of the intervention mechanism—extra tasks are assigned only when disparity exceeds a threshold. In some runs, this condition is met; in others, it is not. Averaging these outcomes results in misleadingly smooth curves that mask the underlying dynamics. Presenting a single representative run allows us to capture and interpret the structural feedback patterns more faithfully, especially when the intervention mechanism induces alternating compensation and non-compensation phases across time steps—as we will see in the simulations below.

Scenario 1: Workers React to Model Unfairness via Retention Only

In the Appreciate-Protest model (Figure 5.7), retention differences widen over time as advantaged (male) workers remain engaged while disadvantaged (female) workers increasingly disengage under unfair task allocations. When extra tasks are granted through appeals, disadvantaged workers tend to stay longer and complete more tasks, temporarily reducing disparities in total task assignments (ML + Extra) as the p values increase. However, when

workers frequently appeal and receive additional tasks—particularly under $p = 0.8$—this triggers oscillations in assignment disparity. The system alternates between assigning and withholding extra tasks depending on whether the disparity in the previous round exceeds the intervention threshold. This cyclical pattern emerges because compensation in one round often closes the gap just enough to suppress further intervention in the next. Yet, the increased participation of disadvantaged workers leads to more biased data being fed into the model, which further exacerbates disparity in subsequent rounds. This feedback loop results in a temporary and unstable pattern of fairness correction.

Looking at trends over time, both retention and ML + Extra task disparities exhibit clear oscillations until approximately $t = 40$. For retention, even the peaks of disparity under high p values remain lower than those under lower p values. In contrast, for ML + Extra task assignments, the peaks of oscillation under high p values (e.g., $p = 0.8$) can produce larger disparities than when no appeals are allowed ($p = 0$), while the troughs reduce disparity below all other conditions. After $t = 40$, oscillations disappear, and both retention and ML + Extra disparities begin to increase again. However, the rate of disparity growth slows as p increases, indicating that higher appeal rates continue to offer some relief. Nevertheless, for ML-only task assignments (subplot 3), disparity increases steadily with higher p across the entire time range, due to increasingly biased model updates. In contrast, for normalized retention (retention divided by assigned tasks), the level of disparity decreases—although they are in increasing trend—consistently as p increases, suggesting that additional tasks allow disadvantaged workers to remain more proportionally engaged.

Finally, examining subplots 3 and 4 in detail, in subplot 3 (ML-only task assignments), higher p values lead to greater disparities, suggesting that allowing disadvantaged workers to appeal can inadvertently worsen the ML model's internal bias over time. This occurs because increased appeal rates retain more disadvantaged workers, who continue to be evaluated unfairly, and these biased ratings then reinforce disparities in subsequent model updates. Subplot 4 shows that total task assignment disparities (ML + Extra) can be somewhat mitigated, particularly at moderate p levels. However, this compensation is unstable. In the case of $p = 0.8$, the early-phase disparity even exceeds that of $p = 0$, due to intervention cycles overshooting or triggering inconsistently. These fluctuations ultimately diminish

as disparities widen beyond the corrective capacity of the intervention. Once model bias becomes too strong, even the maximum compensatory effort cannot restore parity below the fairness threshold. This reveals a structural limitation of relying on task-based appealing interventions alone, as they offer only short-term relief while the underlying unfairness continues to grow. Task-based fairness interventions—while they may somewhat reduce disparities in the short term—seem to be not capable of neutralizing the compounding bias embedded in the ML model's learning process. Over time, the burden of fairness is shifted to the compensatory mechanism, which becomes insufficient to restore equity as model bias deepens.

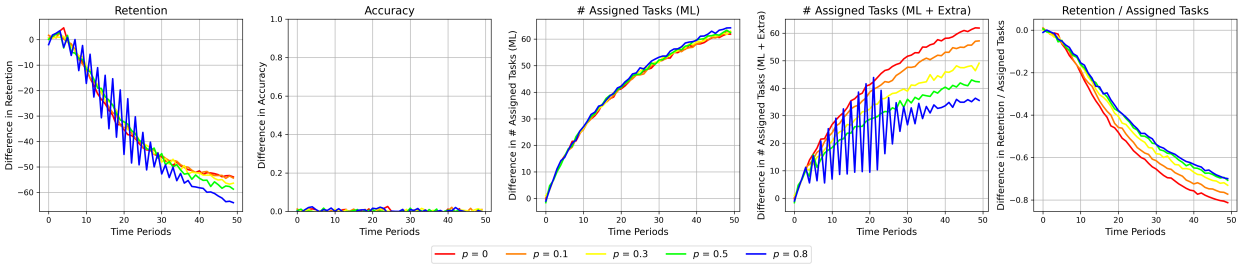


Figure 5.8. Slack-Strive model with retention-based reactions ($S_r = 1, S_q = 0$). Group differences ($M - F$) are tracked over five appealing intervention probabilities. Under this worker model, females stay longer but increasing disparity in final task assignments cannot be controlled over time, even if maximum number of extra tasks are given to female workers.

In the *Slack-Strive* model (Figure 5.8), retention differences widen over time, with disadvantaged (female) workers showing higher retention compared to advantaged (male) workers who increasingly disengage. In other words, in this model female workers remain active and strive under biased task allocations. The retention gap becomes most pronounced under high appeal probabilities (e.g., $p = 0.8$), where frequent compensation encourages continued engagement for female workers. Subplot 4 again shows oscillatory behavior in the ML + Extra task assignment disparity under high p values, reflecting instability in the fairness intervention. When disparity exceeds a threshold, compensation is triggered, restoring parity temporarily. But once parity is achieved, the system halts intervention, allowing disparities

to grow again—creating alternating phases of compensation and non-compensation that fuel the observed oscillations.

Looking at trends across time, the oscillations in both retention and ML + Extra task assignments persist until approximately $t = 30$, especially under $p = 0.8$, where both peaks and troughs are highly pronounced. For retention, the disparity under $p = 0.8$ fluctuates between being the lowest and the highest among all p values. For ML + Extra task assignments, the disparity under $p = 0.8$ similarly oscillates between being lower than all other p values and equaling the disparity observed when $p = 0$. After $t = 30$, the oscillations subside, yet overall disparities in both retention and total task assignments continue to increase. For raw retention, disparities become larger as p increases, suggesting that assigning more tasks to disadvantaged workers encourages them to remain engaged longer. And, for the ML-only task assignment disparities (subplot 3), they steadily increase, showing that the model continues to favor the advantaged group as biased updates accumulate over time.

Specifically, in subplot 3, ML-only task assignment disparities increase steadily over time, and higher p values appear to have little to no effect on altering this trend. Since the model continues to incorporate biased ratings from the engaged disadvantaged group, the disparity accumulates at a similar rate across all p values. This feedback loop reinforces unfair allocations rather than correcting them. In contrast, subplot 4 shows that total task assignment disparities (ML + Extra) are better lowered under moderate p values, but become unstable under high appeal probabilities due to alternating phases of compensation and non-compensation. In early periods, the disparity often drops just below the intervention threshold after compensation, leading to no compensation in the next round and causing the system to revert to ML-only assignments. This back-and-forth dynamic produces visible oscillations, and while the disparities remain bounded, they fluctuate more widely under high p. Although higher p values do eventually lead to reduced disparities later in the simulation—compared to lower p values—the long-term reliance on appealing interventions shifts the responsibility of fairness correction to external interventions, rather than addressing the root cause—bias in the model updates. With higher p values, the ML model strictly increases its bias against disadvantaged workers over time.

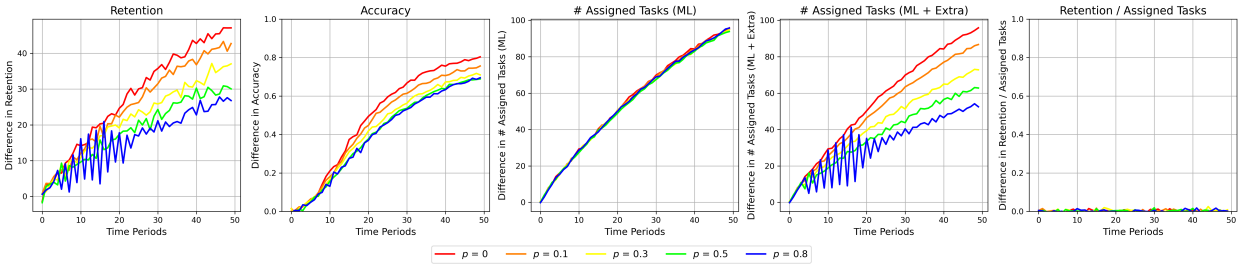


Figure 5.9. Simulation results under the Appreciate-Protest model with quality-based reactions ($S_q = 1, S_r = 0$). Group differences ($M - F$) are tracked over five appealing intervention probabilities. Disadvantaged workers (females) do not change their retention behavior but reduce work quality over time in response to perceived unfairness. As appealing probability p increases, compensatory task assignments for a very short time reduce assignment disparities but also trigger oscillations. However, due to continued biased feedback and decreasing quality signals, disparities quickly grow beyond the intervention's ability to correct.

Scenario 2: Workers React to Model Unfairness via Quality Only

In the *Appreciate-Protest* model (Figure 5.9), where workers respond by adjusting work quality rather than retention, retention differences are driven primarily by disparities in task assignments rather than voluntary disengagement. Over time, male workers receive more total task assignments than female workers, leading to retention differences in favor of males (subplot 1). These disparities widen consistently with decreasing appeal probability, indicating that higher p values (e.g., $p = 0.8$) help reduce—but not eliminate—retention disparities. Subplot 4 shows oscillatory behavior in ML + Extra assignments under high p, reflecting alternating phases of compensation and inaction triggered by the fairness threshold. Although increasing p values decreases the disparity over time, these efforts still result in males receiving more total tasks overall. In terms of accuracy (subplot 2), disadvantaged workers (females) consistently reduce their work quality over time as a form of protest, leading to a growing accuracy gap in favor of males. The gap is most severe when $p = 0$, but narrows slightly as p increases—likely due to the intervention providing more tasks to females and slightly reducing the need to protest. Nevertheless, the quality gap remains substantial across all appeal probabilities.

In subplot 4, ML + Extra task assignment disparities also favor males throughout. Under high p values (especially $p = 0.8$), we observe visible oscillations in the early periods, caused by the conditional nature of the compensation mechanism. When disparity exceeds the threshold, the intervention assigns extra tasks, temporarily closing the gap; but this is followed by a period of non-compensation when the disparity falls below the threshold. This cycle of over- and under-correction continues until approximately $t = 20$, after which the oscillations begin to stabilize. Although the level of disparities become smaller as p value increases, the overall disparity continues to increase gradually in all p value trends, suggesting that the intervention cannot fully counteract the compounding effect of protest-induced low quality.

Finally, subplots 3 and 4 underscore the structural limitations of the intervention. ML-only task assignments (subplot 3) reveal strictly increasing disparities over time, regardless of appeal probability, as the biased model continues to favor males. High p values do somewhat decrease the disparity in ML + Extra task assignments with oscillations early on. But, for the ML model assignments, the overall strict increase is due to additional data from disadvantaged workers—who are delivering low-quality work with increasing data—feeds back into the model and reinforces existing bias.

In the *Slack-Strive* model with quality-based reactions (Figure 5.10), retention differences are again primarily driven by disparities in task assignments, as workers in this setting do not adjust their retention behavior. Subplot 1 shows retention gaps in favor of males, which emerge and stabilize over time. As the p value increases—particularly at $p = 0.8$—oscillations in the retention gap become more pronounced, fluctuating between the lowest and highest disparities across all p values. Similarly, in subplot 4, ML + Extra task assignment disparities exhibit strong oscillations under high p, where the system frequently alternates between assigning and withholding extra tasks. Despite these fluctuations, the overall disparities remain bounded, suggesting that the feedback loop is constrained and does not cause strict increase in disparities over time. In terms of accuracy (subplot 2), the behavioral assumptions of this worker model cause the advantaged group (males) to begin to slack off—gradually reducing their quality—while the disadvantaged group (females) responds by striving and improving their work quality. This results in a growing accuracy gap, where females outperform males.

Notably, under high p (especially $p = 0.8$), the accuracy gap narrows slightly compared to lower p values. This may be because the compensatory task assignments reduce the perceived need for disadvantaged workers to overcompensate through extra effort, softening their striving behavior.

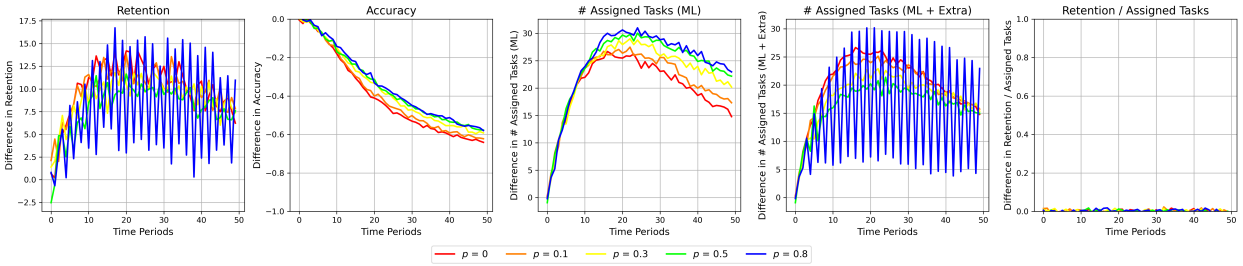


Figure 5.10. Slack-Strive model with quality-based reactions ($S_q = 1, S_r = 0$). Group differences ($M - F$) are tracked over five appealing intervention probabilities. Female workers (disadvantaged group) gradually improve their work quality in response to unfairness, while male workers (advantaged group) reduce theirs over time. As a result, assignment disparities first grow and then partially narrow. When appealing is enabled, oscillations in total task assignment (ML + Extra) always appear but remain within bounds due to the self-correcting nature of the model and behavior.

Unlike previous cases, there is no clear attenuation in the oscillations over time for either retention or ML + Extra task assignments. As p increases, the magnitude of oscillations grows larger—especially under $p = 0.8$—causing disparities to alternately become the worst and best across all p values.

Subplots 3 and 4 show that task assignment disparities initially increase under higher p, but eventually start to decrease. This is due to the partial self-correction of the *Slack-Strive* model: as the quality of disadvantaged workers improves, the model updates begin to reflect this, reducing disparity in ML-only assignments. However, the intervention's impact is mixed. Subplot 3 indicates that assigning extra tasks under high p inadvertently increases disparities in ML-only assignments by amplifying the influence of still-biased updates. Subplot 4 shows that ML + Extra assignment disparities, while bounded, are unstable and can swing to extremely high values under high p. Thus, in a setting where workers strive to improve, the compensation mechanism yields mixed effects. While high levels of appealing

($p = 0.8$) introduce instability and may disrupt the system's natural self-correction, moderate reporting probabilities appear to help mitigate disparities in final task assignments. This suggests that the intervention does not uniformly undermine fairness—rather, its effectiveness depends critically on the rate at which workers appeal.

