FAIR SCREENING WITH RESUME CORRESPONDENCE EXPERIMENTS

AN HONORS THESIS

SUBMITTED ON THE 13th DAY OF MAY, 2022

TO THE DEPARTMENTS OF ECONOMICS AND COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE HONORS PROGRAM

OF NEWCOMB-TULANE COLLEGE TULANE UNIVERSITY

FOR THE DEGREE OF

BACHELOR OF ARTS WITH HONORS IN ECONOMICS

AND BACHELOR OF SCIENCE WITH HONORS IN COMPUTER SCIENCE

BY

Batu El

APPROVED:

Nicholas Mattei
Co-Director of Thesis

Patrick Button
Co-Director of Thesis

Michael Joyce
Third Reader
Algorithms are used in almost every step of the hiring process, assisting hiring decisions, if not directly making them. While the use of algorithms can help us eliminate the effects of the discriminatory tendencies of humans on hiring decisions, it can also amplify the effect of discriminatory hiring practices. Previous studies on algorithmic hiring focused on mitigating discrimination against race and gender, leaving a gap in the literature regarding the methods to address age discrimination in hiring algorithms. Combining the data generation techniques used by resume correspondence studies from the economics of discrimination literature and a particular definition of fairness, counterfactual fairness presented by Kusner et al. (2018), from the computer science literature, I propose an algorithm that is fair with respect to age: I show that counterfactuals in resume correspondence experiments can be used to decompose the effects of age and experience in hiring decisions; moreover, a relaxation of counterfactual fairness can be satisfied by a simple adjustment to the parameters of the machine learning model. The correction in model parameters is designed to address age discrimination, yet this method is of interest to researchers who study other settings where the sensitive attributes, such as age, gender, or race, of the individuals causally affect non-sensitive characteristics relevant to decision-making.
Preface

The genesis of my research proposal was a simple observation: economics and computer science disciplines independently investigated discrimination/fairness, creating possibilities for original research on interesting unexplored questions at the intersection of the two. My thesis is an attempt to integrate the methodologies in economics and computer science literatures to design a causally interpretable and fair screening algorithm.

At the end of my junior year in college, I was extremely fortunate to have the background knowledge to start an interdisciplinary research project on fairness in machine learning: I had worked for Professor Patrick Button as a research assistant for almost two years, studying the economics of discrimination, learned machine learning theory and practices through my classes and independent studies with Professor Jihun Hamm, and reviewed the literature on algorithmic fairness for my research project with Professor Nicholas Mattei. Working on combinatorial problems related to algebraic geometry with Professor Michael Joyce, I was impressed by his professionalism and sincerity, which motivated me to ask him to sit on my committee. In addition to my thesis committee, Professor Aron Culotta advised me on this project throughout my senior year. I have an immense appreciation for the excellent support that I received from my exceptional mentors and the numerous academic opportunities I had as an undergraduate at Tulane.

I am thankful to many peers who have contributed to the development of my ideas in this thesis and provided comments on my drafts. They include Arman Yagci, Can Yesildere, Ege Cavusoglu, David Graber, Jonathan Ogawa, Justin Phillips, and Mark Xiao. I gratefully acknowledge the financial support from the Gordon family whose generous research grant funded my work on this project during the summer of 2021.
## Contents

1 Introduction

1.1 Problem Statement .................................. 1
1.2 Literature Review .................................. 1
1.3 Methodology ........................................ 3
1.4 Contribution ........................................ 4
1.5 Organization ........................................ 4

2 Algorithms in Hiring

2.1 Sourcing ............................................. 6
2.2 Screening ............................................ 7
2.3 Interview ............................................ 8
2.4 Selection ............................................ 9
2.5 Fair Pipeline ........................................ 9

3 Age Discrimination .................................. 11

3.1 Importance ........................................... 11

4 Formal Setting ........................................ 13

4.1 General Setting .................................... 14
4.2 True Y and Prediction .............................. 15
4.3 Score and Predicted Callback .................... 17

5 Survey of Fairness Criteria .......................... 18

5.1 Fairness Criteria in Law ........................... 18
5.2 Discrimination Models in Economics .......... 19
5.3 Fairness in Machine Learning .................... 20
  5.3.1 Fairness through Unawareness ............... 20
  5.3.2 Observational Fairness Criteria ............. 21
## List of Figures

1. The Hiring Pipeline (Rieke, 2017) ................................................. 6
2. Data Generating Process for Callback .......................................... 13
3. Data Generating Process with Generic Variables ....................... 15
4. Data Generating Processes for $Y^*$, $Y$, and $\hat{Y}$ .................. 16
5. Another Plausible Data Generating Process ................................. 17
6. Predictions of a Naive Model Trained on Data with Discrimination . 17
7. Conditional Distribution of Experience Levels Given Age Groups .... 40
8. Model Predictions Before Correction .......................................... 42
9. Model Predictions After Correction ........................................... 43
10. Comparison of Predictions Before and After Correction ............ 43
11. Causal Relationships ............................................................... 47
List of Tables

1  Legally Recognized Protected Classes  . . . . . . . . . . . . . . . . . . .  14
2  Domains Regulated for Fairness  . . . . . . . . . . . . . . . . . . . . . .  14
3  Dataset from Resume Correspondence Experiment by Neumark et al. (2019)  27
4  Applicant Types based on Age and Experience Interactions  . . . . . . . . . .  29
1 Introduction

1.1 Problem Statement

We often don’t have full control over decisions that define our life prospects, such as college admissions and hiring decisions. Inaccurate or unfair decisions made by others, humans or machines, can undermine our ability to access these opportunities. The influence of algorithms on our society is expanding as many important decisions are now automated, and hiring decisions are not an exception. Algorithms are now assisting hiring decisions if they are not directly making them (Miller, 2015a; Rieke, 2017). On the one hand, the use of hiring algorithms has the potential to amplify discriminatory hiring practices (Miller, 2015b; Rieke, 2017; Schumann et al., 2020; Li et al., 2021; Yarger et al., 2020). On the other hand, their usage can help us mitigate the effects of the discriminatory tendencies of humans by establishing an industry-standard in fair hiring (Miller, 2015a; Raghavan et al., 2019). I believe it is essential to address fairness concerns in hiring with foresight, actively trying to prevent discrimination in the decision-making process. My thesis adopts this view by proposing an assistive tool for actively addressing age discrimination in hiring.

1.2 Literature Review

Historically, predilections and biases of humans who make hiring decisions have resulted in unfair hiring practices, undermining the life prospects of certain demographic groups. Unfairness in hiring practices is documented by a large body of work from the economics of discrimination literature, primarily by audit studies, which use trained applicants to test hiring discrimination, and resume correspondence (RC) experiments, which create resumes for fictitious applicants (Fix and Struyk, 1993; Neumark et al., 2019; Bendick et al., 1996, 1999; Riach and Rich, 2006, 2010; Lahey, 2008; Heckman, 1998; Bertrand and Mullainathan, 2004; Quillian et al., 2017; WILSON et al., 1999; Autor, 2009). Initially, algorithms were introduced in the hiring pipeline to automate the decision-making,
hoping that they could offer a correction to this broken process (Miller, 2015a). Today, they are used in almost every step of the hiring process, from identifying potential candidates to determining the terms of their offer (Rieke, 2017). However, experiments with algorithmic hiring showed that the machines do not always make fair decisions, even though they lack predilections and prejudices (Miller, 2015b; Dastin, 2018).

Previous works on algorithmic fairness investigated the channels through which biases can infiltrate decision criteria and proposed methods to address these biases (Barocas et al., 2019). A possible channel is when a machine learning model is trained on a dataset where the outcome variable is contaminated by discrimination: In this case, the model learns a decision criterion that includes the biases in the training set (the data that machine learning model learns from) and replicates the human prejudices in its predictions. Addressing the fairness concerns requires opting for a mathematical fairness criterion from a range of definitions available in the literature. Discrimination is a domain-specific issue; hence the appropriate fairness criterion depends on the sensitive attribute and the use cases of the algorithm (Barocas et al., 2019). Therefore, it is essential to carefully consider different domain and protected class combinations to design fair algorithms.

The literature on algorithmic fairness paid insufficient attention to age discrimination. Ajunwa (2019) and Stypinska (2021) pointed out that age discrimination is overlooked in algorithmic fairness literature. Ajunwa (2019) discussed how Age Discrimination in Employment Act applies to the discrimination concerns in the advertisement, automated hiring, and recruitment platforms, and Stypinska (2021) argued that “The aging population has been largely neglected during the turn to digitality and AI” (Stypinska, 2021; Ajunwa, 2019). Specifically, previous studies on algorithmic hiring focused on mitigating or eliminating discrimination against race and gender, leaving a gap in the literature regarding the methods to address age discrimination in hiring algorithms (Raghavan et al., 2019; Schumann et al., 2020; Li et al., 2021; Yarger et al., 2020). While Alford et al.; Brandao (2019) and Howard et al. (2017) investigated age discrimination in other settings, how to address
ageism in automated hiring still stands as an open question.