5.3 Human Subject Experiment: Experimental Design

To complement our simulation-based findings and better understand how real people respond to evolving ML-driven decisions that affect them, we conducted a randomized human-subject experiment on Prolific. This experiment was designed to reflect real-world algorithmic management scenarios, where crowd workers interact with ML algorithms used by online platforms for task assignment. In particular, we explored how introducing mechanisms that provide workers with more agency—such as the ability to skip tasks or appeal unfair assignments—might shape their behavioral reactions (e.g., such as retention) and perceptions of fairness as they repeatedly interact with these models. Furthermore, we focused on a setting where models are updated based on simulated biased ratings of worker performance, in order to capture real-world challenges where human biases infiltrate algorithmic decision-making, giving rise to feedback loops that influence workers' behaviors and perceptions of fairness, which are then used to update the model further over time.

Our experiment is designed to address the following research questions:

- **RQ1**: As workers interact with ML-based task assignment models under different platform mechanisms (e.g., allowing task skipping or appeals), how do the fairness properties of these models evolve over time and compare to one another?
- **RQ2**: How do changes in the fairness properties of ML-based task assignment models affect workers' retention behavior and the quality of their work under different platform mechanisms?
- **RQ3**: How do evolving fairness properties of ML-based task assignment models influence workers' perceptions of model fairness across different platform mechanisms?

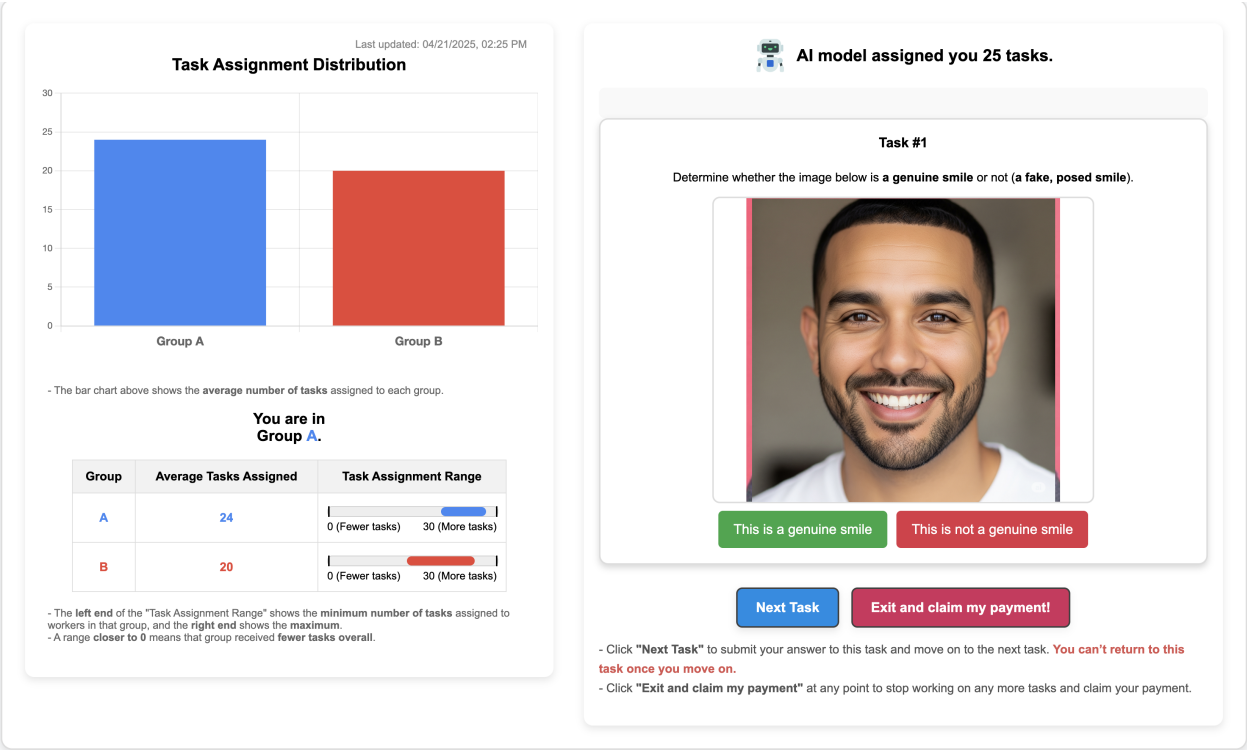


Figure 5.11. Example experiment main view that workers interacted.

5.3.1 Experimental Tasks

In this experiment, we recruited workers to perform an image annotation task in which they classified images of smiling faces as either "genuine" or "fake" (i.e., based on whether the smile engaged the eye muscles). We gathered a dataset of 75 images of genuine and 75 images of fake smiles. An ML model then assigned workers up to 30 tasks, with each task involving the annotation of a single image. Workers were uniformly randomly assigned to one of two groups—Group A or Group B—and the ML model determined the number of tasks each worker received based on their group membership and its estimation of different groups of workers' success rate for this task. By randomly assigning group identities instead of using real demographic characteristics (e.g., "gender"), we ensured that each group had a similar distribution of relevant attributes (e.g., skill level) across tasks. Moreover, when the ML model applied unfair task assignment, it was genuinely unfair because the two groups had similar characteristics.

After tasks were allocated, workers proceeded to a main interface displaying two key views as shown in Figure 5.11: on the left, they saw a dynamic visualization of the ML model's average task assignments to both groups; on the right, they viewed their individually assigned tasks. The left view presented a bar chart showing the average number of tasks assigned by the ML model to each group, alongside line plots indicating the range from the minimum to maximum number of tasks assigned within each group, enabling a clear comparison of both the average assignments and the variability between groups. The left view also indicated the user's group identity. Meanwhile, the right view displayed the total number of tasks assigned to the worker by the ML model at the top, while the bottom part showed the annotation tasks that the workers could complete in a sequential manner. For each task, the workers viewed the image and determined whether it showed a genuine smile. In most cases, workers were required to complete tasks sequentially, submitting a label before proceeding to the next task; however, in some treatments, they were allowed to skip tasks (see Section 5.3.2 for details). At any point, workers had the option to stop completing annotation tasks and exit the study while still being able to claim their payment.

Depending on their assigned treatment, workers could either skip certain tasks or raise complaints about negative experiences they encountered with the ML model. Further details about these are provided in the Section 5.3.2.

5.3.2 Experimental Treatments

Workers were randomly assigned to one of the following three experimental treatments:

1. **Control Treatment**: Workers received task assignments from the ML model and completed them sequentially, with each task requiring submission before the next one became available. Workers were not given any extra options to influence their task flow or workload. For this treatment, we used the regular main experiment view shown in Figure 5.11.

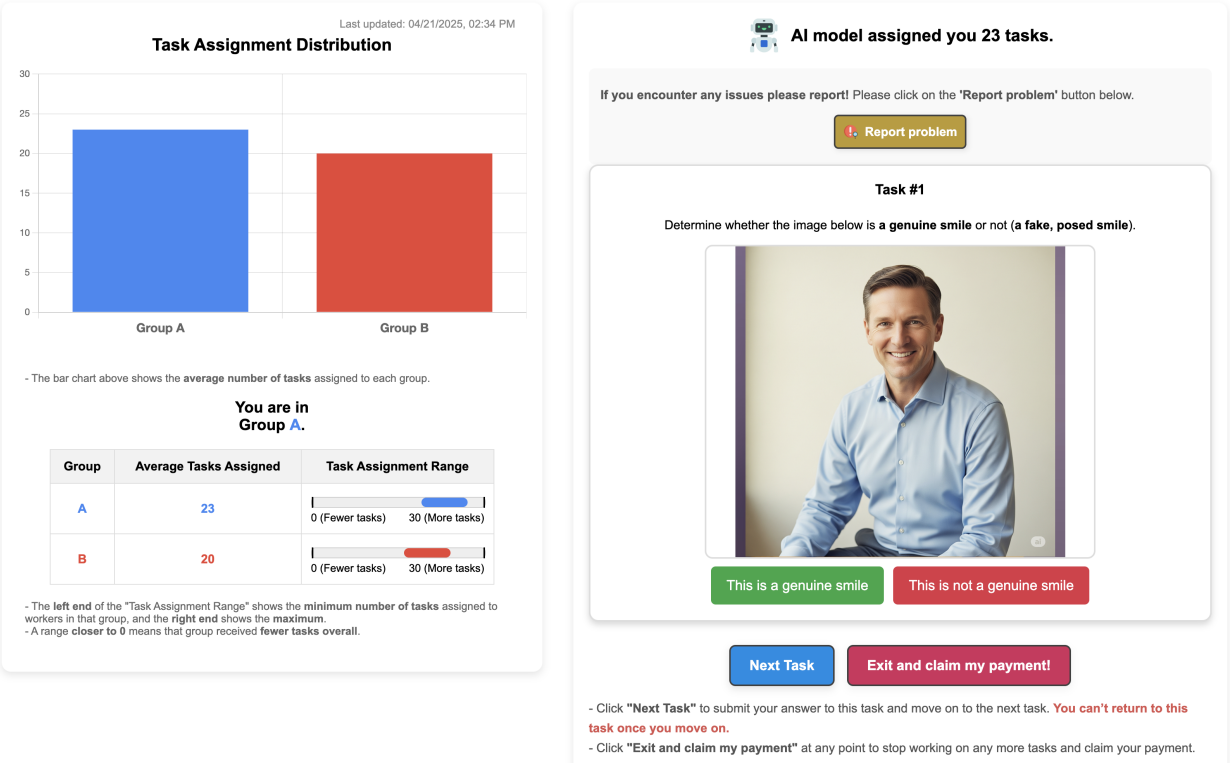


Figure 5.12. Example Main View for "ML Reporting" treatment.

2. **ML Reporting Treatment**: Workers were allowed to report perceived unfair treatment by the ML model. If a complaint was deemed valid—that is, if the worker's group had been assigned significantly fewer tasks (a difference of more than two tasks compared to the other group)—the worker was granted additional tasks (randomly selected between 11 and 20) to offset the biased allocation.

   To facilitate reporting, a "Report Problem" button was introduced, as shown in Figure 5.12. This button allowed workers to select from a predefined list of issues they might have experienced during their interaction with the ML model, such as lack of transparency, biased task allocation, or feelings of dehumanization. If the reported issue related to fairness and was considered valid (i.e., the workers' group received on average at least two tasks fewer than the other group), additional tasks were assigned to the worker.

There are two main reasons we introduced this treatment. First, it reflects real-world platform practices that incorporate grievance mechanisms, enabling workers to report perceived unfair treatment and seek remediation. These mechanisms are essential for providing workers with a sense of agency in algorithmic management systems and may shape both their behavior and perceptions of fairness over time. Second, based on the simulation findings presented in Section 5.2.2.3, this treatment allows us to test those results in a real-world setting, evaluating whether similar feedback dynamics and fairness implications emerge in practice.

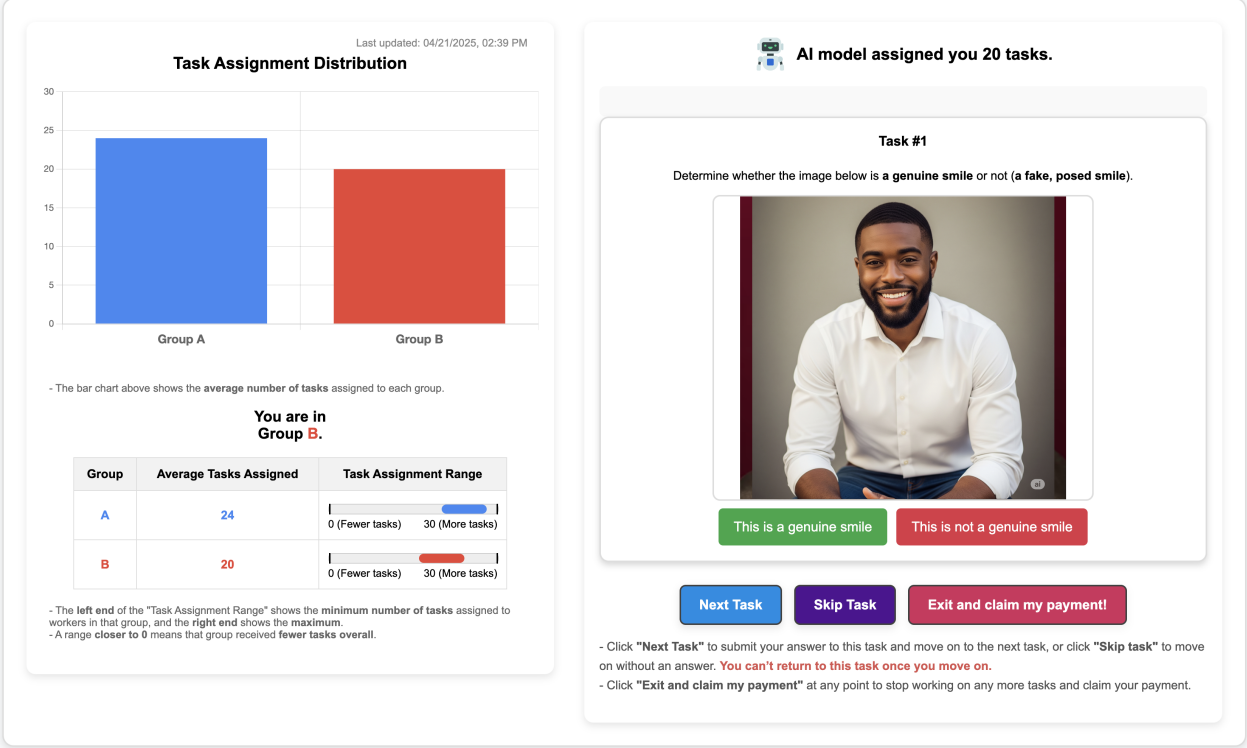


Figure 5.13. Example Main View for "Skip Tasks" treatment.

3. **Skip Tasks Treatment**: Workers were allowed to skip tasks they felt unqualified to complete. Unlike the other two treatments, which required completing each task before progressing, this treatment allowed workers to bypass any tasks they wished to skip. This option enabled workers to exercise greater agency over their task selection, potentially improving their overall work quality by focusing on tasks they felt more confident about. We introduced this treatment to reflect real-world platform mecha-

nisms that allow workers to reject or skip tasks based on perceived suitability, which may influence their engagement levels and perceptions of fairness. For this treatment, we made a minor modification to the main experiment view by adding a "Skip" button option to skip tasks, as shown in Figure 5.13.

5.3.3 ML Model Initialization

All three treatments were powered by their own independently maintained ML model, each initialized similarly to the setup used in the simulation. To initialize each model's belief about group performance (i.e., the likelihood of successful task completion for a particular group), we assumed we had obtained one day of behavioral data from workers and used success/failure counts to parameterize a Beta distribution, Beta(α_g^t, β_g^t) ($g \in \{A, B\}$ refers to worker group being either Group A or Group B), whose mean value $\mu_g = \frac{\alpha_g}{\alpha_g + \beta_g}$ reflected the model's belief about the group g's success rate.

Table 5.1. Initial snapshots of the ML model used for all three treatments. Each group starts with $\alpha + \beta = 990$, assuming 33 workers completing 30 tasks each.

Group	α	β	Mean Success Rate (μ)
A	693	297	0.70
B	693	297	0.70

Specifically, in each day of the experiment, we planned to recruit approximately $200/3 \approx 67$ workers per treatment. Since each treatment involved two groups (based on group identity), each group would receive about 33 workers in one day of experiment. Since a maximum of 30 tasks per worker is possible, this resulted in a total of $30 \times 33 = 990$ trials per group, which were used to initialize the corresponding α_g^t and β_g^t values. To ensure consistent starting conditions across treatments and groups, we set the mean value of the Beta distribution to $\mu_g = 0.7$ for all cases (see Table 5.1). This corresponded to an initial belief that, on average, a worker from any group would complete 70% of their assigned tasks correctly. Based

on this estimated success rate, the ML model determined the number of tasks to allocate to each worker and updated its estimates over time as it obtained more behavioral data from workers.

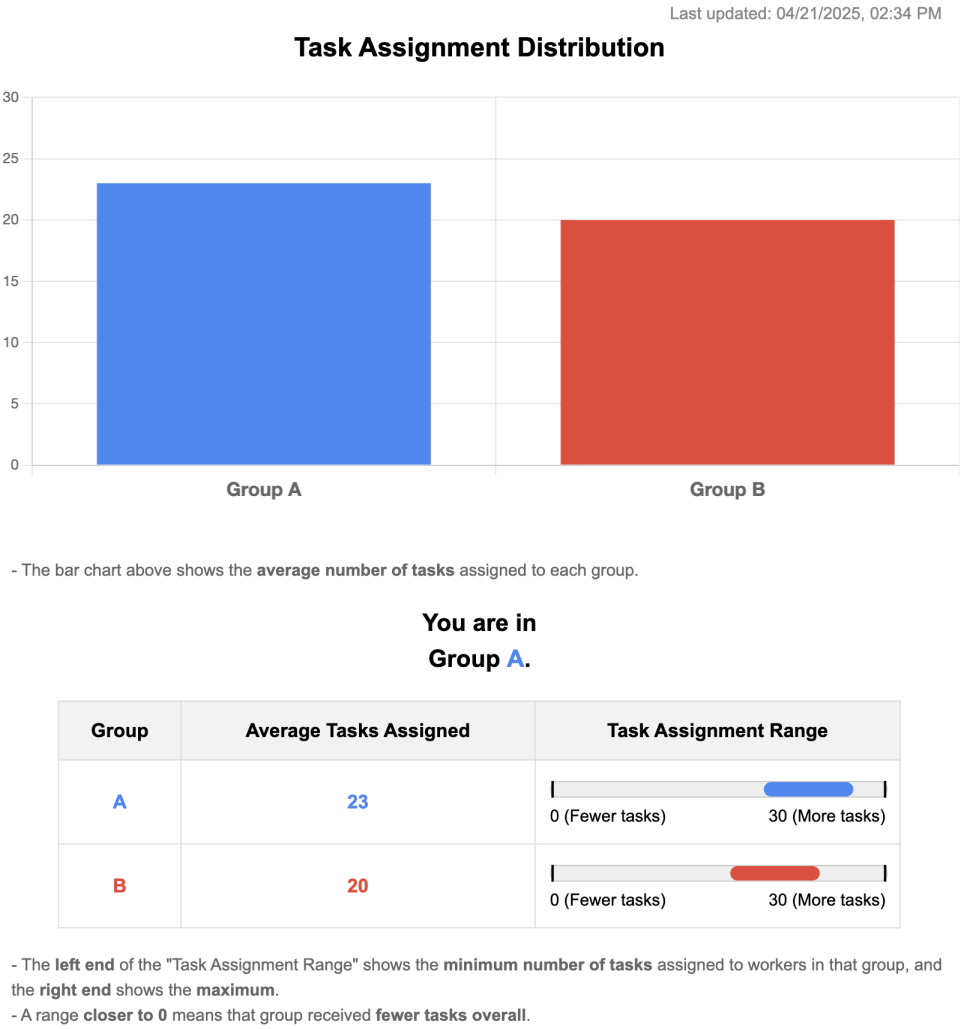


Figure 5.14. Task Assignment Visualization

5.3.4 Experimental Procedure

We launched our study over a period of 4 days, with each day recruiting about 200 workers. Upon their arrival at our study, workers were informed that they would act as gig workers on a platform that distributes microtasks through an algorithmic management system. Each worker was randomly assigned to one of three treatments (i.e., Control, ML

Reporting or Skip Tasks Treatment), with each treatment corresponding to a separate ML model. Workers were also assigned to one of two group identities (Group A or Group B). This group assignment was persistently stored such that if the same worker returned on a later day, they would continue with the same group and treatment.

The ML models were initialized using Beta distributions as described in Section 5.3.3. When a worker entered the study, the server immediately captured a snapshot of the current ML model for their assigned treatment and group. This snapshot—consisting of the group g's current α_g^t and β_g^t values—was then saved for the worker and used to determine their task assignment as well as to populate the bar chart and table visualizations showing task distribution. Specifically, we used these visualizations to help workers understand how the model made task assignment decisions and whether those decisions were fair (see Figure 5.14 for an example). This visualization displayed the average number of tasks assigned to each group, along with the task assignment range. These real-time statistics enabled workers to observe disparities—or parity—in task distribution between groups and adapt their behavior accordingly. To generate these statistics, we simulated 200 workers, each randomly assigned to a group identity. The ML model corresponding to the actual worker's assigned treatment was used to allocate tasks to the simulated workers. The resulting task distribution data was used to generate real-time visualizations shown to workers during their session, allowing them to observe how the ML model was assigning tasks across different groups. A timestamp was also displayed alongside these visualizations, indicating the time of the current snapshot.

Workers then proceeded to the task interface, where they were asked to classify images as either showing a Duchenne (genuine) smile or not. They could complete as many tasks as they wanted to, up to the number of tasks assigned by the ML model and were able to exit at any point by simply clicking the "Exit and claim my payment!" button. Workers were compensated with a \$1 base payment and a \$0.01 bonus for every task they completed. Upon completing their tasks, each worker was evaluated using a biased rating mechanism consistent with the simulation (see Section 5.2.1.6); that is, we drew from a Beta$(2, 1)$ distribution and added the sampled value to the true rating of Group A (or subtracted it from the true rating of Group B). This rating determined the quality score for the worker, which in turn defined the number of perceived "correctly" versus "incorrectly" completed tasks. These outcomes

were used to compute α_w and β_w values for the worker w, which were then added to the corresponding ML model on the server (see Section 5.2.1.7 for model update details). In this way, ML models were updated in real time as workers completed the study, ensuring continual evolution of the task assignment logic.

To add on and reiterate, each treatment had its own workflow. In the Control Treatment, the workers sequentially completed tasks. In the ML Reporting Treatment, workers were provided with a "Report Problem" button to express concerns regarding their experience. If the concern was related to perceived unfair task distribution and the group-level difference in mean assigned tasks exceeded a predefined threshold (e.g., 2 tasks), additional tasks were assigned to the disadvantaged group to offset the disparity. The number of tasks granted in response was set to a maximum of 20 tasks. In the Skip Tasks Treatment, workers could simply skip tasks they deemed irrelevant or unpreferred.

At the end of the study, workers completed a two-part survey. The first part collected demographic information, including age, gender, education level, and race or ethnicity. The second part assessed workers' preferences regarding algorithmic management (i.e., their comfort with ML models assigning tasks instead of humans), perceptions of the model's fairness, and general sensitivity to fairness (i.e., the extent to which individuals value fairness in decision-making systems). The survey included both Likert-scale items and open-ended questions, and two attention checks were embedded to ensure data quality.

5.3.5 Analysis Methods

Our dependent variables begin with *assigned tasks*, defined as the total number of tasks assigned to each worker by the ML model, ranging from 0 to 30. *Retention fraction* is defined as the number of tasks a worker chose to complete divided by their total assigned tasks. For example, if a worker was assigned 15 tasks and completed 12, their retention fraction would be computed as $\frac{12}{15}$. *Accuracy* refers to the fraction of completed tasks that were annotated correctly, calculated as the number of correctly completed tasks divided by the total number of completed tasks. *Perceived fairness* is measured by summing a worker's responses to

survey questions evaluating how fair they found the ML model; the score ranges from 0 to 24, with scores above 12 indicating a positive perception of fairness.

5.3.5.1 ML Model Dynamics

To examine how the ML models evolved, we extracted each worker's model parameters governing their task assignment probabilities (α_g^t and β_g^t). The expected task assignment probability was computed as the mean of the corresponding Beta distribution:

$$\mu = \frac{\alpha_g^t}{\alpha_g^t + \beta_g^t}$$

We plotted the expected percentage of tasks assigned over each worker arrival order, separately for Group A and Group B, to visualize how assignment dynamics diverged across treatments.

5.3.5.2 Across-Group Disparity Visualization

To analyze temporal trends in group-level disparities across treatments, we used a sliding window approach to compute and visualize the average differences in dependent variables between Group A and Group B. Within each treatment, workers were first chronologically ordered based on their task start time. We then constructed overlapping sliding windows of a fixed size (60 workers per window), shifting forward by one worker per step. For each window, we separately extracted the dependent variable values of Group A and Group B workers and computed the mean difference (Group A - Group B). To estimate the variability of this gap, we applied non-parametric bootstrapping within each window: we independently sampled with replacement from Group A and Group B values in the current window (bootstrap sample size equaled number of A and B workers). We then computed the difference of sample means, and repeated this process 1,000 times. From the resulting empirical distribution, we extracted the 2.5th and 97.5th percentiles to form the 95% confidence interval of the group gap. These bootstrapped mean differences and confidence intervals were plotted over sliding window indices to capture how group disparities evolved over time for each treatment.

5.3.5.3 Statistical Analysis

To examine how dependent variables varied by treatment, group identity, and over time, we conducted a series of linear mixed-effects regressions. For each dependent variable, we specified a mixed-effects model with fixed effects for group (A vs. B; with A as the reference category), treatment (Control, ML Reporting, Skip Tasks; with Control as the reference category), and day (treated as a continuous numeric variable from 1 to 4), along with all interaction terms.

Models were fit using restricted maximum likelihood (REML), and significance was assessed using Wald z-tests with 95% confidence intervals. Separate models were fit for each dependent variable using the following general specification:

$$\mathrm{DV}^{\mathrm{i}}_{w} \sim \mathrm{Group}_{w} \times \mathrm{Treatment}_{w} \times \mathrm{Day}_{w}$$

For worker w, the model specification above, with $\mathrm{DV}^{\mathrm{i}}_{w}$, represents each of the four dependent variables i, which are analyzed in separate models.

5.4 Experimental Results

We collected a total of 789 data points from 717 unique subjects. Approximately 200 workers participated each day, resulting in a balanced sample across the four-day period. Specifically, we recorded 274 responses under the Control Treatment, 255 under the ML Reporting Treatment, and 260 under the Skip Tasks Treatment. Among those in the ML Reporting Treatment, 50 workers reported fairness issues—30 from Group A and 20 from Group B. Additionally, 6 workers skipped tasks with 4 from Group A and 2 from Group B.

5.4.1 RQ1: Effects on the ML Model's Task Assignment Fairness Evolvement

To understand how fairness in task allocation evolved over time, we conducted two complementary analyses: one based on the *expected task assignment* inferred from the ML models' internal "belief" updates of how well a worker from a particular group would successfully complete a task, and another based on the *actual task assignments* observed in the exper-

imental data. The expected assignment analysis reveals how the ML model's assignment decisions evolve theoretically based on biased input signals, while the actual assignment analysis captures how these decisions translated into real task distributions experienced by workers. The key difference between the two is that actual task assignments may reflect greater noise and randomness typical of real-world systems, whereas theoretical assignments represent the model's idealized expectations based solely on its internal learning dynamics.

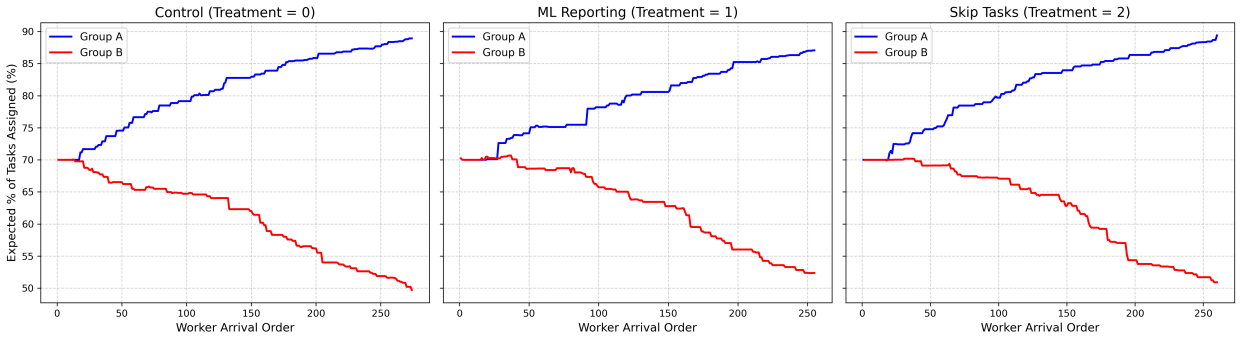


Figure 5.15. Expected percentage of tasks assigned over time for Group A and Group B across treatments.

Figure 5.15 illustrates how the *theoretical* number of tasks assigned to each group evolved over time across the three treatments, computed using the ML model's estimated success rates at each worker's arrival. By sorting the workers with respect to their arrival time and then using their ML snapshot α_g^t and β_g^t values, we calculate and plot $\mu = \alpha_g^t/(\alpha_g^t + \beta_g^t)$ as "Expected % of Tasks Assigned." In all treatments, workers from Group A received an increasing number of tasks over time, while Group B's task assignments steadily declined. This divergence emerged as a result of biased ratings used to evaluate worker performance, despite the models being initialized with identical success rates for both groups.

Figure 5.16 shows how the theoretical difference in expected task assignment rates between Group A and Group B evolved over time under each treatment condition. To generate this figure—similarly as the Figure 5.15—we sorted workers by their arrival order and calculated the group-level difference in expected task assignment rates using the ML model's belief parameters at the time of each worker's arrival. Specifically, we computed $\mu = \alpha_g^t/(\alpha_g^t + \beta_g^t)$ for each group and plotted the difference (Group A minus Group B) across treatments. As

seen in the figure, the expected assignment gap consistently grows over time in all three treatments, highlighting the compounding effect of biased feedback. While the Control and Skip Tasks treatments show the steepest increases in assignment disparity, the ML Reporting treatment exhibits a comparatively slower growth curve. This pattern suggests that allowing workers to report perceived unfairness provided some, albeit limited, mitigation against the widening gap. Still, the intervention was insufficient to fully correct the model's biased evolution, as Group A continued to receive a disproportionately larger share of expected task assignments.