1.3 Methodology

My thesis investigates age discrimination in automated hiring and provides a possible answer to this open question. I focus on the screening stage of the hiring pipeline, where employers have received the applicants’ resumes and are choosing which applicants they will interview. I treat screening as a supervised learning problem, assuming that we can access a dataset containing resumes for a set of applicants and associated decisions to offer an interview or reject the applicant by employers. However, the decisions by employers are contaminated by ageism. The machine learning model aims to learn the decision criteria of the employers while eliminating the existing ageism from its decision criteria.

Studying age discrimination in hiring algorithms is especially tricky because of the causal relationship between the work experiences of applicants, which is a factor that employers use in making hiring decisions, and the applicants’ age, which is illegal to use in making hiring decisions. Identifying appropriate fairness criteria for age discrimination is particularly challenging because of this relationship: Age can be a barrier that reduces the applicant’s chances of getting an interview, while it can also serve as a factor that helps the applicant as older applicants often have more work experiences. I survey the different fairness criteria from the computer science literature to identify one that is well suited for addressing age discrimination in hiring.

Combining the data generation technique from Neumark et al. (2019), a resume correspondence study that quantifies hiring discrimination against older workers, and a particular definition of fairness, counterfactual fairness presented by Kusner et al. (2018), I propose an algorithm that is fair with respect to age. I use the relaxation of the definition of counterfactual fairness and present a new approach to achieve fairness, as the previous methods relied on distributional assumptions, limiting their practicality (Kilbertus et al., 2018; Nabi and Shpitser, 2018; Kusner et al., 2018). I show that counterfactuals in resume correspon-
dence experiments can be used to decompose the effects of age and experience in hiring decisions, and relaxation of counterfactual fairness can be satisfied by a parameter adjustment on the model. Employing the dataset from Neumark et al. (2019), I demonstrate that this methodology effectively uses the work experiences of applicants in predictions without discriminating against age groups despite the existence of a strong causal relationship between the ages and experience levels of the applicants.

1.4 Contribution

The achievements of this thesis are threefold. Firstly, I discuss how the different fairness criteria in economics, law, and computer science literature relate to each other. The purpose of this is to make the thesis accessible to a wide range of audiences across disciplines in mathematical and computational sciences as well as social sciences. I then explain how the definitions in literature can be applied to age discrimination. Secondly, I survey the fairness criteria in the computer science literature and propose a plausible definition of fairness to address age discrimination in algorithmic hiring. Thirdly, I present an algorithm that is fair with respect to age and discuss its possible use cases. While the correction in model parameters is designed to address age discrimination in screening, researchers who study other settings where the sensitive attribute of the individuals causally affects non-sensitive characteristics of the applicant can improve and modify my approach to be applied in other settings.

1.5 Organization

In Sections 2 and 3, I provide the general background that motivates my study. Sections 4 and 5 define the terminology and explain the causal relations that are needed to understand the ideas in this thesis. I discuss the dataset in Section 6. Then, in Section 7, firstly, I describe the connection between counterfactual fairness from the computer science literature and the ways that economists have interpreted discrimination (7.1), then I propose an algo-
algorithm that achieves a relaxation of counterfactual fairness (7.4). In Section 8, I demonstrate through experiments that the methodology that I propose factors the work experiences of applicants in decision making while avoiding the usage of their age. Finally, I discuss the limitations and practicality of this methodology in Section 9.
2 Algorithms in Hiring

Hiring is not a single decision point but a series of assessments and decisions that collectively result in a rejection or a job offer for an applicant (Rieke, 2017). Figure 1 demonstrates the four main steps in the hiring pipeline discussed by Rieke (2017): sourcing, screening, interview, and selection.

2.1 Sourcing

A job posting is often the first interaction of a candidate with an employer. At the sourcing step of the hiring pipeline, the methods and tools used by the employers and platforms influence the demographics of the people who submit an application to a job opening. There are two main avenues through which biases can jeopardize the fairness of the sourcing step: The first avenue governs who gets to see the job posting and the second concerns who are encouraged to apply by the posting.

Firstly, the pool of candidates who see the posting on a platform is often determined by a recommendations system. Recommender systems match the applicant pool and the employer pool on a platform. Using collaborative filtering (predictions based on preferences of similar users) and content-based filtering (predictions based on user’s activity), these algorithms usually provide applicants and employers with a ranked list of most likely matches (Rieke, 2017; Ekstrand et al., 2011; van Meteren, 2000). Yao and Huang (2017) notes that collaborative filtering algorithms are sensitive to discrimination that exists in the historical data (Yao and Huang, 2017). In addition to the biases that exists in the historical data, Farnadi et al. (2018) discussed observation bias which recommender systems “due to a feedback loop which causes the model to learn to only predict recommendations similar to previous ones” (Farnadi et al., 2018).
Secondly, the tone of the job description in the posting can play an important role in determining who chooses to apply Rieke (2017); Gaucher et al. (2011); Collier and Zhang (2016). Gaucher et al. (2011) proposed that gendered wording, i.e., the usage of masculine-themed and feminine-themed words, in the job recruitment materials can act as an institutional mechanism that reinforces and perpetuates existing gender inequalities (Gaucher et al., 2011). While their study focused on gender discrimination, the wording of the posting can also act as a factor that maintains other group-based inequalities. Similarly, Collier and Zhang (2016) discussed the beneficial use of gender-fair language in job postings (Collier and Zhang, 2016).

Moreover, Rieke (2017) noted that employers could use targeted advertisements to announce a job opening to the subset of the potential applicants, which can introduce a bias in the process. Additionally, biases can also arise from the platform that employers use to advertise the job opening if the users of the platform are a demographically imbalanced subset of the pool of potential candidates.

### 2.2 Screening

Screening is the first assessment in the hiring pipeline after employers receive the applications. When there are excess applications for a job opening, in the screening step of the hiring pipeline, automated methods are used to restrict the pool of applicants to a number that the employer can afford to interview. After a resume is received and reviewed by the employers, the applicants receive a response, which can be a direct rejection or an invitation for the assessment or interview in the next step of the hiring pipeline. Following the convention in the economics of discrimination literature, I will refer to the decision made by the employer at this stage as *callback* throughout the thesis.

**Definition 2.1 (Callback)** An invitation to return for a second audition or interview (oxf, 2022).
Rieke (2017) discusses the use of assessments, or “games,” that employers send to the candidates to inform the screening process. Evaluating the fairness of those assessments requires data on how the responses from demographic groups to the questions on the assessments vary. If a demographic group tends to give a particular answer to questions on the assessment or engage in a particular way with the game, which is associated with lower scores, then it is possible that the assessments will have biased outcomes. Wilson et al. (2021) outlined a framework for algorithmic auditing, using pymetrics, a startup that uses such assessments to recommend job candidates to their clients, as a case study (Wilson et al., 2021). On the other hand, Rieke (2017) argues that pymetrics and similar assessments “exemplify some of the most fundamental concerns about predictive technology used in hiring,” because the criteria they use to differentiate low and high performers reflect subjective evaluations, hence their methods can mirror undesirable social patterns (Rieke, 2017). A major concern is that the traits inferred in these assessments are not necessarily causally related to the job performance of the candidates.

In this thesis, I focus on the screening algorithms that take the information the candidate reports on their resume as input and predict callback based on the skills and work experiences of the applicants, which are assumed to be causally related to the job performance. Models that take the information on how the candidates engage with the assessment “games” are outside the scope of my investigation.

2.3 Interview

After screening, the next step in the hiring pipeline is the interview. The interview stage is vulnerable to the explicit and implicit biases of the hiring managers who conduct the interviews. There have been efforts to automate and assist the interview process by requesting structured video interviews from the applicants. For instance, HireVue scores pre-recorded video interviews and grades the applicants’ responses (Rieke, 2017). The algorithms that extract signals such as facial expressions and speech from videos are likely to have bi-
ases due to the discrepancies in their accuracy for different skin tones and different accents (Cavazos et al., 2020; Feng et al., 2021).

2.4 Selection

Finally, while making the final decisions in the selection stage, employers do background checks and try to optimize profitable terms for the job offer. Rieke (2017) notes that algorithms used to predict the likelihood that a candidate has a tendency to violate workplace policies based on the background checks can be a potential source of bias at this stage.

2.5 Fair Pipeline

Ensuring the fairness of even a single step in the hiring pipeline is challenging; however, it is not sufficient to make the hiring process fair. Furthermore, Dwork and Ilvento (2018) showed that satisfying fairness criteria for individual steps in a pipeline also does not guarantee that the fairness criteria will be satisfied for the pipeline, highlighting that “classifiers that are fair in isolation do not necessarily compose into fair systems and also that seemingly unfair components may be carefully combined to construct fair systems” (Dwork and Ilvento, 2018).