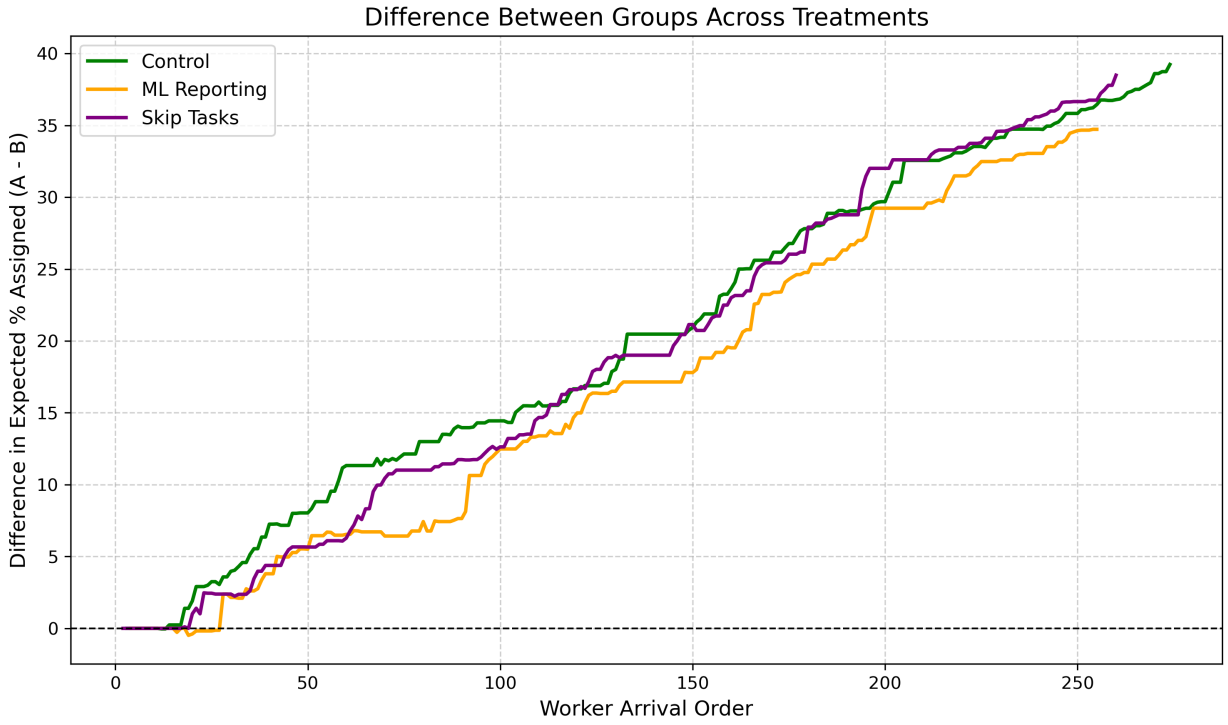


Figure 5.16. Difference in ML models' task assignments over time between Group A and Group B.

To formally test whether the magnitude of expected assignment gaps differed across treatments and changed with worker arrival order, we conducted an ANCOVA with treatment, arrival order, and their interaction as predictors (see Table 5.2). The analysis revealed a significant main effect of *treatment* ($F(2, 777) = 385.19$, $p < 0.001$), indicating that the average disparity in expected task assignments varied across conditions. Post hoc comparisons

showed that the ML Reporting treatment led to significantly smaller average group-level disparities (16.61%) compared to both the Control (20.47%, $p < 0.001$) and Skip Tasks treatments (18.74%, $p < 0.001$), suggesting that offering a mechanism for workers to report perceived unfairness reduced the overall assignment gap. The difference between the Control and Skip Tasks treatments was also statistically significant ($p < 0.001$), with the Control condition exhibiting the largest disparity.

Table 5.2. ANCOVA results comparing assignment gap growth rates across treatments.

Effect	Sum of Squares	df	F	p-value
Treatment	874.82	2	385.19	< 0.001
Arrival Order	103580.30	1	91214.78	< 0.001
Treatment × Arrival Order	137.86	2	60.70	< 0.001
Residual	882.33	777	–	–

We also found a strong main effect of *arrival order* ($F(1, 777) = 91214.78$, $p < 0.001$), indicating that assignment gaps increased significantly over time regardless of treatment. Critically, the *treatment* × *arrival order* interaction was significant ($F(2, 777) = 60.70$, $p < 0.001$), demonstrating that the pace of assignment gap increase varied across treatments. This interaction suggests that although disparities grew in all conditions, they accelerated more rapidly in the Control and Skip Tasks treatments than in the ML Reporting treatment. In other words, providing workers with a reporting mechanism seemed to somewhat reduce the overall disparity and also slowed the compounding disadvantage over time.

Having examined how disparities emerged in the ML models' theoretical assignment decisions, we now turn to the actual task assignment data to assess how these modeled biases manifested in practice. Figure 5.17 displays the evolution of group-level disparities in actual task assignments (Group A minus Group B) over time under each experimental treatment. Each line represents the average difference in the number of task assigned to Group A versus Group B, computed within sliding windows of workers sorted by their start time (i.e., a moving average of Group A minus Group B assignments). The shaded bands

show 95% bootstrapped confidence intervals, capturing uncertainty in the estimated group-level assignment gap at each time window.

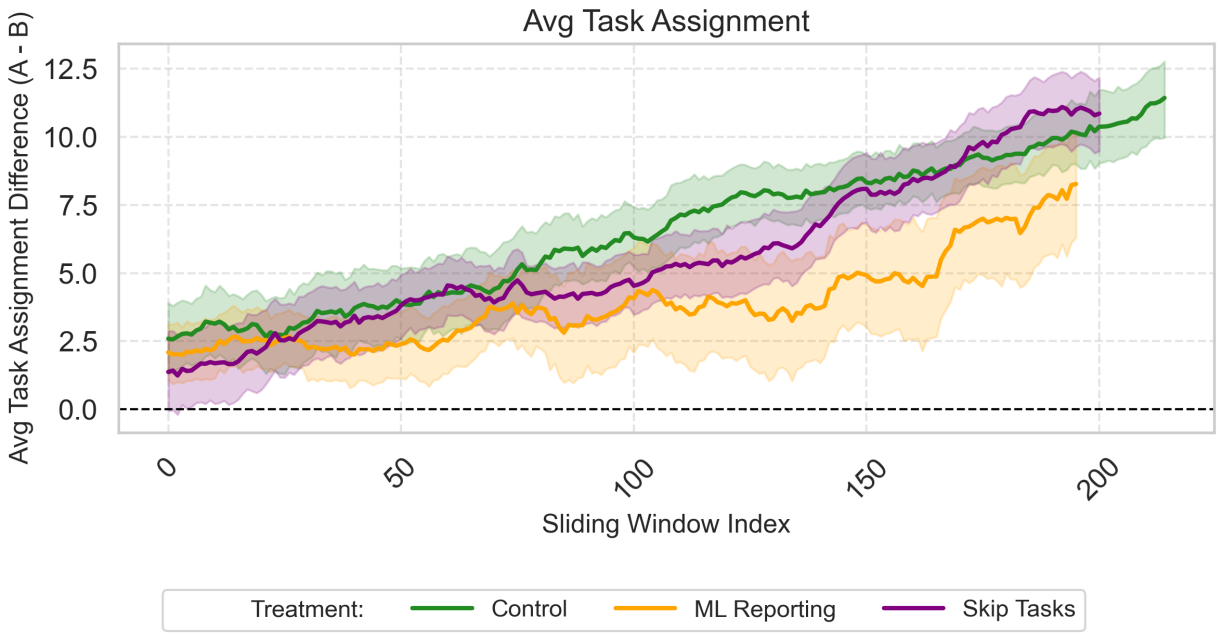


Figure 5.17. Difference in ML models' actual task assignments over time between Group A and Group B.

Similar to the expected task assignment analysis, we observe growing disparities in actual task assignments between Group A and Group B over time, with patterns varying across treatments. To ensure a proper comparison—given that each treatment had a different number of workers, which affects the length of the curves—we focus on a common range of sliding windows (approximately indices 0 to 180) where all treatments have overlapping data. Within this shared range, the *Control* and *Skip Tasks* treatments exhibit a pronounced upward trend in assignment gaps, surpassing a difference of 8 to 10 tasks towards the end. In contrast, the *ML Reporting* treatment maintains a mostly flat trajectory throughout this interval, with group-level disparities remaining closer to 4–5 tasks on average and at most 8 towards the end.

Table 5.3. Linear Mixed Effects Regression on Task Assignments

Predictor	Coef.	SE	z	p	95% CI
Main Effects					
Group B	0.223	0.606	0.37	0.713	[–0.96, 1.41]
ML Reporting	0.203	0.660	0.31	0.758	[–1.09, 1.50]
Skip Tasks	–0.744	0.522	–1.43	0.154	[–1.77, 0.28]
Day (continuous)	1.443	0.169	8.56	<0.001	[1.11, 1.77]
Two-Way Interactions					
Group B × ML Reporting	0.309	0.994	0.31	0.756	[–1.64, 2.26]
Group B × Skip Tasks	1.202	0.832	1.44	0.149	[–0.43, 2.83]
Group B × Day	–2.751	0.238	–11.58	<0.001	[–3.22, –2.29]
ML Reporting × Day	–0.112	0.276	–0.41	0.685	[–0.65, 0.43]
Skip Tasks × Day	0.207	0.247	0.84	0.402	[–0.28, 0.69]
Three-Way Interactions					
Group B × ML Reporting × Day	0.883	0.382	2.32	0.021	[0.14, 1.63]
Group B × Skip Tasks × Day	–0.138	0.341	–0.41	0.685	[–0.81, 0.53]
Intercept	20.699	0.342	60.52	<0.001	[20.03, 21.37]
Random Effects (Subject ID)	Variance = 4.050				

These temporal trends are quantitatively supported by the linear mixed effects model results shown in Table 5.3. The model reveals a strong main effect of time, with task assignments significantly increasing over days ($\beta = 1.443$, $p < .001$). Importantly, the negative interaction between *Group B* and *Day* ($\beta = -2.751$, $p < .001$) indicates that, relative to Group A, Group B workers received fewer assignments as time progressed—consistent with the visual trajectory in the control group. Furthermore, a statistically significant three-way interaction between *Group B*, *ML Reporting*, and *Day* ($\beta = 0.883$, $p = .021$) suggests that ML reporting helped slow down the compounding disadvantage for Group B, partially offsetting the growing assignment gap. Other two-way and three-way interactions are not statistically significant.

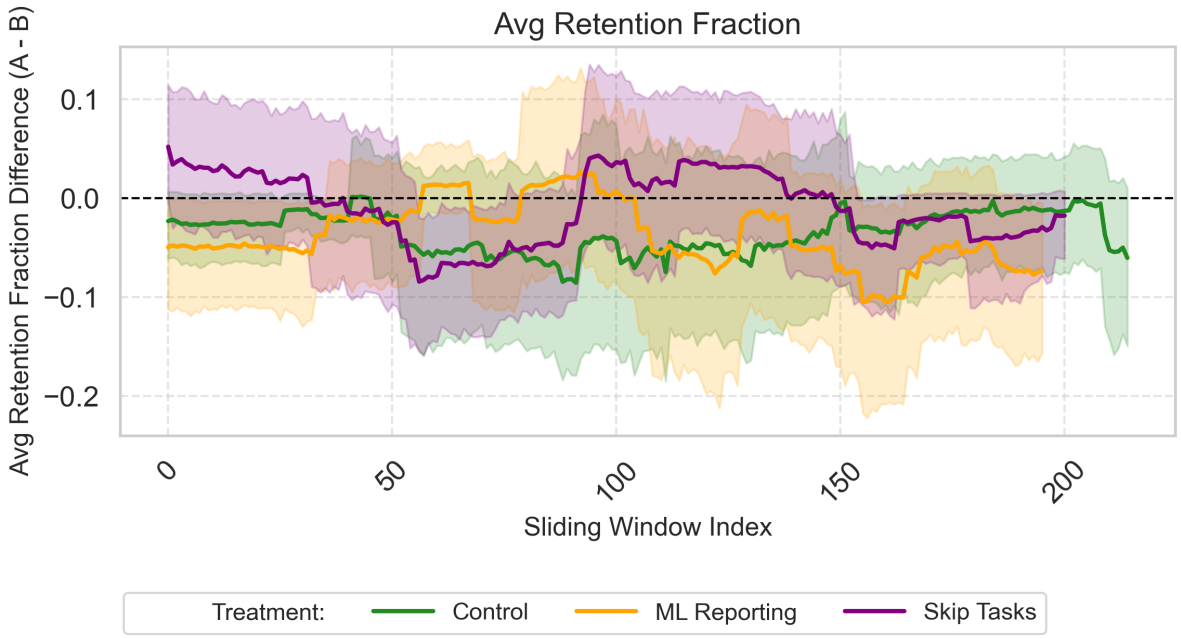


Figure 5.18. Difference in retention fraction over time between Group A and Group B, across treatments.

5.4.2 RQ2: Effects on Retention and Accuracy Across Time

In this section, we analyze how the evolving fairness properties of ML models affect workers' retention and quality behaviors over time.

Retention Fraction Gap Analysis. To more closely examine whether worker behavior in terms of task completion diverges between groups as fairness properties evolve, we look at both visualizations and regression results.

Figure 5.18 shows the evolution of retention fraction differences between Group A and Group B across the study period, visualized through a sliding window approach, and the shaded bands show 95% bootstrapped confidence intervals. Across all treatments, the retention gap remains centered around zero for much of the timeline, suggesting a relatively balanced level of engagement across groups in terms of task completion. While temporary fluctuations emerge, particularly in the *Skip Tasks* and *ML Reporting* treatments, no consistent pattern of divergence is observed over time. Notably, the Control treatment remains

stable and close to parity throughout, and although Skip Tasks appears to briefly have Group A have more retention mid-way through the study, these effects do not persist.