Designing a fair hiring pipeline is a challenging task that requires careful study of fairness in each step of the pipeline and the appropriate methods to combine them. This thesis aims to contribute to the efforts of the many researchers who worked on this topic. My investigation is limited to the screening stage of the hiring pipeline, a major step where the candidate pool is significantly reduced. My model learns from a dataset that contains the information from the resumes of the applicants and the associated decisions by the employers. The decision is either to offer an interview (callback = 1) or to reject the applicant (callback = 0).

While I propose an algorithm that can be used to make automated hiring decisions, I highlight that this algorithm should be perceived as an assistive tool that notifies the
decision-makers, humans or other algorithms, when their decision seems to be discrimina-
tory, i.e., the decisions of the decision-maker and the fair predictions of the model presented
by this thesis are in conflict.


3 Age Discrimination

3.1 Importance

Age discrimination manifests in hiring as treating an applicant less favorably because of their age. Addressing age discrimination against older workers in hiring is critical for three major reasons.

Firstly, age discrimination against older workers is illegal. The Age Discrimination in Employment Act (ADEA) protects the rights of workers who are 40 years old or older. While it is illegal for an employer to favor a younger worker over an older one, the law does not prohibit favoring older applicants over younger ones (EEOC). There is no federal policy that covers workers under the age of 40, although some states have laws that protect younger applicants from age discrimination (EEOC). The law prohibits age discrimination not only in hiring but also in “firing, pay, job assignments, promotions, layoff, training, benefits, and any other term or condition of employment” (EEOC). This thesis focuses on age discrimination in hiring.

Secondly, undermining the employment prospects of older workers, age discrimination encourages them to leave the labor force at earlier stages of their lives. In many countries, population aging is leading to rising dependency ratios, i.e., the ratio of people who benefit from the social security programs (children and older people) to the workforce that provides income for those programs (working-age population) (Land and Lamb, 2008). This creates difficulties in financing social security programs. Older individuals’ participation in the labor force is, therefore, essential to maintaining a balanced budget for public programs. However, age discrimination forces senior workers to leave the labor force at an earlier age, adding to the financial distress in covering the costs of social security programs due to population aging and rising dependency ratios.

Finally, age discrimination cases have significant legal costs: Neumark et al. (2019) estimates that the costs of potential and actual hiring cases under the Age Discrimination
in Employment Act are around $3.29 billion per year. Hence, addressing age discrimi-
nation concerns before the discriminatory tendencies of the decision-makers influence the 
outcome can help cut a significant amount of legal costs.

Addressing age discrimination in the context of algorithmic fairness is especially tricky 
because of the strong causal relationship between age and the years of experience of an 
aplicant (Neumark et al., 2019). When the former is used in the decision-making process, 
it leads to discrimination, whereas the latter is a relevant factor that informs the employers 
about an applicant’s productivity and fitness for the job. Therefore, my goal is to design an 
algorithm that can use the information on the work experiences of an applicant as a factor 
in decision making while not using their age.
4 Formal Setting

The directed acyclic graph (DAG) in Figure 2 demonstrates a possible data-generating process of callback decisions for a set of candidates who applied to a given job opening at company Z. We assume that the composition of the pool of applicants for the position is representative of the population. For this example, assume that the only sensitive attribute of an applicant is their age; hence all the observations are from a pool of applicants with a fixed gender, race, ethnicity, etc. The red arrows on the DAG show the casual effects of age on other variables, including callback. Skills and experiences of an applicant constitute the observed non-sensitive attributes, and they are shown as two separate nodes in the DAG due to the difference in their causal relationships with age: We assume that the work experiences of an applicant are causally dependent on their age, while their skills are not causally dependent on the applicant’s age. The observed non-sensitive attributes are presented in the resumes of the applicants, and they are informative about an applicant’s productivity and fitness for the job. Therefore, it is desirable to use observed non-sensitive attributes as features when predicting callback. Unobserved skills on the DAG refers to the unobserved attributes of an applicant, which can potentially be informative about the applicant’s productivity and fitness for the job. Callback decisions are a function of all other attributes of the applicant.
<table>
<thead>
<tr>
<th>Protected Class</th>
<th>Legal Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td>Color</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td>Sex</td>
<td>Equal Pay Act of 1963</td>
</tr>
<tr>
<td>Religion</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td>National origin</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td>Citizenship</td>
<td>Immigration Reform and Control Act</td>
</tr>
<tr>
<td>Age</td>
<td>Age Discrimination in Employment Act of 1967</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>Pregnancy Discrimination Act</td>
</tr>
<tr>
<td>Familial status</td>
<td>Civil Rights Act of 1968</td>
</tr>
<tr>
<td>Disability status</td>
<td>Rehabilitation Act of 1973</td>
</tr>
<tr>
<td>Veteran status</td>
<td>Vietnam Era Veterans’ Readjustment Assistance Act of 1974</td>
</tr>
<tr>
<td>Genetic information</td>
<td>Genetic Information Nondiscrimination Act</td>
</tr>
</tbody>
</table>

Table 1: Legally Recognized Protected Classes

<table>
<thead>
<tr>
<th>Regulated Domain</th>
<th>Legal Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit</td>
<td>Equal Credit Opportunity Act</td>
</tr>
<tr>
<td>Education</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td></td>
<td>Education Amendments of 1972</td>
</tr>
<tr>
<td>Employment</td>
<td>Civil Rights Act of 1964</td>
</tr>
<tr>
<td>Housing</td>
<td>Fair Housing Act</td>
</tr>
<tr>
<td>Public Accommodation</td>
<td>Civil Rights Act of 1964</td>
</tr>
</tbody>
</table>

Table 2: Domains Regulated for Fairness

4.1 General Setting

While this thesis focuses on addressing age discrimination, our methodology is applicable to other settings where similar causal relationships between the attributes of the individuals and the outcome variable exist. Therefore, I present a generalization of the causal model.

In Figure 3, we can think of the $A$ node as a matrix whose columns are the sensitive attributes that indicate membership in different demographic groups, such as a gender, a race, and ethnicity, a nationality, or other protected classes described in Table 1. Table 2 provides a list of the domains regulated for fairness, which constitute the possible use cases of similar algorithms. We assume that sensitive attributes are observed.
We use $X$ to denote the observed non-sensitive attributes of an individual that influence the outcome $Y$. The columns of the $X$ matrix have two subsets: $X_i$, which stands for $X_{\text{independent}}$ and are the features that are not causally influenced by the sensitive attributes $A$, and $X_d$, or $X_{\text{dependent}}$, which denotes the non-sensitive attributes that are causally dependent on $A$ through $\alpha_x$. $X_i$ influences the outcome $Y$ through $\beta_i$, and the effect of $X_d$ on $Y$ is captured by $\beta_d$. The columns of the $U$ matrix are non-sensitive attributes of the applicant that are not observed but are influential in determining $Y$. Their effect on $Y$ is through $\beta_u$. Some of the unobserved non-sensitive unobserved attributes, i.e., some of the columns of $U$, can be causally dependent on $A$, and $\alpha_u$ captures that relationship. The direct effect of the sensitive characteristic on the outcome variable $Y$ is $\alpha_y$. When sensitive characteristics are directly used in decision making, we will have a nonzero $\alpha_y$.

### 4.2 True $Y$ and Prediction

Figure 4 introduces two new variables: $Y^*$ and $\hat{Y}$. Note that the DAG in the middle is the same as 3 and demonstrates the data generating process for $Y$, which is an outcome variable that is contaminated by discrimination. The variable $Y^*$, or true $Y$, is a variable that truly captures whether the candidate is a “good hire” for our problem. It is a function of observed and unobserved attributes of the applicant and is not directly affected by
A. Ideally, we want to train our model on $Y^*$ rather than $Y$; however, $Y^*$ is often not observable. For the screening problem, a variable that is a good proxy for $Y^*$ can be the job performance scores of the previous hires. However, there are two major problems with using performance scores. Firstly, we can not be sure that they are not contaminated by workplace discrimination or unfair examinations. Secondly, using the performance scores of the applicants requires using the pool of current employees of a company $Z$ to train our model. However, the current composition may not be representative of the demographics of the population or may not include people from certain age groups. In general, $Y$ is an estimate of $Y^*$ but is contaminated by discrimination. In hiring, usually $Y \neq Y^*$ because:

1. The callback decisions are influenced by implicit biases of the hiring managers, which is captured by $\alpha_y$ in Figure 4.
2. Employers inaccurately use $A$ as a proxy for $U$
3. Hiring decisions are not always accurate as potential “good hires” can be rejected due to the inaccuracies in decisions by hiring managers. Hence, even in the absence of items 1 and 2, $Y$ is a noisy estimate of $Y^*$.

I aim for the predictions of our model, $\hat{Y}$, to be an accurate estimate of $Y^*$; however, if we train a model on $Y$ ignoring the fairness concerns, the decision criteria of our model will include discrimination and will estimate $Y$ instead of $Y^*$. We can not use the unobserved variables, $U$, in predictions; therefore, $\hat{Y}$ will be a function of $X_i, X_d$, and $A$. 