Table 5.4. Linear Mixed Effects Regression on Retention Fraction

Predictor	Coef.	SE	z	p	95% CI
Main Effects					
Group B	0.029	0.035	0.81	0.415	[–0.040, 0.098]
ML Reporting	0.023	0.020	1.19	0.233	[–0.015, 0.062]
Skip Tasks	–0.006	0.032	–0.19	0.851	[–0.068, 0.056]
Day (continuous)	–0.006	0.007	–0.76	0.446	[–0.020, 0.009]
Two-Way Interactions					
Group B × ML Reporting	–0.023	0.050	–0.46	0.645	[–0.121, 0.075]
Group B × Skip Tasks	–0.042	0.053	–0.79	0.431	[–0.147, 0.063]
Group B × Day	–0.001	0.013	–0.09	0.930	[–0.027, 0.025]
ML Reporting × Day	–0.009	0.010	–0.85	0.393	[–0.028, 0.011]
Skip Tasks × Day	0.006	0.013	0.45	0.652	[–0.019, 0.031]
Three-Way Interactions					
Group B × ML Reporting × Day	0.008	0.019	0.41	0.680	[–0.030, 0.046]
Group B × Skip Tasks × Day	0.009	0.020	0.48	0.635	[–0.029, 0.048]
Intercept	0.968	0.016	61.30	<0.001	[0.937, 0.999]
Random Effects (Subject ID)	Variance = 0.009				

To better understand these trends, we look at the results of the linear mixed effects regression presented in Table 5.4. None of the main effects—including *Group*, *Treatment*, or *Day*—are statistically significant, and all estimated coefficients are small in magnitude. For example, the main effect of *Day* is near zero ($\beta = -0.006$, $p = .446$), suggesting no systematic increase or decrease in retention fraction over time. Furthermore, the two-way and three-way interactions between group identity, treatment, and day are all statistically non-significant, indicating no evidence that the relationship between group and retention changed as a function of either treatment or time. These findings suggest that different treatments had a minimal impact on retention behavior throughout the study.

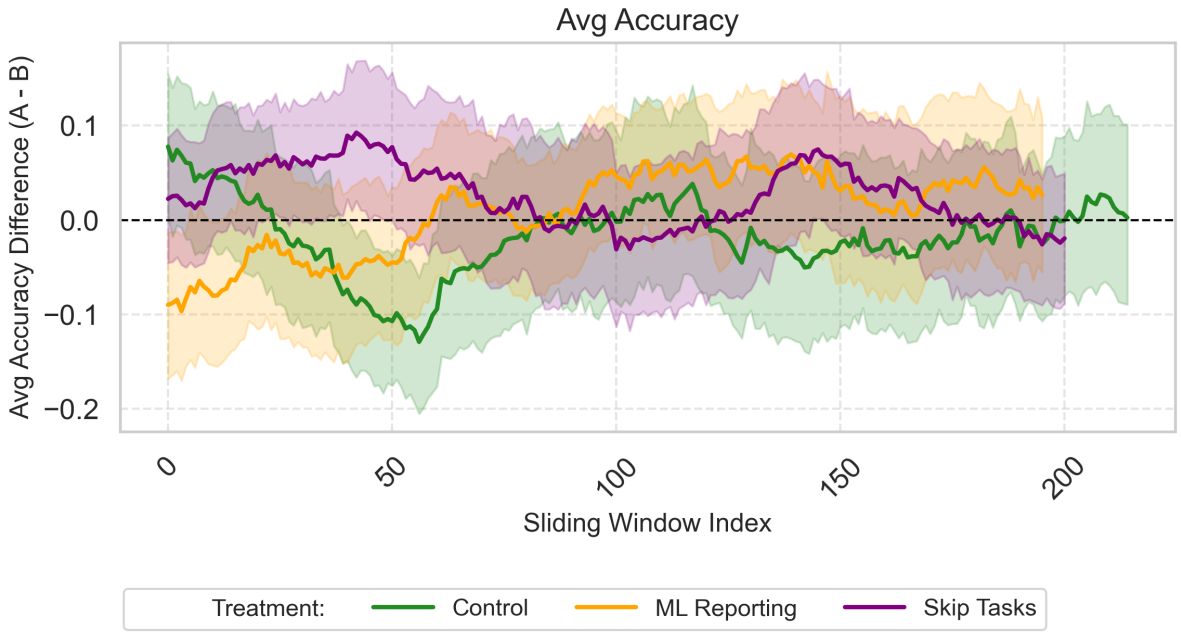


Figure 5.19. Difference in accuracy over time between Group A and Group B, across treatments.

Accuracy Gap Analysis. We then examine whether worker behavior, in terms of work quality (accuracy), diverges between groups as fairness properties evolve; accordingly, we again look at both visualizations and regression results.

Figure 5.19 illustrates the evolution of average accuracy differences between Group A and Group B across different treatments over time, visualized using a sliding window approach. The shaded bands represent 95% bootstrapped confidence intervals, capturing uncertainty in the estimated group-level differences. Although fluctuations occur within each treatment group, none of the treatments exhibit a consistent, widening disparity favoring either group. The Control treatment shows an early period where Group A performs slightly worse than Group B, followed by gradual convergence. The ML Reporting treatment shows a relatively consistent upward trend, where accuracy differences initially indicate that Group B outperformed Group A, but over time this pattern reverses, with Group A achieving higher accuracy levels than Group B. In contrast, the Skip Tasks treatment displays more fluctuation without a persistent pattern of disparity in either direction.

Table 5.5. Linear Mixed Effects Regression on Accuracy

Predictor	Coef.	SE	z	p	95% CI
Main Effects					
Group B	–0.024	0.046	–0.51	0.608	[–0.114, 0.067]
ML Reporting	–0.055	0.042	–1.31	0.190	[–0.138, 0.027]
Skip Tasks	0.030	0.047	0.65	0.517	[–0.061, 0.121]
Day (continuous)	0.011	0.011	0.94	0.347	[–0.011, 0.033]
Two-Way Interactions					
Group B × ML Reporting	0.119	0.056	2.11	0.035	[0.008, 0.230]
Group B × Skip Tasks	–0.031	0.067	–0.46	0.642	[–0.163, 0.101]
Group B × Day	0.010	0.016	0.64	0.521	[–0.021, 0.042]
ML Reporting × Day	0.025	0.015	1.64	0.102	[–0.005, 0.056]
Skip Tasks × Day	0.002	0.017	0.09	0.927	[–0.032, 0.035]
Three-Way Interactions					
Group B × ML Reporting × Day	–0.049	0.021	–2.34	0.020	[–0.090, –0.008]
Group B × Skip Tasks × Day	0.003	0.024	0.12	0.902	[–0.044, 0.050]
Intercept	0.734	0.033	22.34	<0.001	[0.670, 0.799]
Random Effects (Subject ID)	Variance = 0.013				

For more quantitative results, we look at the regression results in Table 5.5 to have a deeper understanding. The main effects of group, treatment, and day are all statistically insignificant, indicating no clear directional change in accuracy tied to these factors individually. However, we do observe a significant two-way interaction between Group B and ML Reporting ($\beta = 0.119$, $p = .035$), suggesting that during the early stage of the experiment (i.e., Day 1), Group B workers in the ML Reporting treatment exhibited relatively higher accuracy. Moreover, we also find a significant three-way interaction between Group B, ML Reporting, and Day ($\beta = -0.049$, $p = .020$) revealing that the relative accuracy for Group B under ML Reporting actually diminishes over time. These two observations suggest that as the model's task assignments became increasingly biased against Group B, their work quality declined, while Group A's work quality improved over time. No other two- or three-way interaction terms reach statistical significance. These findings suggest that while treatment and group effects on accuracy were limited, there may be nuanced patterns that can emerge through interaction effects over time.

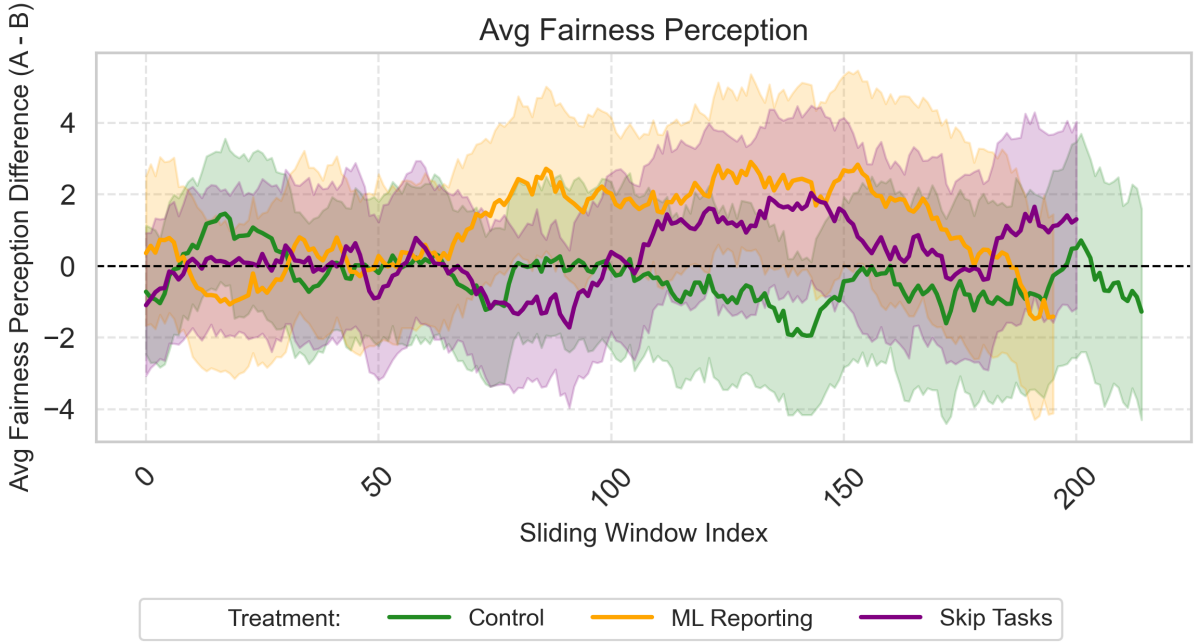


Figure 5.20. Difference in fairness perceptions over time between Group A and Group B, across treatments.

5.4.3 RQ3: Effects on Perceived Fairness

Lastly, we explore how the evolution of ML models' fairness properties—and their resulting task assignment behavior—influenced workers' fairness perceptions over time across different treatments.

Figure 5.20 visualizes the temporal trend of perceived fairness differences between Group A and Group B across the three treatments. The figure is generated using a sliding window approach, with shaded regions indicating 95% bootstrapped confidence intervals. In the *Control* treatment, the perceived fairness gap between Group A and Group B remains relatively close to zero throughout the experiment, with no consistent trend in either direction; the fluctuations fall within overlapping confidence intervals, suggesting no reliable group-level difference in fairness perceptions. In contrast, both the *ML Reporting* and *Skip Tasks* treatments exhibit an upward shift in perceived fairness for Group A relative to Group B. The *ML Reporting* treatment, in particular, displays the most pronounced and sustained high fairness perception for Group A during the middle phase of the experiment. Although

this diminishes towards the end, the overall trend suggests that favored group's overall advantage, mostly influenced them to report the model as more fair especially in these two treatments.

Table 5.6. Linear Mixed Effects Regression on Fairness Perception

Predictor	Coef.	SE	z	p	95% CI
Main Effects					
Group B	0.451	1.203	0.38	0.708	[–1.907, 2.808]
ML Reporting	1.833	1.186	1.55	0.122	[–0.492, 4.158]
Skip Tasks	–0.978	0.864	–1.13	0.258	[–2.671, 0.715]
Day (continuous)	–0.522	0.280	–1.87	0.062	[–1.071, 0.026]
Two-Way Interactions					
Group B × ML Reporting	–1.995	1.925	–1.04	0.300	[–5.768, 1.778]
Group B × Skip Tasks	0.697	1.581	0.44	0.659	[–2.403, 3.796]
Group B × Day	–0.021	0.453	–0.05	0.962	[–0.908, 0.866]
ML Reporting × Day	–0.507	0.470	–1.08	0.281	[–1.428, 0.414]
Skip Tasks × Day	0.386	0.394	0.98	0.327	[–0.386, 1.159]
Three-Way Interactions					
Group B × ML Reporting × Day	0.382	0.704	0.54	0.587	[–0.997, 1.761]
Group B × Skip Tasks × Day	–0.624	0.614	–1.02	0.309	[–1.827, 0.579]
Intercept	16.612	0.560	29.64	<0.001	[15.514, 17.710]
Random Effects (Subject ID)	Variance = 11.223				

When we turn to the regression results in Table 5.6 to better understand differences in fairness perceptions, we observe that the main effects of group, treatment, and day are all statistically insignificant. Although the coefficient for *ML Reporting* is positive ($\beta = 1.833$), suggesting higher perceived fairness under this treatment, the effect does not reach statistical significance ($p = .122$). None of the two-way interaction terms—including those between group identity and treatment or time—are statistically significant, and neither of the three-way interactions involving group, treatment, and day show reliable effects. Overall, the regression results indicate that workers' perceptions of fairness did not differ significantly across treatments, groups, or over time. While there is a positive, though non-significant, trend suggesting that the ML Reporting treatment may be associated with slightly higher perceived fairness, this effect is not strong enough to draw firm conclusions. Likewise, nei-

ther group identity nor its interaction with treatment or time meaningfully shaped fairness perceptions. These findings suggest that, despite measurable disparities in task assignment, workers' subjective perceptions of fairness remained relatively stable and were not significantly influenced by the experimental manipulations.

5.5 Conclusions and Discussions

This chapter examined how fairness in ML-driven task assignment systems dynamically evolves through repeated interactions between algorithmic decision-making and human behavior. Building on earlier chapters that focused on non-updating ML models, this chapter advanced the analysis by explicitly modeling and empirically testing the full feedback loop formed by three critical stages in the ML pipeline: decision-making, behavioral data generation, and model updating. To do so, we designed a two-part investigation combining a large-scale behavioral simulation with a randomized human-subject experiment. The simulation modeled a platform where task allocation decisions were driven by ML models that learned from workers' performance ratings—either fair or biased—while workers, in turn, responded to perceived fairness through changes in retention and work quality, which then shaped subsequent performance ratings and future model updates. In parallel, the experiment deployed a real-world platform where crowd workers interacted with live ML models under different treatments that either allowed recourse (reporting unfairness), strategic restraint (skipping tasks), or no intervention. By comparing the simulation and empirical approaches, this chapter showed that fairness in ML-driven systems can erode over time due to biased feedback and human behavior-model interactions—even when models start out fair. While the simulation revealed how fixed worker behavior models can lead to compounding disparities, the empirical study showed that real workers adapt their behavior dynamically—for instance, shifting from high to low work quality and vice-versa in response to perceived unfairness. Together, these findings indicate that mitigating unfairness in evolving systems requires interventions that not only address biased feedback but also account for the fluid and adaptive nature of human behavior.

5.5.1 Simulation Findings and Implications

The simulation study offered a controlled environment to trace how disparity emerges and compounds in ML-driven task assignment systems. Even with equal initial priors (i.e., model's initial belief of successful task completion) to ML models for both groups, we found that biased performance signals—such as skewed ratings disproportionately penalizing one group—caused task assignments to diverge sharply over time. As the model updated its beliefs of successful task completion for each worker group using this biased feedback, workers from the disadvantaged group were progressively assigned fewer tasks, while advantaged group members received more. In other words, the dynamic illustrates how early-stage bias in data collection can propagate through model retraining cycles and amplify inequality, even in systems that are initially fair by design. This echoes prior findings in algorithmic management research that highlight how small disparities in visibility, evaluation, or work assignment can escalate over time through feedback loops [131, 134, 136]. Specifically, the simulation incorporated two distinct response models—*Appreciate-Protest* and *Slack-Strive*—each modeling different behavioral reactions. In the *Appreciate-Protest* model, disparities were further reinforced by workers' reactions to perceived unfairness: disadvantaged workers reduced their retention and work quality, while advantaged workers remained engaged or improved over time. This asymmetry fed increasingly divergent signals back into the model, creating a vicious cycle of compounding disparity. In contrast, the *Slack-Strive* model exhibited a different dynamic. Disadvantaged workers who stayed in the system and gradually improved their work quality, influenced the model to assign tasks more evenly. In this setting, the feedback loop showed partial self-correcting properties—assignment disparities somewhat began to shrink, particularly when workers increased their quality and remained in the system long enough to influence the model.

These results carry several implications. First, they demonstrate that feedback loops in ML models are not only shaped by the presence of bias in performance evaluations that are fed in, but also by how workers react to perceived unfairness. This aligns with prior theoretical work showing that workers are not passive recipients of algorithmic decisions but adapt their strategies based on incentives, perceptions, and system design [137, 141]. Second,

the simulation findings signifies that fairness interventions must be tailored to the behavior-environment dynamics in which they operate. For example, mechanisms that simply increase task allocation to disadvantaged groups without first addressing biased evaluations or correcting worker behavior may backfire. Third, the emergence of partial self-correction in the *Slack-Strive* setting suggests that under certain conditions, worker persistence and behavioral improvement can help stabilize fairness outcomes—pointing to the value of designing systems that encourages long-term engagement and learning. Taken together, these findings demonstrate the importance of explicitly modeling the feedback loop between algorithmic updates and human behavior when evaluating fairness interventions. Without accounting for human reactions and their evolving dynamics, interventions that appear fair in static settings may fail—or even worsen outcomes—when deployed in real-world systems that learn from user data over time.

5.5.2 Human-subject Experiment Findings and Implications

In the human-subject experiment, we observed similar trends to that of simulation. Specifically, task assignment disparities between groups grew over time across all treatments, but were least pronounced in the ML Reporting treatment. Workers under this treatment were able to report perceived unfairness as a remediation option, which appeared to partially slow the model's group-level drift. However, this effect was limited; disparities still grew over time, echoing prior findings that transparency and recourse can help but are not always sufficient on their own [123, 126].

Workers' retention behaviors and fairness perceptions were relatively stable and did not significantly differ across treatments. However, a critical insight regarding work quality emerged from the real-world results. Basically, a key difference arised compared to the simulation results, as *behavioral flexibility* were observed among workers. Specifically, the simulation assumed fixed behavior models—such as "appreciate-protest" or "slack-strive"—which reflect idealized and consistent worker strategies. In contrast, real workers displayed more dynamic responses. For instance, in the ML Reporting treatment, Group B workers initially increased their work quality, but this trend reversed in later periods—suggesting an early

effort to stay engaged and perform well, followed by eventual discouragement as disparities accumulated. This kind of behavioral shift—unaccounted for in the simulation—illustrates the limitations of fixed behavior modeling assumptions and emphasizes the need for incorporating more nuanced, adaptive behavioral models in future simulation work. It also connects to emerging literature on worker resistance and adaptation in AM systems [137, 139, 140], which show that workers actively respond to perceived injustice, but do so in different, often non-linear ways (i.e., employing different strategies from time to time).

Moreover, our findings expand on existing work by showing that fairness interventions can affect not only individual perceptions but also the structural dynamics of model evolution. While studies have proposed design features such as recourse [146], protective optimization technologies [147], and co-designed tools [142, 143], empirical evidence on how these interventions affect fairness over time remains limited. Our experiment contributes to filling this gap by demonstrating that worker-facing fairness mechanisms—like reporting options—can influence to slow down the temporal trajectory of unfairness for sometime, even when group-level disparities persist.

5.5.3 Final Reflections

In summary, this chapter conveys that fairness in ML-based task assignment is not a static property, but an emergent one—shaped by biased feedback, human behavior, and iterative model updates. Simulations are essential for uncovering structural vulnerabilities and long-term consequences, but real-world experimentation is also crucial for revealing behavioral complexity and contextual nuance. These two perspectives, emphasize the importance of hybrid approaches in understanding and addressing fairness in algorithmic management systems. Future work should further integrate adaptive behavioral modeling and stakeholder-informed interventions to design AM systems that are not only fair at a moment in time, but resiliently fair over time.

5.6 Acknowledgments

This work is currently in preparation. I would like to thank my advisor, Professor Ming Yin, for her continued guidance and feedback throughout this project. I also acknowledge the support of the Department of Computer Science at Purdue University, as well as the crowd workers who participated in the study.

6. CONCLUSION & FUTURE WORK

Machine learning (ML) models are now ubiquitous in many fields, fitting into everyday routines and essential industrial processes. However, the human-led world is inherently marked by bias, and unfortunately, this bias can infiltrate ML models, leading them to act in ways that may not be fair or just. To better understand bias and fairness issues in ML models, adopting a human-in-the-loop perspective in the ML development pipeline is essential, as humans play various key roles in the development of ML models, appearing at its various stages.

This dissertation explores a comprehensive and multidisciplinary investigation of bias and fairness in ML systems, thoroughly exploring the multifaceted roles humans assume throughout the entire ML development and deployment lifecycle. By adopting a human-in-the-loop perspective, this dissertation highlights the intrinsic complexity of fairness, portraying it as a dynamic, emergent property resulting from continuous interactions between algorithmic systems and human behaviors. Specifically, the dissertation demonstrates how humans significantly influence ML fairness at various pipeline stages, including data annotation, deployment, and iterative model updating. It emphasizes that biases introduced by humans—whether through cognitive biases during dataset annotation or through behavioral data that reflects strategic responses to model decisions—can profoundly shape the fairness outcomes of ML models, particularly when such data is used for model training/updating. Through theoretical modeling and empirical understanding, this work emphasizes the cyclical and reciprocal nature of human and algorithmic interactions, advocating for a deep integration of human-centric considerations in the design and governance of fair ML systems.

6.1 Summary of Contributions

Starting with Chapter 2, I address the foundational stage of the ML pipeline—data annotation—by introducing an bias-aware label aggregation algorithm designed specifically to account for annotators' confirmation bias. Confirmation bias, a common cognitive bias, significantly influences annotators' decisions, leading them to disproportionately favor information aligning with their existing beliefs or values. To address this, the chapter proposes a

probabilistic graphical model that explicitly captures annotators' political stances, degrees of confirmation bias, and inherent tendencies to favor specific labels. This approach leverages an expectation-maximization algorithm to simultaneously infer annotators' parameters, along with the parameters of the tasks they interact with—including the ground-truth labels of these tasks. Empirical evaluations using real-world data annotations, particularly on political tasks such as differentiating factual statements from opinion statements in a gun control setting, demonstrate that bias-aware label aggregation outperforms traditional label aggregation techniques. Additionally, simulation analyses reveal that the proposed algorithm works best in scenarios where annotators exhibit higher levels of confirmation bias or when annotators' values are highly polarized or dispersed. In sum, these findings underscore the critical need to explicitly model cognitive biases early in the ML pipeline, as doing so can significantly improve data quality and prevent fairness issues from propagating to downstream stages of the pipeline.

After ML models are trained on data generated during the annotation stage and subsequently deployed to interact with users, Chapters 3 and 4 examine fairness not as a one-time property, but as a long-term experience shaped by users' repeated interactions with ML systems over time. Specifically, Chapter 3 examines how decision subjects perceive the fairness of ML model decisions that affect them, and consequently, how these perceptions shape their strategic reactions to continue engaging with the system over time. Through randomized human-subject experiments in a ML-based loan lending system, the chapter reveals that fairness perceptions are not solely driven by the model's group-wide fairness, but more so influenced by whether the system favors the subject's own group. For instance, subjects who observe that their group is systematically disadvantaged are more likely to leave the system. The study also demonstrates that subjects' responses are nuanced and vary by qualification level: highly qualified subjects tend to exhibit higher retention regardless of fairness disparities, suggesting that perceived utility can partially override their fairness concerns. Finally, subjects who are highly sensitive to fairness report the system as less fair and are less likely to remain engaged, but their sensitivity does not substantially moderate the influence of biased treatment on retention. In summary, Chapter 3 signifies the importance of considering long-term both individual and group-level fairness dynamics in deployed systems.

Chapter 4 directly builds on these insights by incorporating the possibility for individuals to strategically improve their qualification—a realistic feature in many ML-mediated decision-making environments such as hiring or education. Through another human-subject experiment, this chapter investigates how the availability and design of improvement opportunities interact with group identity and system fairness. The findings reveal that offering opportunities for qualification improvement increases overall subject engagement and encourages long-term participation, but fairness perceptions still remain heavily dependent on whether favorable decisions are allocated to subjects' own group. Specifically, when improvements are less accessible for disadvantaged groups, subjects tend to interpret such systems as systematically more unfair. We find that this is still the case, even when qualification improvements are easily accessible, as in the study subjects still reported the underlying bias against their own group. These results emphasize that fairness perceptions still serve as reliable indicators of a system's fairness in this context. However, from the flip side, these results also suggest that interventions—such as improvement opportunities or skill-building—though well-intentioned, may inadvertently obscure underlying disparities in usage data if ML practitioners focus solely on interaction metrics and overlook users' fairness perceptions and overall well-being. Two chapters emphasize that fairness in ML systems cannot be fully captured by a single metric or a one-shot interaction. Instead, it requires a long-term perspective—one that actively engages with users and considers how group membership, the availability and accessibility of improvement opportunities, and strategic behaviors may intersect to create complex dynamics in ML-based decision systems.

Chapter 5 deepens the analysis of fairness in ML systems by explicitly examining how human reactions to algorithmic decisions can influence subsequent model updates, creating feedback loops that shape long-term fairness outcomes. This chapter contributes a novel two-part investigation—comprising both simulations and randomized human-subject experiments—to explore how fairness evolves in ML-based task assignment systems that are periodically retrained using behavioral data. The simulation models a gig-work environment where an ML model assigns tasks to workers who may react and choose how many tasks to complete (retention) and at what quality. These completed tasks are then rated by simulated raters, whose evaluations may be biased against certain demographic groups, particularly

minorities. These ratings, whether fair or biased, feed back into the model to determine future task assignments. The simulation introduces both retention and quality as behavioral responses to perceived fairness and examines how different levels of reaction to unfair treatment influence the long-term evolution of fairness. Results show that even initially fair models can degrade over time into highly biased systems if updated with data derived from biased raters and highly reactive workers. Notably, strong behavioral reactions—such as decreased or increased task completion and work quality—can alter the volume and nature of data available for model updates. Depending on the underlying worker dynamics, these reactions can either slow the growth of disparities by limiting biased updates or, conversely, exacerbate inequality by reinforcing skewed performance signals. In some cases, particularly when disadvantaged workers remain engaged and improve their performance, the feedback loop exhibits partial self-correction. To validate and extend the simulation findings, a complementary empirical experiment was conducted in which real workers engaged with an ML-based task assignment system across multiple days. Workers were assigned to different experimental treatments, where they could skip tasks, or raise complaints. The ML models were dynamically updated based on worker performance and interactions, creating a real-world deployment scenario. This empirical study showed that—compared to the simulation assumptions—worker reactions to perceived unfairness are not static; instead, workers adapt their behavior over time—initially increasing or sustaining work quality, and later reducing it as disparities persist. The findings also suggest that interventions such as assigning more tasks to disadvantaged groups, while seemingly fair, offer only limited corrective power if the underlying evaluation of work quality fed into the model remains biased. In these cases, additional task assignments may slow the growth of disparities but are insufficient to reverse them, and may still reinforce inequities by amplifying the influence of flawed performance signals in future model updates. Overall, the combination of simulation and empirical findings demonstrates how fairness is not static but dynamically co-constructed through human-ML interactions over time. This chapter thus emphasizes the need for fairness-aware interventions that account not only for initial model conditions and metrics but also for dynamic behavioral feedback loops that emerge in deployed ML systems.

6.2 Connections Between Chapters

To reiterate, the chapters of this dissertation creates a narrative that explores bias and fairness in machine learning from a lifecycle and human-in-the-loop perspective. Together, they articulate a dynamic view of fairness—one that emerges through iterative interactions between humans and algorithmic systems and evolves as those systems are trained, deployed, and retrained using human behavioral data.

The dissertation begins at the very first stage of the ML pipeline: data annotation. Chapter 2 demonstrates that bias and fairness concerns can originate even before an ML model is trained—within the labels provided by human annotators. These labels are not neutral; rather, they may encode deep-seated cognitive biases such as confirmation bias, especially in subjective or ideologically charged domains like political discourse. The chapter introduces a novel probabilistic graphical model that quantitatively captures such biases by inferring both annotator ideologies and their annotation tendencies. The resulting bias-aware label aggregation algorithm outperforms standard approaches, especially in high-bias scenarios, underscoring the critical need to address human biases at the point of data creation. Without such interventions, the ML pipeline begins with distorted inputs, compromising model fairness from the outset.

Building on this, Chapters 3 and 4 shift the focus from data to deployment—examining how people experience, interpret, and respond to ML models that make consequential decisions about their lives when these models are trained using such annotations. These chapters examine stage 3 of the ML pipeline (Figure 1.1): the moment of ML decisions meet human reactions to them. Specifically, in Chapter 3, users strategically engage repeatedly with an ML-based loan approval system, and the experiments reveal that fairness is not a static or global construct for decision subjects. Instead, it is relational and self-referential—shaped by whether the user's own group is favored or disadvantaged. These fairness perceptions, in turn, influence user retention: disadvantaged users are more likely to leave, while advantaged users are more likely to stay and benefit, regardless of the model's overall fairness.

Chapter 4 further deepens this analysis by introducing the option for users to strategically improve their qualifications. This addition reflects a realistic and increasingly common

feature in algorithmic systems—such as reskilling in gig-work platforms or credit rebuilding in financial technologies. The chapter explores how different improvement opportunity structures interact with fairness perceptions, retention behavior, and group identity. It finds that while retention is not affected by model fairness in this setting—even when qualification improvement opportunities are available—fairness perceptions remain strongly tied to how one's group is treated. Emphasizing that usage data alone is insufficient to detect fairness disparities, this finding highlights the importance of incorporating users' well-being—particularly their fairness perceptions—when evaluating the fairness of algorithmic systems.