Figure 4: Data Generating Processes for $Y^*$, $Y$, and $\hat{Y}$
One of the limitations of the causal models in Figure 3 and Figure 4 is the assumption that $U$ is not influenced by $X_i$ or $X_d$. In the real world, it is possible that an unobserved non-sensitive attribute is causally dependent on $A$, as well as $X$. Therefore, the causal relationships can, in fact, be similar to those in Figure 5.

4.3 Score and Predicted Callback

Figure 6 shows an example output $\hat{Y}$ from a logistic regression trained on a dataset with $Y$ as the target variable. The $x$-axis shows the score that our model assigns to applicants. The score of the applicants $R$ is a continuous output variable that the model assigns to each applicant. For the purposes of this example, $t = 0.6$ is arbitrarily chosen as a threshold for the callback. In general, $t \in [0, 1]$. The model suggests providing positive callback (callback = 1) to the applicants whose scores lie above the threshold, i.e. whenever $R > t$, and reject (callback = 0) the applicant if $R \leq t$. This process defines a binary output variable, $C \in \{0, 1\}$, which I call predicted callback. $R$ and $C$ are two different encodings of $\hat{Y}$. 
5 Survey of Fairness Criteria

The definition of discrimination and interpretation of fairness has changed significantly throughout history. Different groups in the population faced morally unjustified treatment for centuries. The lessons from history and the present motivated different disciplines to come up with their interpretations of the dynamics that drive discrimination, informing scientists’ intuition when formulating effective methods to combat systemic unfairness. In this section, I provide a survey of the definitions of discrimination and fairness from the legal sphere, economics, and computer science literature.

5.1 Fairness Criteria in Law

There are two different definitions of discrimination in the legal system under the Civil Rights Act: “Disparate Treatment” and “Disparate Impact”.

**Definition 5.1 (Disparate Treatment)** Treatment of an individual in a regulated domain that is less favorable than treatment of others because of individual’s membership to protected classes. (*Merriam-Webster*)

Disparate treatment is an interpretation of discrimination based on a narrow definition of equality that aims to ensure the decision-making process treats similar people, who belong to different demographic groups, similarly. Disparate treatment occurs when the decision-maker explicitly considers protected class membership as a criterion for determining the outcome for an individual. In hiring, we say there is disparate treatment when “members of a minority are treated differently (less favorably) than members of a majority group with identical productive characteristics” (*Autor, 2009*). This includes “intentional” attempts to discriminate against a class without direct reference to their protected attributes, although implicit biases can also result in disparate treatment of individuals. However, preventing disparate treatment is not sufficient to achieve distributive justice and minimize inequality of outcomes. For that, the decision-maker also needs to avoid disparate impacts.
Definition 5.2 (Disparate Impact) “A test or other tool used for selection that, though appearing neutral, actually has an adverse effect on a particular protected class of individuals” (cor).

Disparate impact concerns the outcomes where selection rates for groups are not balanced. In particular, the \( \frac{4}{3} \) rule states if the selection rate for a certain group is less than 80% of the rate of the group with the highest selection rate, there is disparate impact (Que). This definition is based on a broader notion of equality that is concerned with reorganizing society in a way that also promotes equality in the long run. It requires fair treatment of people whose dissimilarities resulted from systemic injustices in the past.

In Figure 3, the effect of \( A \) on outcome trough \( \alpha_y \) captures disparate impact since it is the direct effect of \( A \) on \( Y \). Moreover, the effect of \( A \) on \( Y \) through the path \( \alpha_x \beta_d \) is another example because that channel also requires employers to make callback decisions only based on their observations of \( A \), although theoretically, they are using information from \( A \) as a proxy for the unobserved characteristics that they do not observe. On the other hand, the effect of \( A \) on \( Y \) through \( \alpha_x \beta_d \) is a potential source of disparate impact.

5.2 Discrimination Models in Economics

The notions of fairness discussed in economics literature fall under two main categories: taste-based discrimination models and statistical models of discrimination.

Definition 5.3 (Taste-based Discrimination) In the economics models of taste-based discrimination, hiring applicants that belong to a certain demographic group diminishes the utility of employers (Button, 2021; Becker, 1971).

Under taste-based discrimination, members of the protected class have to compensate by obtaining superior skills to get the same chances of getting hired. By definition, taste-based discrimination results in disparate treatment of the members of the protected class. In Figure 3, the channel \( \alpha_y \) corresponds to taste-based discrimination.
The second class of fairness models in the economics literature capture the other channel for disparate treatment: \( \alpha_i \ast \beta_i \). While taste-based discrimination models assume that discrimination is not a consequence of economic decision-making but rather a phenomenon that arises from the tastes of the employers, statistical discrimination models interpret discrimination as a signal extraction problem.

**Definition 5.4 (Statistical Discrimination)** Statistical discrimination occurs when employers have imperfect information about the productivity-related characteristics of an applicant, so they estimate the applicant’s unobserved productivity-related characteristics (U) based on the applicant’s membership in protected classes, which is then used to make hiring decisions (Phelps, 1972; Arrow, 1971).

In doing so, they use the difference between the means of the productivity-related characteristics across different demographic groups, as well as the difference between variances (Autor, 2009; Button, 2021).

### 5.3 Fairness in Machine Learning

In this section, I will provide the definitions of several fairness criteria grouped under four categories: Fairness through Unawareness, Group-based Fairness Criteria, Individual Fairness Criteria, and Counterfactual Fairness.

#### 5.3.1 Fairness through Unawareness

Ruf and Detyniecki (2020) notes that “[many of the] current legal standards demand to remove sensitive attributes from data in order to achieve *fairness through unawareness*”.

**Definition 5.5 (Fairness through Unawareness(FTU))** A predictor is said to achieve fairness through unawareness if protected attributes are not explicitly used in the prediction process (Gajane, 2017).
FTU is an unsophisticated suggestion, as it ignores possible causal relationships between the protected class and other input variables. If the causal structure of the model is as in Figure 3, then ignoring $A$ will have a confounding effect which will lead to a biased estimate of the model parameters. Economists can recognize this as the textbook example of the omitted variable bias. A detailed explanation is provided in Appendix A.

5.3.2 Observational Fairness Criteria

Observational fairness criteria evaluate the model predictions on a sample from the population to assess the fairness of a given model. Many different observational fairness criteria are defined in the literature (Gajane, 2017). Barocas et al. (2019) point out that most of them can be grouped under three categories: independence ($\hat{Y} \perp A$), separation ($\hat{Y} \perp A|Y^*$), and sufficiency ($Y^* \perp A|\hat{Y}$).

Definition 5.6 (Independence) The random variables $(A, \hat{Y})$ satisfy independence if

$$\hat{Y} \perp A$$

Independence is equivalently defined by other researchers as demographic parity, statistical parity, group fairness (Hardt et al., 2016; Jiang et al., 2021; Kusner and Loftus, 2020; Srivastava et al., 2019; Corbett-Davies and Goel, 2018; Besse et al., 2021; Zemel et al., 2013; Gajane, 2017). The goal of establishing independence is to eliminate (in the case of strict independence) disparate impact or to mitigate it (using a soft condition that is incorporated in the loss function). Independence imposes that we hire an equal proportion of the candidates from each protected class $A$. Hence, it results in predictions such that

$$Pr(C = 1|A = 1) = Pr(C = 1|A = 0)$$

In binary classification problems, this characterization is equivalent to the definition of
independence. Separation and its derivatives constitute the second main group of observational fairness criteria.

**Definition 5.7 (Separation)** *The random variables $(\hat{Y},A,Y^*)$ satisfy separation if*

\[(\hat{Y} \perp A|Y^*)\]

Separation requires both the false-positive rates and the false-negative rates to be equal between demographic groups (Barocas et al., 2019). In other words, for a binary classification problem, it requires

\[Pr(C = 1|A = 1,Y^* = 1) = Pr(C = 1|A = 0,Y^* = 1)\]  \hspace{1cm} \text{Equalized True Positive Rates}

\[Pr(C = 1|A = 1,Y^* = 0) = Pr(C = 1|A = 0,Y^* = 0)\]  \hspace{1cm} \text{Equalized False Positive Rates}

The two conditions together are also called “equalized odds” in other works (Goel et al., 2018; Hardt et al., 2016; Gölz et al., 2019). The first condition that forces the true positive rates to be equal for the demographic groups is also referred to as “Equality of Opportunity” because it equates the rates at which the qualified candidates from demographic groups are accepted (Hardt et al., 2016).

**Definition 5.8 ( Sufficiency)** *The random variables $(\hat{Y},A,Y^*)$ satisfy sufficiency if $Y^* \perp A|\hat{Y}$ (Barocas et al., 2019).*

In the case of binary classification, this requires

\[P(Y^* = 1|\hat{Y} = r,A = 1) = P(Y^* = 1|\hat{Y} = r,A = 0)\]

When the protected attribute $A$ also takes only two distinct values, the definition is charac-
terized by the following conditions:

\[ P(Y = 1|C = 1, A = 1) = P(Y = 1|C = 1, A = 0) \quad \text{Equalized True Positive Rates} \]
\[ P(Y = 1|C = 0, A = 1) = P(Y = 1|C = 0, A = 0) \quad \text{Equalized False Negative Rates} \]

### 5.3.3 Individual Fairness Criteria

One of the early works that pointed out that awareness of the protected class is necessary for fairness was Dwork et al. (2011). Their formulation of individual fairness requires that if two applicants are similar to each other, based on some metric \( d(\cdot, \cdot) \), then the predictions of the model for the two applicants should also be similar in order to satisfy individual fairness Dwork et al. (2011); Kusner et al. (2018); Gajane (2017).

**Definition 5.9 (Individual Fairness (IF))** Given a distance metric \( d(\cdot, \cdot) \) that captures the similarity of the applicants, if individuals \( i \) and \( j \) are similar, then predictions for individuals \( i \) and \( j \) should also be similar, i.e.

\[ d(i, j) < \varepsilon \implies \hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)}) \]

for some small \( \varepsilon \).

### 5.3.4 Counterfactual Fairness

The definition of counterfactual fairness is based on the idea that a decision is fair towards an individual if it is the same in the actual world and in the counterfactual world where the individual belongs to a different demographic group (Kusner et al., 2018). The counterfactual statement “the value of \( Y \) if \( A \) had taken the value \( a \)”, for two observable variables \( Y \) and \( A \) correspond to the question how would the callback values change if the applicant was old. Mathematically, this counterfactual statement is denoted as \( Y_{A \leftarrow a} \) or \( Y_a \), indicating that we make an intervention on \( A \) to modify its value.
Counterfactual fairness compares the scores of two applicants in our threshold model (see 6): The first applicant, applicant $i$, can be any applicant from the dataset. The second applicant, applicant $j$, is a hypothetical applicant that we create by changing the protected class membership of the applicant $i$ with an intervention, which is equivalent to applying the $do()$ operator from Pearl and Mackenzie (2018); Pearl (2009). After the intervention on $A$, we let all the characteristics of the applicant that are causally dependent on age change due to the intervention to obtain applicant $j$. A model is said to be counterfactually fair if the scores that it assigns to applicant $i$ and applicant $j$ are similar for any applicant $i$ that we can choose from the dataset.

**Definition 5.10 (Counterfactual Fairness)** A model’s prediction, $\hat{Y}$, is counterfactually fair if under any context where $X = x$ and $A = a$,

$$P(\hat{Y}_{A \leftarrow a} = y|X = x, A = a) = P(\hat{Y}_{A \leftarrow a'} = y|X = x, A = a)$$

for all values of $y$ and for all values of $a$.

$\hat{Y}_{A \leftarrow a}$ refers to the value of $A$ in a world where we set the value of $A$ equal to some $a$ with an intervention, i.e., $do$ operator from Pearl and Mackenzie (2018); Pearl (2009). Kusner et al. (2018) highlights that counterfactual fairness captures the idea of an agent, external to the system, modifying the value of $A^{(i)}$ by forcefully assigning it to be some $a'$, “for example as in a randomized experiment” Kusner et al. (2018). The conditioning on $A = a$ means that before the intervention, we had $A = a$. We change $A$ with an intervention while keeping all the values that do not causally depend on $A$ constant. We let the other attributes that are causally dependent on $A$ change due to the intervention we did on $A$. Hence, we know that $X_i$ will stay constant for the applicants before and after the intervention, whereas $X_d$ will change due to the change in $A$ (see Figure 3).
5.4 Critique of Fairness Criteria

It is important to note that none of these criteria are perfect: Disparate treatment is a minimum threshold to classify decision-making as illegal, while it is not sufficient to achieve distributive justice and minimize inequality of outcome. Disparate impact, which aims to address those concerns, only declares decision-making illegal if the acceptance rate for a certain group is less than $\frac{4}{3}$ of the group with the highest acceptance rate. The definitions I provided from the literature on the economics of discrimination seem to be especially narrow, as they only focus on addressing disparate treatment. Furthermore, earlier works on statistical discrimination argued that it could, in fact, increase efficiency, presenting an unpleasant trade-off to the society (Schwab, 1986; Norman, 2003; Phelps, 1972; Arrow, 1971). Fairness through unawareness also does not provide a solution for settings where the input variables are correlated with $A$, as the parameter estimates, and therefore the model’s predictions would be confounded by the membership to a protected class. Observational fairness criteria have received significant attention in the computer science literature; however, separation, sufficiency, and their derivatives often require knowing the true value of the output variable, $Y^*$, which is not feasible in many settings. Independence provides an idealistic approach, yet it can be an overkill to enforce independence with respect to age on the test set (a sample of candidates that is representative of the pool of potential applicants). Moreover, it is not possible to satisfy the three observational fairness criteria in non-degenerate data sets; hence, collectively, they do not provide a conclusive answer to the fairness question. Counterfactual fairness seems to be a reasonable proposition; however, applying it to real-world situations requires assuming the causal model. Before moving on to the discussion of the appropriate fairness criteria for addressing age discrimination in screening algorithms, in the next section, I will discuss the dataset used in this study.
6 Dataset: A Resume Correspondence Experiment

The focus of the debate on age discrimination policy is older workers who struggle to find new jobs (Neumark et al., 2019). Previous studies on age discrimination in economics literature have compared the difference in the callbacks for old and young applicants that are “matched,” meaning that they are similar in all attributes, including their work experiences, except their age (Bendick et al., 1996, 1999; Riach and Rich, 2006, 2010; Lahey, 2008). Neumark et al. (2019) highlighted that this standard methodology for resume correspondence experiments that assigned older and younger applicants similar labor market experiences could potentially bias the estimates of age discrimination because of two main reasons:

1. The absence of relevant experience proportional to applicant’s age may be a negative signal to the employers.
2. On real-world resumes, applicants’ experiences are often proportional with their ages.

Hence, Neumark et al. (2019) argued “for both policy and legal reasons, the right comparison for measuring age discrimination is between younger applicants and older applicants who have experience commensurate with their age,” and conducted a resume correspondence experiment where they assigned ranges of experience levels to applicants from different age groups. I employ their dataset that is unique in its generation of counterfactuals regarding the interactions between age and experience levels and contains over 40,000 resumes of fictitious applicants and the responses that the applicants received from different job openings. The variables in dataset from Neumark et al. (2019) are described in Table 3.

According to Neumark et al. (2019), two objectives guided their resume generation process: (i) making the resumes realistic to support the external validity of their results, and (ii) having data for valid comparisons of older and younger applicants. Neumark et al. (2019) applied only to low-skilled jobs as high-skilled jobs often require specific work experiences, which are difficult to control for in an experiment. The four low-skilled occu-
<table>
<thead>
<tr>
<th>Generic Variables</th>
<th>Variables in the Dataset</th>
<th>Description</th>
</tr>
</thead>
</table>
| $Y$               | callback                 | 1 if positive callback from employer  
0 otherwise |
| $A$               | young                     | applicant is between the ages 29 and 31  
middle-aged       | applicant is between the ages 49 and 51  
senior(old)       | applicant is between the ages 64 and 66 |
| $X_i$             | spanish                   | applicant speaks spanish  
computer          | applicant has computer skills  
liscense           | applicant has driver’s license  
certificate       | applicant has certificate  
customerservice   | applicant has customer service experience  
cpr                | applicant knows cardiopulmonary resuscitation  
technskills       | applicant has tech skills  
wpm                | word per minute included on resume  
grammar            | grammatical errors on the resume  
college            | applicant graduated from college  
employeemonth     | applicant was chosen employee of the month  
volunteer          | applicant has volunteer experience  
emp                | applicant is currently employed |
| $X_d$             | low experience            | applicant doesn’t have significant work experience  
low experience     | applicant doesn’t have significant work experience  
high experience    | applicant has significant work experience  
early bridge       | applicant started working at a bridge job early in their career  
late bridge        | applicant started working at a bridge job late in their career |
| Cities            | city1-city12             | 12-big cities in the US where the resumes are sent |
| Occupations       | occupation               | janitor, administrative assistant, retail sales, or security guard |

Table 3: Dataset from Resume Correspondence Experiment by Neumark et al. (2019)
occupations targeted by the study are (i) retail sales, including retail salespersons and cashiers, (ii) administrative assistant, including secretaries and administrative assistants, receptionists and information clerks, office clerks and file clerks, (iii) janitors, including janitors and building cleaners, and (iv) security guards, including security guards and gaming surveillance officers. I will focus on only one of these occupations, administrative assistants, in my experiments. In the dataset, all administrative assistants are female, and Neumark et al. (2019) used Caucasian sounding names on the resumes to control for the potential effects of signaled race on callback. Hence, the only sensitive attribute, $A$, present in the dataset is assumed to be the applicants’ ages. The experiment used three age groups ($A$): young (29–31), middle (49–51), and old (64–66). Each fictitious applicant was first randomly assigned to an age category with equal probability.