Chapter 5 considers the feedback loop within in the ML development pipeline, by turning attention to also including stages 4 and 5 of the ML pipeline: the transformation of human reactions into behavioral data, and the subsequent use of this data in retraining the model. Through both simulations and real human-subject experiments, this chapter investigates how seemingly fair systems can degrade into unfairness over time due to biased feedback loops. The simulation models workers and raters on a task assignment platform. Workers choose how many tasks to complete and with what quality, sometimes responding to perceived model unfairness through disengagement or deliberate underperformance, and other times through increased engagement and overperformance. Raters, in turn, introduce bias into quality ratings (assessments)—systematically overrating some and underrating others. This generates data that the ML model uses in its updates. Over time, the system can drift as initially disadvantaged groups receive fewer tasks, leading to sparser data and further marginalization in subsequent model updates. In some cases, this reduction in data slows the model's ability to reinforce disparities. In others, self-correcting feedback loops emerge—particularly when disadvantaged workers stay engaged and improve performance—partially correcting disparity over time. The human-subject experiment validates some of these insights. It reveals that remediation strategies—such as compensating disadvantaged users by assigning more tasks—may provide limited relief and ultimately fall short if the evaluation mechanisms used to update the model remain biased. The findings show the limitations of rule-based interventions and highlight the necessity of designing dynamic fairness-aware systems that are more robust to feedback loops and evolving user behavior.

To summarize, the chapters of this dissertation outline a recursive, human-centered view of the bias and fairness in the machine learning development pipeline. Chapter 2 shows how biases enter the pipeline through data generation (Stage 1). Chapters 3 and 4 illustrate how these biases manifest in lived experiences of users during deployment (Stage 3), and how users strategically respond. Chapter 5 shows how human responses feed back into the system, potentially amplifying unfairness over time through continuous biased retraining (Stages 3, 4 and 5). Thus, this dissertation completes the cycle depicted in Figure 1.1 and highlights the central thesis: fairness issues in ML models originates from human biases in data and feedback loops, which affect model updates. Addressing these issues requires understanding and modeling human behavior at each stage.

The human-in-the-loop framing across all chapters also serves to position fairness as a joint responsibility—distributed across data creators, system designers, users, and model maintainers. Fairness is not just an optimization goal to be achieved post-hoc or a metric to be minimized during training. It is a socio-technical construct, emerging through feedback loops, strategic behaviors, and continuous adaptation. This dissertation provides both a modeling framework and a behavioral understanding of how bias and fairness evolve as ML models interact with humans over time. In this sense, each chapter feeds into the next not only chronologically but conceptually. By the time of arrival at Chapter 5, it is clear that the solution to fairness cannot lie in isolated technical fixes but requires a holistic, lifecycle-oriented, and human-aware perspective. This concluding insight provides a foundation for future work in fairness-aware ML system design, calling for interventions that are anticipatory rather than reactive, systemic rather than local, and more robust to the long-term dynamics of human behavior.

6.3 Future Directions

Building upon the insights and contributions from this dissertation, several promising and intellectually rich research directions emerge—each reinforcing the broader theme that fairness in ML systems is dynamic, context-sensitive, and deeply shaped by ongoing human-model interactions. The work presented in this dissertation reveals fairness not as a fixed

metric to be optimized once, but as an evolving phenomenon shaped by a recursive cycle of design, deployment, user reaction, and model retraining. Future work must thus address fairness considering ML development pipeline with a long-term human-centered perspective, incorporating behavioral science, participatory design, and machine learning.

- **Broader Bias Modeling**: While this dissertation focused on confirmation bias in annotation, real-world decision-making and labeling processes are influenced by a broader range of cognitive biases. Future research should aim to model additional biases—such as anchoring (the tendency to rely too heavily on the first piece of information seen), framing effects (how the presentation of information alters judgment), group conformity (peer-influenced decision-making), and negativity bias (the overweighting of negative information). Extending bias-aware aggregation algorithms to these contexts can yield more robust and less biased datasets for sensitive tasks such as misinformation labeling, risk assessment, and hate speech detection.

- **Real-world Fairness Applications**: While this dissertation relies on controlled simulations and empirical studies, future work must validate fairness interventions in real-world, high-stakes domains such as healthcare diagnostics, employment screening, judicial risk assessment, and education technology. These contexts offer unique challenges—such as legal constraints, accountability demands, and direct consequences for individuals—that require fairness interventions to be not only effective but interpretable, auditable, and socially accepted. Deploying fairness-aware algorithms in such environments will also provide insight into organizational adoption barriers and long-term social impact.

- **Advanced Causal Methods**: User reactions to model fairness—ranging from protest behaviors to continued engagement and adaptive overperformance—complicate fairness evaluations by making the observed data endogenous to the system itself. As shown in Chapter 5, these reactions are not static; workers may initially respond with increased effort but later become a low performer as disparities accumulate. This behavioral variability signifies the need for future research to apply causal inference and counterfactual

modeling techniques to disentangle genuine skill or qualification signals from artifacts of human reaction to perceived algorithmic unfairness. For instance, causal tools could help determine whether lower performance among a disadvantaged group reflects true ability gaps or results from reduced effort in response to perceived unfairness. Such causal decomposition is critical for interpreting fairness metrics meaningfully and for designing interventions that address root causes rather than surface-level symptoms.

- **Participatory Design Approaches**: Fairness cannot be defined in a vacuum. Future work must expand participatory design frameworks that actively include the voices of affected communities—especially those historically marginalized by algorithmic systems. These frameworks should involve stakeholders in defining what fairness means in specific contexts, identifying relevant harms, and co-designing interventions. Incorporating these perspectives not only increases the legitimacy and acceptance of fairness solutions but also helps uncover context-specific nuances that purely technical definitions often miss. Participatory approaches will be especially vital in culturally sensitive or global deployments of ML systems.

- **Robustness Against Users' Strategic Behaviors**: Strategic behaviors in human-algorithm systems are not mere anomalies—they are emergent and often unpredictable responses to ML model's fairness structure. While the studies in this dissertation observed these behaviors in relatively consistent manner—such as increased retention when the model favors a worker's group or persistent engagement when quality improvement opportunities exist—future systems must move beyond assuming consistent user behavior. There is a growing need for formal models that can anticipate when and how users may react strategically, introducing behavioral noise that distorts model learning and feedback loops. Tools from game theory, adversarial machine learning, and mechanism design offer promising avenues to formalize and simulate these dynamics. For example, future work could identify equilibrium strategies under perceived unfairness, adversarial patterns of protest (e.g., degrading work quality or selective disengagement), or develop incentive-aligned mechanisms that promote honest, quality engagement. Designing ML systems that are robust not only to random noise

but also to adaptive, strategic human behavior is essential for preserving fairness and stability over time.

- **Longitudinal Fairness Monitoring and Intervention**: A key insight from this dissertation is that fairness can drift over time, especially in systems that learn from human behavior. Future work should focus on building infrastructure for longitudinal fairness monitoring—real-time systems that visualize fairness metrics, detect early warning signals of drift, and adaptively deploy fairness-preserving interventions. Such systems would not only enable more responsible ML practices but would also support regulatory compliance in contexts like credit scoring, hiring, and health diagnostics. Incorporating users' feedback to improve these systems, could further increase transparency and user trust.

- **Intervening in the Feedback Loop**: Finally, the simulation and empirical evidence provided in Chapter 5 reveal that interventions, such as assigning more tasks to compensate for bias, may somewhat work to fix the disparities forming in the existence of biased feedback loops. This method also carries the risk of exacerbating the disparities already formed in the system. Thus, future research must develop proper fairness-aware retraining protocols and model update mechanisms that explicitly incorporate the risk of feedback amplification. This may involve discounting biased observations, reweighting underrepresented user data, or integrating fairness constraints during retraining. Designing such interventions requires a deeper understanding of the full pipeline—where human behaviors generate data, which shapes model decisions, which influence future human behaviors.

In sum, addressing bias and fairness in ML requires both an understanding of the full development pipeline and a long-term, human-centered perspective—one that views fairness not as a one-time goal, but as an ongoing process that demands continuous monitoring, accounts for human roles and adaptability, and acknowledges the evolving influence of human-ML interactions over time. Future research should embrace this complexity by building tools and frameworks that respond not only to technical challenges but also to the lived

experiences and behavioral dynamics of the people who play different roles embedded in ML systems.

REFERENCES

[1] OpenAI, *Gpt-3.5: Language model by openai*, Accessed April 20, 2025, 2021. [Online]. Available: https://openai.com.

[2] *Machine learning in spam detection*, [Accessed 27-10-2023]. [Online]. Available: https://bdtechtalks.com/2020/11/30/machine-learning-spam-detection/.

[3] *Real-World Examples of Machine Learning (ML)*, [Accessed 27-10-2023]. [Online]. Available: https://www.tableau.com/learn/articles/machine-learning-examples.

[4] Admin. "Top 15 industrial applications for image classification - deeplobe." Accessed on 27-10-2023. (2023), [Online]. Available: https://deeplobe.ai/top-15-industrial-applications-for-image-classification/.

[5] N. Mesa, *Can the criminal justice system's artificial intelligence ever be truly fair?* May 2021. [Online]. Available: https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/.

[6] G. Smith and I. Rustagi. "When good algorithms go sexist: Why and how to advance ai gender equity." (2021), [Online]. Available: https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity.

[7] "Chatgpt and large language model bias." Accessed April 20, 2025. (Mar. 2023), [Online]. Available: https://www.cbsnews.com/news/chatgpt-large-language-model-bias-60-minutes-2023-03-05/.

[8] A. Misra, A. Gooze, K. Watkins, M. Asad, and C. A. Le Dantec, "Crowdsourcing and its application to transportation data collection and management," *Transportation Research Record*, vol. 2414, no. 1, pp. 1–8, 2014.

[9] A. Gupta and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *Journal of biomedical informatics*, vol. 108, p. 103 500, 2020.

[10] X. Zhang *et al.*, "How do fair decisions fare in long-term qualification?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 457–18 469, 2020.

[11] X. Zhang, M. Khaliligarekani, C. Tekin, *et al.*, "Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] M. A. Gemalmaz and M. Yin, "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*, ser. AIES '22, Oxford, United Kingdom: Association for Computing Machinery, 2022, pp. 295–306, ISBN: 9781450392471. DOI: 10.1145/3514094.3534201. [Online]. Available: https://doi.org/10.1145/3514094.3534201.

[13] C.-M. Chiu, T.-P. Liang, and E. Turban, "What can crowdsourcing do for decision support?" *Decision Support Systems*, vol. 65, pp. 40–49, 2014.

[14] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in neural information processing systems*, vol. 22, 2009.

[15] M. A. Gemalmaz and M. Yin, "Accounting for confirmation bias in crowdsourced label aggregation.," in *IJCAI*, 2021, pp. 1729–1735.

[16] V. Carlson, "Overestimation in size-constancy judgments," *The American journal of psychology*, vol. 73, no. 2, pp. 199–213, 1960.

[17] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," *CoRR*, vol. abs/2001.09604, 2020. arXiv: 2001.09604. [Online]. Available: https://arxiv.org/abs/2001.09604.

[18] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon, "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*, 2023, pp. 1–21.

[19] C. A. Bail *et al.*, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[20] C. Hube, B. Fetahu, and U. Gadiraju, "Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[21] D. La Barbera, K. Roitero, G. Demartini, S. Mizzaro, and D. Spina, "Crowdsourcing truthfulness: The impact of judgment scale and assessor bias," in *European Conference on Information Retrieval*, Springer, 2020, pp. 207–214.

[22] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3150–3158.

[23] L. Hu and Y. Chen, "A short-term intervention for long-term fairness in the labor market," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1389–1398.

[24] Y. Hu and L. Zhang, "Achieving long-term fairness in sequential decision making," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 9549–9557.

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

[26] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 525–534.

[27] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," *IEEE Internet Computing*, vol. 17, no. 2, pp. 76–81, 2013.

[28] J. Otterbacher, P. Barlas, S. Kleanthous, and K. Kyriakou, "How do we talk about other people? group (un) fairness in natural language image descriptions," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 106–114.

[29] A. Biswas, M. Kolczynska, S. Rantanen, and P. Rozenshtein, "The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 97–104.

[30] C. Eickhoff, "Cognitive biases in crowdsourcing," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 162–170.

[31] H. Zhuang, A. Parameswaran, D. Roth, and J. Han, "Debiasing crowdsourced batches," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1593–1602.

[32] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[33] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," *Advances in neural information processing systems*, vol. 22, pp. 2035–2043, 2009.

[34] P. Welinder, S. Branson, P. Perona, and S. Belongie, "The multidimensional wisdom of crowds," *Advances in neural information processing systems*, vol. 23, pp. 2424–2432, 2010.

[35] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," *Advances in neural information processing systems*, vol. 25, pp. 692–700, 2012.

[36] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 469–478.

[37] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" *Proc. VLDB Endow.*, vol. 10, no. 5, pp. 541–552, Jan. 2017, ISSN: 2150-8097. DOI: 10.14778/3055540.3055547. [Online]. Available: https://doi.org/10.14778/3055540.3055547.

[38] C.-J. Ho, A. Slivkins, S. Suri, and J. W. Vaughan, "Incentivizing high quality crowdwork," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 419–429.

[39] S. Doroudi, E. Kamar, E. Brunskill, and E. Horvitz, "Toward a learning science for complex crowdsourcing tasks," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 2623–2634.

[40] Y. Zheng, J. Wang, G. Li, R. Cheng, and J. Feng, "Qasca: A quality-aware task assignment system for crowdsourcing applications," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1031–1046.

[41] W. Tang, M. Yin, and C.-J. Ho, "Leveraging peer communication to enhance crowdsourcing," in *The World Wide Web Conference*, 2019, pp. 1794–1805.

[42] A. Braylan and M. Lease, "Modeling and aggregation of complex annotations via annotation distances," in *Proceedings of The Web Conference 2020*, 2020, pp. 1807–1818.

[43] Y. Li, H. Sun, and W. H. Wang, "Towards fair truth discovery from biased crowdsourced answers," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 599–607.

[44] E. Newell and D. Ruths, "How one microtask affects another," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 3155–3166.

[45] J. Huang *et al.*, "Sequential biases on subjective judgments: Evidence from face attractiveness and ringtone agreeableness judgment," *Plos one*, vol. 13, no. 6, e0198723, 2018.

[46] M. Coscia and L. Rossi, "Distortions of political bias in crowdsourced misinformation flagging," *Journal of the Royal Society Interface*, vol. 17, no. 167, p. 20 200 020, 2020.

[47] X. Duan, C.-J. Ho, and M. Yin, "Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, 2020, pp. 155–158.

[48] A. Mitchell, J. Gottfried, M. Barthel, and N. Sumida, *Distinguishing between factual and opinion statements in the news*, https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/, Accessed: 2021-05-21, Jun. 2018.

[49] D. Zhou, S. Basu, Y. Mao, and J. Platt, "Learning from the wisdom of crowds by minimax entropy," *Advances in neural information processing systems*, vol. 25, pp. 2195–2203, 2012.

[50] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 155–164.

[51] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, "Multiobjective evolutionary algorithms for the riskreturn tradeoff in bank loan management," *International Transactions in Operational Research*, vol. 9, pp. 583–597, 2002.

[52] M. Bogen and A. Rieke, "Help wanted: An examination of hiring algorithms, equity, and bias," 2018.

[53] A. Kalhan, "Immigration policing and federalism through the lens of technology, surveillance, and privacy," *Political Institutions: Federalism & Sub-National Politics eJournal*, 2013.