### 6.1 Skills ($X_i$)

Neumark et al. (2019) assigned skills to the applicants independent of their age. Half of the applicants from each age category were chosen to be high-skilled, meaning that they will have five (randomly chosen) of the seven skills discussed below. The first five are general skills that were used for candidates for all occupations, and the last two are specific skills for administrative assistants. The other half of the applicants are low-skilled, meaning that their resumes do not contain any of the seven skills.

1. **College Degree:** bachelor of arts for sales, administrative assistant, and security guards; associate of arts for janitors
2. **Spanish:** fluency in Spanish as a second language
3. **Employee of the month:** an “employee of the month” award on the most recent job
4. **Volunteer:** Volunteering experience (food bank, homeless shelter, or animal shelter)
5. **Grammar:** an absence of typographical errors
6. **Word per Minute:** typing 45, 50, or 55 words per minute listed on the resume
7. **Computer Skills:** facility with relevant computer software, a random mix of Quick-
books, Microsoft Office, and inventory management software

When assigning colleges and universities for the high-skilled resumes, Neumark et al. (2019) avoided the top-tier universities, choosing randomly from the local schools, colleges, and universities that have been operating since 1960. Artificial candidates were then randomly assigned to current employment status, unemployed or employed, with equal probability, independent of their age. While it is illegal to discriminate against applicants based on their employment status, for my experimentation, I treat it as a part of $X_i$ because it is not feasible to remove it from the dataset. I highlight that ideally it should not be included as a part of $X_i$ as it is a sensitive attribute.

### 6.2 Experience Levels ($X_d$)

Neumark et al. (2019) noted that plausible experience levels for each age group were different. As shown in Table 4, young workers can only have low experiences (Y) due to the number of years between their school-leaving age and the time of the application. In addition to high and low experience levels, resumes with “bridge jobs”, which convey rising skill levels in their early career followed by a transition to lower-skill jobs, were assigned to middle-aged and old applicants. Middle-aged applicants can only display a transition to a bridge job early in their career, whereas senior applicants can transition to a bridge job either early (OBE) or at a later stage (OBL) in their career. When we decompose the dataset based on the experience levels and age groups of the applicants, we get eight applicant “types”, as demonstrated in Table 4.
6.3 Callback (Y)

Neumark et al. (2019) sent these resumes to jobs in 12 different cities in the US, and the responses from the employers were collected. There are three types of coded responses from the firms. First, a typical unambiguous positive response indicates the candidate has passed the screening process by offering to set up an interview; second, an ambiguous response asks for additional information from the applicants; and third, an unambiguous negative response is rejection. Following Neumark et al. (2019), I treated the ambiguous responses as positive callbacks since they convey a willingness to move forward with the applicant in the hiring process. Hence, each resume in the dataset is associated with an output variable, callback, that takes the value 1 or 0.
7 Methodology

7.1 Appropriate Fairness Criteria for Ageism in Screening

Choosing the appropriate fairness criteria for age in hiring algorithms is especially tricky because of the causal relationship between the work experiences of applicants and their age. While the work experience of an applicant is a factor that employers use in making hiring decisions, using the age of the applicant in decision-making is illegal. On the one hand, age can be a barrier that reduces the applicant’s chances of getting an interview. On the other hand, it can also serve as a factor that helps the applicant because older applicants often have more work experience.

7.1.1 The Effect of Ignoring Age

Consider the effect of satisfying fairness through unawareness when training a model on a dataset where the outcome variable is contaminated by age discrimination. Firstly, assuming we know for a fact that age causally affects the experience level of an applicant, we can show that the parameter estimate from a model that does not take age into account will still be influenced by age (see Appendix A). Secondly, older applicants have higher experience levels on average. If discrimination is present in the dataset and the older applicants more often receive a negative callback, it is possible that the model will learn to reject applicants with high experience when we do not give the model information on the ages of the applicants. Thirdly, when we do not provide information on the ages of the applicants, the model sees a senior applicant with five years of experience to be the same person as a young applicant with five years of experience and the same skills. Therefore, the model will assume that any difference in their callback is due to randomness or unobserved factors. This is not desirable as it jeopardizes the learning process and can reduce the accuracy of our predictions.
7.1.2 Population Level Criteria Against Ageism

Note that an algorithm needs to give a positive callback to an equal proportion of applicants from each age group to satisfy independence. That is,

\[ Pr(callback = 1 | Age = \text{old}) = Pr(callback = 1 | A = \text{middle}) = Pr(callback = 1 | A = \text{young}) \]

Satisfying demographic parity in the test set enforces that we accept an equal proportion of applicants from each age group, even if a particular age group in our applicant pool has significantly high skills, which are not causally dependent on the ages of the applicants. Moreover, satisfying demographic parity in training is only reasonable when the training set is representative of the pool of potential applicants. If our training set is not representative of the pool of potential applicants, then satisfying demographic parity in training does not translate to satisfying it in a random sample from the pool of applicants.

Both separation and sufficiency require us to know the true callback variable, one that is not contaminated by discrimination. The true callback is a variable that captures whether the applicant is a good hire. Obtaining information on true callback is usually an unrealistic assumption in the hiring setting. One possible way through which we can obtain the true callback is by looking at the performance scores of the applicants after they spend some time working in the company. However, it is difficult to know that those scores aren’t also contaminated by discrimination.

7.1.3 Individual Criteria and Counterfactual Fairness against Ageism

Remember that counterfactual fairness compares the scores of two applicants: applicant \( i \), any applicant in the dataset, and applicant \( j \), who is the counterfactual version of \( i \) and belongs to a different age group. It is necessary to use a causally interpretable fairness criterion to distinguish between the attributes of the applicant that are causally dependent on age, and the ones that aren’t. Moreover, employing a causally interpretable fairness
definition allows the algorithm to provide feedback to the applicants on how they can improve their applications.

Neumark et al. (2019) stated that “[For] both policy and legal reasons, the right comparison for measuring age discrimination is between younger applicants and older applicants who have experience commensurate with their age” (Neumark et al., 2019). There are significant similarities between the fairness criteria articulated by Neumark et al. (2019) and the definition of counterfactual fairness under the assumption that the age of applicants causally affects their experiences, denoted by \( X_d \), but not applicants’ other skills, \( X_i \).

Following the interpretation of counterfactuals by Neumark et al. (2019) and considering other benefits of causally interpretable models, I propose a relaxation of counterfactual fairness to be an appropriate fairness criterion in addressing age discrimination.

### 7.2 Design Goals

#### 7.2.1 Similar Accuracy, Precision, and Recall across Age Groups

Accuracy, precision, and recall are measures that are used to evaluate the performance of a classifier, a machine learning model with binary output variable, given a test dataset. They are based on four basic statistics calculated for the applicants in the test set: (i) the number correct callbacks (TP), applicants with \( \hat{Y} = 1 \) and \( Y^* = 1 \), (ii) the number of incorrect callbacks (FP), applicants with \( \hat{Y} = 1 \) and \( Y^* = 0 \), (iii) the number of correct rejections (TN), applicants with \( \hat{Y} = 0 \) and \( Y^* = 0 \), and (iv) the number of incorrect rejections (FN), applicants with \( \hat{Y} = 0 \) and \( Y^* = 1 \).

**Definition 7.1 (Accuracy)** Accuracy of a classifier, evaluated on a test set, is the percentage of the data points that the model correctly classifies, i.e.,

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

**Definition 7.2 (Precision)** Precision of a classifier, evaluated on a test set, is the percent-
age of qualified candidates among the candidates who received positive callbacks, i.e.,

\[ \text{precision} = \frac{TP}{TP + FP} \]

**Definition 7.3 (Recall)** Recall of a classifier, evaluated on a test set, is the percentage of the candidates who received positive callbacks among the candidates who are qualified, i.e.,

\[ \text{recall} = \frac{TP}{TP + FN} \]

Objective 1. Similar accuracy, precision, and recall for applicants that belong to different age groups and have different experience levels. This is not a strict requirement for satisfying separation and sufficiency but an effort to improve the model’s fairness based on those criteria.

Inaccurate decision-making can result in qualified candidates getting rejections, i.e., low recall. The discrepancy in the model’s accuracy, recall, and precision for different age groups is a possible avenue through which bias can pervade the hiring process. Buolamwini and Gebru (2018) investigated a similar concern regarding the discrepancy in accuracy of facial recognition software across different races and genders. They noted that such discrepancies could arise from the under-representation of attributes correlated with the darker skin tones and gender display norms in the dataset, highlighting the importance of transparency in the demographic composition of training and benchmark datasets (Buolamwini and Gebru, 2018). In light of their findings, I suggest using a training set that contains an equal number of observations from different age groups, where the age groups have a balanced composition of different experience levels. While this helps the model achieve similar accuracy levels for different age groups, it complicates the relationship between the model’s fairness in the training set and the test set as age is not uniformly distributed across individuals in the population.

In an effort to satisfy Objective 1, I use a training set that contains an equal number
of applicants that belong to different age groups such that the number of applicants with different experience levels is balanced for each age group. While this is a useful step in achieving similar accuracy, precision, and recall across age groups, it does not guarantee the model’s performance in observational fairness metrics, separation, and sufficiency. However, by satisfying this requirement, we create a training set that is not representative of the population because age groups are not uniformly distributed.

7.2.2 Unlearning the Discrimination in the Training Set

Objective 2. Any dataset labeled by humans will contain explicit and implicit discriminatory tendencies of the humans as a part of the decision rule. Hence, when we naively train a model on this dataset, the model will learn the biases in the training set as a part of the decision criteria. We do not want our decision rule to contain these biases.

The present the algorithm that satisfies a relaxation of counterfactual fairness in Section 7.4, and elaborate on how it achieves a fairness in Section 7.5.

7.2.3 Accurately Estimating the Model Parameters

Objective 3. One of the obstacles in estimating accurate model parameters is the spurious correlations in the dataset. For instance, if age is correlated with an attribute $x$ in the training data but is not causally related to it, this correlation can result in inaccurate estimation of the parameters associated with the attribute $x$. I aim to eliminate these spurious correlations in the training set to enhance the accuracy of the parameters of our model.

In the population, age can correlate with skills that are not causally dependent on age, i.e., spurious correlations. If our observations in training set display a similar pattern, then our parameter estimates associated with those skills can be biased, resulting in inaccurate predictions. To avoid the effect of spurious correlations in the training dataset, I suggest
using datasets that have compositions similar to resume correspondence studies in economics, such as Neumark et al. (2019). Similarly, structuring the training dataset ensures that $A$ is not correlated with the factors that do not causally influence. This allows us to isolate the effect of skills on callback controlling for all other variables.

### 7.3 Procedure

The procedure presented in this chapter addresses (i) the concerns related to the causal relationship between age and experience levels, (ii) the imbalance in the accuracy of the model for different age groups, and (iii) the biases in model parameters that can arise due to spurious correlations. While I focus on age discrimination, to address the fairness concerns for other protected classes, such as race and gender, I suggest selecting different thresholds for each class such that demographic parity is satisfied in the pool of applicants.

**Step 1.** Firstly, to minimize the effect of statistical discrimination on decisions, I propose surveying the employers to identify the relevant attributes for decision making. From the survey, I suggest determining the number of age categories, $A$, and the categories for different experience levels, $X_d$, that are relevant in the decision-making process.

**Step 2.** Using the data generation process similar to the study by Neumark et al. (2019), we generate a set of resumes of artificial applicants whose ages are uniformly distributed and the experience levels within each category are balanced. However, it is desirable that the variation in the skills of the applicants is significantly higher than the variation in the datasets from resume correspondence experiments to enhance the accuracy of the model.

**Step 3.** Hiring managers across firms label the resumes with callback decisions.

**Step 4.** We create interaction variables, $A \times X_d$. For addressing age discrimination, we create interaction variables for each age category, experience-level pair, $(A = Age, X_d = Experience)$. 

36
Step 5. We train the model on the dataset with $X_i$, columns of $A$ other than age categories, and interaction variables $A \times X_d$ as input and callback, $Y$, as the target.

Step 6. For the correction, we take the conditional distribution of $X_d$ given $A$ as our input. Applying the correction discussed in the next section to the model parameters, we enforce a relaxation of counterfactual fairness on the model.

Step 7. When making predictions on a pool of new applicants, we use demographic parity to ensure fairness with respect to protected classes other than age.

7.4 Algorithm

Our algorithm takes two inputs: (i) the training data that consists of $X_i, X_d, A$ and the target $Y$, and (ii) the conditional distribution of $X_d$ given $A$ as inputs.

While it is difficult to access population statistics such as the conditional distribution of $X_d$ given $A$ for researchers, if a platform decides to use our methodology to prevent ageism in hiring, they will likely have access to that information from their data on the users.

After training the model on the training data, we correct the coefficients associated with the age/experience interactions such that for all young applicants, the scores assigned to a young applicant $i$ is equal to the expected score of the counterfactual applicant $j$, who is the counterfactual older version of applicant $i$ (see Definition 7.4).

We use the conditional distribution of $X_d$ given $A$ to calculate proper adjustments for the coefficients of the applicant types, i.e., age group and experience level interactions. We correct all the coefficients for a given age group by adding a constant. The constant is different for each age group and is equal to the difference between the weighted average of the coefficients for the types that belong to the youngest age group and the weighted average of the coefficients for types for the group concerned. The weights associated with the types are the probabilities from the conditional distribution of $X_d$ given $A$. 
7.5 Satisfying Fairness: A Case Study using the RC Experiment

Remember the age categories and associated experience levels in our dataset from Table 4. It is convenient to demonstrate how the algorithm works using a simple linear model.

\[ \hat{Y} = \text{skills}_i \beta + \pi_1 Y_L + \pi_2 M_L + \pi_3 M_H + \pi_4 M_B + \pi_5 O_H + \pi_6 O_L + \pi_7 O_B + \pi_8 O_B_L \]

Satisfying counterfactual fairness requires us to compare two applicants: an applicant \( i \) that we choose from the dataset, and a hypothetical applicant, applicant \( j \), who is the counterfactual version of applicant \( i \). The model will be counterfactually fair if it assigns these two applicants similar scores.

Let applicant \( i \) be a young applicant, and applicant \( j \) be a middle-aged applicant. Then, the score that the model assigns to applicant \( i \) is \( Y_i = \text{skills}_i \beta + \pi_1 \), and the expected score for the applicant \( j \) is

\[ Y_j = \text{skills}_j \beta + \pi_2 P(M_L|\text{Age} = \text{middle}) + \pi_3 P(M_H|\text{Age} = \text{middle}) + \pi_4 P(M_B|\text{Age} = \text{middle}) \]

We assume that the skills are not causally affected by age, so \( \text{skills}_i = \text{skills}_j \). Hence, we need

\[ \pi_1 = \pi_2 P(M_L|\text{Age} = \text{middle}) + \pi_3 P(M_H|\text{Age} = \text{middle}) + \pi_4 P(M_B|\text{Age} = \text{middle}) \]

to satisfy counterfactual fairness for these two applicants. Note that if this holds, counterfactual fairness is satisfied for any applicant \( i \) who is young, and their counterfactual middle aged version, applicant \( j \).

Similarly, for the same applicant \( i \), if applicant \( k \) is an old applicant who is counterfactual version of \( i \), to satisfy counterfactual fairness, we need

\[ \pi_1 = \pi_5 P(O_H|\text{Age} = \text{old}) + \pi_6 P(O_L|\text{Age} = \text{old}) + \pi_7 P(O_B|\text{Age} = \text{old}) + \pi_8 P(O_B_L|\text{Age} = \text{old}) \]
However, satisfying these conditions does not suffice to satisfy counterfactual fairness. For that, we need this equality to hold for all applicants \( i \) and their counterfactual version \( j \) in the dataset. That means the definition should also be satisfied for any middle-aged applicant \( j \) and old applicant \( k \). But this is more tricky as we know that if a person has high experience in middle-age, when we increase the age with an intervention, we know that the probability that their counterfactual version is an old applicant with low experience is 0. Hence, we can not use the same probability distribution to correct the model parameters. This motivates my definition of counterfactual fairness with respect to a base class, a relaxation of counterfactual fairness.

**Definition 7.4 (Counterfactual Fairness with respect to a Base Class)** We say \( \hat{Y} \) is counterfactually fair with respect to base class \( a_1 \) if under any context where \( X = x \) and \( A = a_1 \),

\[
P(\hat{Y}_{A \leftarrow a} = y|X = x, A = a_1) = P(\hat{Y}_{A \leftarrow a'} = y|X = x, A = a_1)
\]

for all values of \( y \) and for all values of \( a \) and \( a' \).

We use this relaxation of counterfactual fairness that requires counterfactual fairness to hold for the base class of young applicants, i.e., the applicant \( i \) in the definition is always from the young age group.

The parameter adjustment that we do in the coefficients fixes \( \pi_1 \), and shifts \( \pi_i \) for \( i = 2, 3, 4, 5, 6, 7, 8 \) so that

\[
\pi_1 = \pi_2 P(ML|Age = middle) + \pi_3 P(MH|Age = middle) + \pi_4 P(MB|Age = middle)
\]

and

\[
\pi_1 = \pi_5 P(OH|Age = old) + \pi_6 P(OL|Age = old) + \pi_7 P(OBE|Age = old) + \pi_8 P(OBL|Age = old)
\]
Figure 7: Conditional Distribution of Experience Levels Given Age Groups

This can be achieved by adding

$$\pi_1 - \pi_2 P(ML|Age = middle) + \pi_3 P(MH|Age = middle) + \pi_4 P(MB|Age = middle)$$

to $\pi_2, \pi_3,$ and $\pi_4$. Similarly, we add

$$\pi_1 - (\pi_5 P(OH|Age = old) + \pi_6 P(OL|Age = old) + \pi_7 P(OBE|Age = old) + \pi_8 P(OBL|Age = old))$$

to $pi_j$ for $j \in \{5, 6, 7, 8\}$

7.6 From Discrete to Continuous

While our dataset contains only three age groups, it is possible to apply the same methodology with more age groups and experience levels. In Figure 7, I provide an example of possible composition of experience levels across age groups. The colors represent different experience levels and associated probabilities: blue is 0 years, orange is 1 to 4 years, etc. Each bar describes a particular age group, showing the distribution of the experience levels of applicants who belong to that age group. I include Figure 7 here to demonstrate that if
we know the conditional distribution of experience levels given ages, which is a data easily accessible for platforms, our model is flexible and can work with a greater number of age groups and experience levels. We can easily correct the parameters, taking the youngest age group as our base class and using the probabilities from the conditional distribution.