[54] N. Vigdor, "Apple card investigated after gender discrimination complaints," *The New York Times*, vol. 10, 2019.

[55] H. Ledford. "Millions of black people affected by racial bias in health-care algorithms." (2019), [Online]. Available: https://www.nature.com/articles/d41586-019-03228-6.

[56] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[57] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[58] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*, PMLR, 2013, pp. 325–333.

[59] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[60] A. Agarwal, A. Beygelzimer, M. Dudk, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*, PMLR, 2018, pp. 60–69.

[61] H.-F. Cheng *et al.*, "Soliciting stakeholders fairness notions in child maltreatment predictive systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, May 2021. DOI: 10.1145/3411764.3445308. [Online]. Available: https://doi.org/10.1145/3411764.3445308.

[62] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "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*, 2019, pp. 99–106.

[63] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 392–402.

[64] M. Srivastava, H. Heidari, and A. Krause, "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*, 2019, pp. 2459–2468.

[65] M. Yaghini, A. Krause, and H. Heidari, "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*, 2021, pp. 1023–1033.

[66] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[67] O. Gillath, T. Ai, M. S. Branicky, S. Keshmiri, R. B. Davison, and R. Spaulding, "Attachment and trust in artificial intelligence," *Computers in Human Behavior*, vol. 115, p. 106 607, 2021, ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2020.106607. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074756322030354X.

[68] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–52.

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

[70] K. Okamura and S. Yamada, "Adaptive trust calibration for human-ai collaboration," *Plos one*, vol. 15, no. 2, e0229132, 2020.

[71] B. Dietvorst, J. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, vol. 64, pp. 1155–1170, Mar. 2018. DOI: 10.1287/mnsc.2016.2643.

[72] S. M. Merritt, "Affective processes in human–automation interactions," *Human Factors*, vol. 53, no. 4, pp. 356–370, 2011.

[73] Z. Lu and M. Yin, "Human reliance on machine learning models when performance feedback is limited: Heuristics and risks," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[74] C.-W. Chiang and M. Yin, "Youd better stop! understanding human reliance on machine learning models under covariate shift," in *13th ACM Web Science Conference 2021*, 2021, pp. 120–129.

[75] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 295–305.

[76] X. Wang, Z. Lu, and M. Yin, "Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1697–1708.

[77] A. Rechkemmer and M. Yin, "When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14.

[78] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. arXiv: 1908.09635. [Online]. Available: http://arxiv.org/abs/1908.09635.

[79] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18, Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7, ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. [Online]. Available: https://doi.org/10.1145/3194770.3194776.

[80] S. Mitchell, E. Potash, S. Barocas, A. DAmour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual Review of Statistics and Its Application*, vol. 8, no. 1, pp. 141–163, Mar. 2021, ISSN: 2326-831X. DOI: 10.1146/annurev-statistics-042720-125902. [Online]. Available: http://dx.doi.org/10.1146/annurev-statistics-042720-125902.

[81] P. Gajane and M. Pechenizkiy, *On formalizing fairness in prediction with machine learning*, 2018. arXiv: 1710.03184 [cs.LG].

[82] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, *Fairness through awareness*, 2011. arXiv: 1104.3913 [cs.CC].

[83] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[84] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human Decisions and Machine Predictions*," *The Quarterly Journal of Economics*, vol. 133, no. 1, pp. 237–293, Aug. 2017, ISSN: 0033-5533. DOI: 10.1093/qje/qjx032. eprint: https://academic.oup.com/qje/article-pdf/133/1/237/30636517/qjx032.pdf. [Online]. Available: https://doi.org/10.1093/qje/qjx032.

[85] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of ml fairness notions," *arXiv preprint arXiv:2006.16745*, 2020.

[86] J. Hannan, H.-Y. W. Chen, and K. Joseph, "Who gets what, according to whom? an analysis of fairness perceptions in service allocation," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 555–565.

[87] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions," in *Proceedings of the 2018 Chi conference on human factors in computing systems*, 2018, pp. 1–14.

[88] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19, Marina del Ray, California: Association for Computing Machinery, 2019, pp. 275–285, ISBN: 9781450362726. DOI: 10.1145/3301275.3302310. [Online]. Available: https://doi.org/10.1145/3301275.3302310.

[89] J. Schoeffer, Y. Machowski, and N. Kuehl, "A study on fairness and trust perceptions in automated decision making," *arXiv preprint arXiv:2103.04757*, 2021.

[90] N. Van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov, "Effect of information presentation on fairness perceptions of machine learning predictors," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[91] N. Grgi-Hlaa, A. Weller, and E. M. Redmiles, "Dimensions of diversity in human perceptions of algorithmic fairness," *arXiv preprint arXiv:2005.00808*, 2020.

[92] M. Kasinidou, S. Kleanthous, P. Barlas, and J. Otterbacher, "I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 690–700.

[93] L. T. Liu, A. Wilson, N. Haghtalab, A. T. Kalai, C. Borgs, and J. Chayes, "The disparate equilibria of algorithmic decision making when individuals invest rationally," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 381–391.

[94] P. Nokhiz, A. K. Ruwanpathirana, N. Patwari, and S. Venkatasubramanian, "Precarity: Modeling the long term effects of compounded decisions on individual instability," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 199–208.

[95] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine bias. propublica, may 23, 2016*, 2016.

[96] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, PMLR, 2018, pp. 77–91.

[97] H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong, "Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–22, 2020.

[98] C. Kam, "Risk attitudes and political participation," *American Journal of Political Science*, vol. 56, Oct. 2012. DOI: 10.1111/j.1540-5907.2012.00605.x.

[99] M. A. Gemalmaz and M. Yin, "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*, 2022, pp. 295–306.

[100] J. Dastin, *Amazon scraps secret ai recruiting tool that showed bias against women*, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[101] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 275–285.

[102] N. A. Saxena, "Perceptions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19, Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 537–538, ISBN: 9781450363242. DOI: 10.1145/3306618.3314314. [Online]. Available: https://doi.org/10.1145/3306618.3314314.

[103] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of machine learning fairness notions," *SIGKDD Explor. Newsl.*, vol. 23, no. 1, pp. 14–23, May 2021, ISSN: 1931-0145. DOI: 10.1145/3468507.3468511. [Online]. Available: https://doi-org.ezproxy.lib.purdue.edu/10.1145/3468507.3468511.

[104] H.-F. Cheng *et al.*, "Soliciting stakeholders fairness notions in child maltreatment predictive systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[105] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on fairness, accountability and transparency*, PMLR, 2018, pp. 149–159.

[106] M. K. Lee and K. Rich, "Who is included in human perceptions of ai?: Trust and perceived fairness around healthcare ai and cultural mistrust," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.

[107] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A qualitative exploration of perceptions of algorithmic fairness," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.

[108] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "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*, ser. AIES '19, Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 99–106, ISBN: 9781450363242. DOI: 10.1145/3306618.3314248. [Online]. Available: https://doi.org/10.1145/3306618.3314248.

[109] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions," ser. CHI '18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14, ISBN: 9781450356206. DOI: 10.1145/3173574.3173951. [Online]. Available: https://doi.org/10.1145/3173574.3173951.

[110] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," *CoRR*, vol. abs/1901.07694, 2019. arXiv: 1901.07694. [Online]. Available: http://arxiv.org/abs/1901.07694.

[111] J. Schoeffer, Y. Machowski, and N. Kuehl, *A study on fairness and trust perceptions in automated decision making*, 2021. DOI: 10.48550/ARXIV.2103.04757. [Online]. Available: https://arxiv.org/abs/2103.04757.

[112] N. van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov, "Effect of information presentation on fairness perceptions of machine learning predictors," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966. DOI: 10.1145/3411764.3445365. [Online]. Available: https://doi.org/10.1145/3411764.3445365.

[113] K. Kieslich, B. Keller, and C. Starke, "Artificial intelligence ethics by design. evaluating public perception on the importance of ethical design principles of artificial intelligence," *Big Data & Society*, vol. 9, no. 1, p. 20 539 517 221 092 956, 2022.

[114] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 903–912.

[115] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 10–19.

[116] R. N. Spreng, M. C. McKinnon, R. A. Mar, and B. Levine, "The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures," *Journal of personality assessment*, vol. 91, no. 1, pp. 62–71, 2009.

[117] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, 2016. arXiv: 1607.06520 [cs.CL].

[118] A. Nadeem, B. Abedin, and O. Marjanovic, "Gender bias in ai: A review of contributing factors and mitigating strategies," 2020.

[119] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science facultys subtle gender biases favor male students," *Proceedings of the national academy of sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.

[120] K. M. Elsesser, "Gender bias against female leaders: A review," *Handbook on well-being of working women*, pp. 161–173, 2016.

[121] C. Lee, "Gender bias in the courtroom: Combating implicit bias against women trial attorneys and litigators," *Cardozo JL & Gender*, vol. 22, p. 229, 2015.

[122] M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Computers in Human Behavior*, vol. 70, pp. 1–13, 2017.

[123] J. Meijerink and T. Bondarouk, "Fairness in algorithmic management: How practices promote fairness and redress unfairness on digital labor platforms," *Human Resource Management Journal*, 2021.

[124] K. C. Kellogg, M. Valentine, and S. Suri, "Algorithmic management in the gig economy: A systematic review and research integration," *Journal of Organizational Behavior*, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/job.2831.

[125] K. C. Kellogg, M. A. Valentine, and A. Christin, "Perceived organizational support in the face of algorithmic management: A conceptual model," *Academy of Management Discoveries*, 2020.

[126] C. Rzepka and E. S. Berger, "Fairness in algorithmic management: Bringing platform-workers into the fold," *Journal of Business Ethics*, 2022.

[127] A. Setiawan and D. S. Suhendi, "Workers' affective commitment in the gig economy: The role of is quality, organizational support, and fairness," *International Journal of Information Management*, 2023.

[128] B. Rogers, "Discriminating tastes: Uber's customer ratings as vehicles for workplace discrimination," *Berkeley Journal of Employment and Labor Law*, 2017.

[129] A. Tassinari and V. Maccarrone, "Algorithmic paranoia: Gig workers' affective experience of abusive algorithmic management," *New Technology, Work and Employment*, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/ntwe.12317.

[130] H. Zhang, L. Chen, Y. Sun, and T. Sun, "The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments," *Management Science*, vol. 69, no. 4, pp. 2093–2117, 2023. DOI: 10.1287/mnsc.2022.4485.

[131] E. Bokányi and A. Hannák, "Ride-share matching algorithms generate income inequality," *arXiv preprint arXiv:1905.12535*, 2019. [Online]. Available: https://arxiv.org/abs/1905.12535.

[132] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr," in *Proceedings of CSCW*, 2017.

[133] T. Sun, "Does the gig economy discriminate against women? evidence from physicians in china," *Labour Economics*, 2022.

[134] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr," *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pp. 1914–1933, 2017. DOI: 10.1145/2998181.2998327.

[135] M. L. Gray and S. Suri, "Discrimination in the gig economy: The experiences of black online english teachers," *International Journal of Communication*, 2022.

[136] T. Sun, "Does the gig economy discriminate against women? evidence from physicians in china," *Labour Economics*, 2022.

[137] Z. Zhang, Y. Huang, B. Li, and S. Pan, "The coin of ai has two sides: Matching enhancement and information revelation effects of ai on gig-economy platforms," in *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2023.

[138] N. Seaver, "The transparency-resistance paradox in algorithmic management," *Computers in Human Behavior*, 2024, Forthcoming. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0747563224002711.

[139] A. Rosenblat and L. Stark, "Algorithmic management and the politics of demand: Control and resistance at uber," *Big Data & Society*, 2016.

[140] G. Newlands, "Algorithmic management and worker autonomy: Resistance and support in the ride-hailing sector," *New Technology, Work and Employment*, 2021.

[141] D. Schneider and K. Harknett, "Algorithmic management and the politics of control: Gig workers strategic responses," *Socius: Sociological Research for a Dynamic World*, 2020.

[142] M. Y. Liu, A. Xu, S. Guha, and M. Dontcheva, "A bottom-up end-user intelligent assistant approach to empower gig workers against ai inequality," in *CHIWORK '23: Symposium on Human-Computer Interaction for Work*, 2023.

[143] A. Zhang, A. Boltz, J. Lynn, C.-W. Wang, and M. K. Lee, "Stakeholder-centered ai design: Co-designing worker tools with gig workers through data probes," in *arXiv preprint arXiv:2303.03367*, 2023. [Online]. Available: https://arxiv.org/abs/2303.03367.

[144] C. Toxtli, A. Richmond-Fuller, and S. Savage, "Reputation agent: Prompting fair reviews in gig markets," *arXiv preprint arXiv:2005.06022*, 2020. [Online]. Available: https://arxiv.org/abs/2005.06022.

[145] K. C. Kellogg, M. A. Valentine, and A. Christin, "Algorithmic management in the workplace," *Academy of Management Annals*, vol. 14, no. 1, pp. 366–410, 2020.

[146] S. Joshi and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 26 385–26 397.

[147] B. Kulynych, R. Overdorf, C. Troncoso, and S. Gürses, "Protective optimization technologies," *arXiv preprint arXiv:1806.02711*, 2018.

[148] Algorithmic Justice League, *Crash: Community reporting of algorithmic system harms*, https://www.ajl.org/crash-project, Accessed: 2025-04-27, 2022.

A. THE EFFECT OF QUALIFICATION IMPROVEMENT ON DECISION SUBJECTS' REPEATED INTERACTIONS WITH ML MODELS APPENDIX

This appendix provides the demographic data for and survey questions used in all experiments presented in Chapter 4.

A.1 Participant Demographics

Table A.1. 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.

%	Study 1	Study 2 (Easy to hard)	Study 2 (Hard to easy)	Study 3
Gender				
Male	54%	51%	57%	55%
Female	45%	48%	42%	45%
Other	1%	1%	1%	0%
Age				
18–24	17%	12%	15%	15%
25–34	29%	26%	27%	26%
35–44	24%	26%	25%	25%
45+	30%	36%	33%	34%
Race				
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%

A.2 Survey Questions

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) Bachelors degree

  (i) Masters 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 ML 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 ML 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 ML 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 ML 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
  (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 ML Fairness Questions**

- **The bank's ML system is fair. (Positive)**
  (a) Strongly disagree
  (b) Disagree
  (c) Neutral
  (d) Agree
  (e) Strongly agree
- **The bank's ML system is fair to loan applicants. (Positive)**
  (a) Strongly disagree
  (b) Disagree
  (c) Neutral
  (d) Agree

(e) Strongly agree

- **The bank's ML 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 ML system will be fair. (Positive)**
  (a) Strongly disagree
  (b) Disagree
  (c) Neutral
  (d) Agree
  (e) Strongly agree
- **The bank's ML 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 ML system will make mistakes. (Negative)**
  (a) Strongly disagree
  (b) Disagree
  (c) Neutral
  (d) Agree
  (e) Strongly agree