One concern that arises from increasing the number of age groups and experience levels is that it leads to an exponential increase in the number of interaction variables included in our model, which raises two concerns: overfitting and representation of each category in the training data. Since our training data is artificially generated, it is possible to adjust it to ensure representation; however, we will also need to increase the number of observations in the dataset to have enough data points for each category. Techniques from the machine learning literature can be employed, such as Support Vector Machine (SVM) regression and cross-validation can be employed to avoid overfitting. While it is difficult to quantify, there will be an optimum number of age groups and experience levels that we want to include in the model so that the model has enough sensitivity while not having too many parameters. This is a potential avenue for future research.

8 Experiments

In my experiments, I use a subset of the dataset that only includes female administrative assistants, and has 21,000 observations. Firstly, I train a logistic regression model with X and A as input variables and observe the scores of applicants across different age groups and for different types of applicants. Secondly, I correct the parameters of the model so that they satisfy counterfactual fairness with respect to the base class of young applicants and observe the changes in the predictions. For simplicity, in the experiments, I assume that age is uniformly distributed in the population, and the following probabilities characterize the conditional distribution of experience levels given age.

1. $P(YL|\text{Age} = \text{young}) = 1$
Figure 8: Model Predictions Before Correction

2. \( P(ML|\text{Age} = \text{middle}) = 0.33 \)
3. \( P(MB|\text{Age} = \text{middle}) = 0.33 \)
4. \( P(MH|\text{Age} = \text{middle}) = 0.33 \)
5. \( P(OL|\text{Age} = \text{old}) = 0.25 \)
6. \( P(OBE|\text{Age} = \text{old}) = 0.25 \)
7. \( P(OH|\text{Age} = \text{old}) = 0.25 \)
8. \( P(OBL|\text{Age} = \text{old}) = 0.25 \)

Figure 8 shows the predictions when we use a naive logistic regression model. The horizontal axis shows the scores that the model assigns to the candidates. The scores of each age group are displayed at a different level and with a different color code. The threshold can be chosen depending on the number of interviews that the employer wants to offer. All applicants who lie above the threshold receive callback, and those who score below get rejections. We know from the data generating process that the age groups are matched in their skill levels. However, we can observe that there is a discrepancy between the
Figure 9: Model Predictions After Correction

Figure 10: Comparison of Predictions Before and After Correction
average score received by the age groups. This points to discriminatory decision-making by the model. The second graph in Figure 8 shows the decomposition of scores for eight different types of applicants for age-experience interactions. We can observe the workers with high experience got a higher score on average compared to the workers with low experience. This is an indication that the model captures the willingness of employers to hire experienced workers.

In Figure 9, we observe the predictions of the model after the correction that enforces counterfactual fairness with respect to the base class of young applicants. We see that the scores are matched across age groups. When we require the means for the coefficients associated with the age groups to be equal, we force the model to assign the same average score for each of the age groups, but we do not interfere with the influence of experience that the algorithm learned from the dataset. Hence, the model uses experience as a factor in decision-making while not using age. For instance, a closer look at the graph that displays the scores for eight different types reveals that now, old workers with high experience receive, on average, slightly higher scores than the young group, while the old workers with low experience still receive lower scores on average. Figure 10 presents a closer look at the improvement in fairness achieved by imposing counterfactual fairness with respect to the base class of young applicants.

The training accuracy and cross-validated test accuracy of the model range from 60% to 90% for different subsets of the dataset from the resume correspondence experiment when we select different occupations and cities. The online appendix contains a detailed description of the experiments.
9 Conclusion

The algorithm I proposed satisfies a relaxation of counterfactual fairness with respect to age. To do so, it uses a dataset that resembles datasets from resume correspondence experiments and the conditional distribution of experience levels given age. The algorithm satisfies counterfactual fairness with respect to a base class through a simple parameter adjustment after training.

I highlight that this algorithm can be an assistive tool that companies use to promote fair screening. The procedure unfolds as follows: Firstly, hiring managers review a set of fictitious resumes that are designed to extract precise information on the effect of attributes of the candidates on callback. This provides the dataset needed to train the algorithm. Secondly, platforms can provide information on the conditional distribution of experience levels given age groups. The counterfactually fair screening model can be trained using the dataset and information on the probability distribution. Then, when the hiring managers are evaluating candidates, they can see the fair score assigned to the candidate by the model. While the accuracy of the model in experiments with resume correspondence study was low, the dataset from Neumark et al. (2019) included observations from many different firms and cities. It is reasonable to think that different firms will have slightly different selection criteria, and jobs in different cities may have different skills that they seek from candidates, such as a second language in some places. These factors introduce noise to the outcome variable, and if the labeling of the dataset is done at the firm level, those noises can be eliminated, increasing the accuracy of our predictions.

My thesis has three main contributions. Firstly, I present how different fairness criteria in economics, law, and computer science literature relate to each other and explain how those fairness criteria apply to the case of ageism in screening algorithms. Secondly, I propose counterfactual fairness as the appropriate criteria to address age discrimination in algorithmic hiring. I formulate a new definition of fairness, a relaxation of counterfactual fairness, which addresses the same fairness concerns and is much simpler to satisfy with
a parameter adjustment in the model after training. Thirdly, I present an algorithm that satisfies the relaxation of counterfactual fairness for age. Although the correction in model parameters is designed to address age discrimination in screening, I use generic notation throughout the thesis to highlight that it is of interest to researchers who study other settings where the sensitive attribute of the individuals causally affects non-sensitive characteristics of the applicant. Future research can improve and modify the approach proposed in this thesis to be applied in other settings where similar causal structures exist.
A Fairness Through Unawareness and Confounding

In Figure 11, A, X and Y are defined as in Section 4. The main difference from Figure 3 is that here X includes both X_i and X_d, and δ includes the direct effect of A on Y, as well as the effect of A on Y through U. We assume that we know the causal model in Figure 11 is the true causal model that generates our dataset. Based on this information, I will show that excluding A from the model will result in biased parameter estimates. Furthermore, it is possible to show that if we exclude A, the parameter estimates will still be dependent on A. In our case study of hiring discrimination, A stands for age, X is the characteristics of the applicant that are justifiably used in decision making, and Y is the callback variable that is contaminated by discrimination. Excluding age from the model results in estimation of model parameters that still depend on A.

\[ Y = X\beta + A\delta + \epsilon \quad \text{True Model for predicting } Y, \ E(\epsilon) = 0 \]
\[ Y = X\beta^S + \epsilon^S \quad \text{Model for calculating the effect of } X \text{ on } Y \]
\[ A = X\pi + \epsilon^a \quad \text{auxiliary regression to calculate } \pi \]

For simplicity, assume Gauss-Markov assumptions hold. The true model for Y is a linear function of A and X. Hence, Ordinary Least Squares (OLS) is appropriate to quantify the effect of X and A on Y and to predict the true scores of the applicants. Then, picking a threshold, employers can offer interviews to all applicants that are above the threshold and reject other applicants. In this setting, Gauss-Markov Assumptions are
GM.1 True model is linear in Parameters: \( Y = X\beta + A\delta + \epsilon \) with \( E[\epsilon] = 0 \)

GM.2 The columns of \( X \) and \( A \) are linearly independent from each other, i.e. no perfect multicollinearity

GM.3 Zero Conditional Mean: \( E[\epsilon|X,A] = 0 \), which implies uncorrelatedness between \( [X|A] \) and \( \epsilon \)

GM.4 Homoskedasticity and no autocorrelation: \( \text{Var}(\epsilon|X,A) = E[\epsilon\epsilon^T|X,A] = \sigma^2 I \), where \( \sigma \) is a scalar and \( I \) is the \( n \) by \( n \) identity matrix

Note that \( X \) contains columns of 1s, hence \( \beta, \beta^S, \) and \( \pi \) includes intercepts. Then, if we use OLS, we can estimate \( \hat{\beta}^S = (X^TX)^{-1}Y \). Plugging in the true model we get

\[
\hat{\beta}^S = (X^TX)^{-1}X^T Y = (X^TX)^{-1}X^T (X\beta + A\delta + \epsilon) = \beta + (X^TX)^{-1}X^TA\delta + (X^TX)^{-1}X^T \epsilon
\]

Therefore, the expected value of \( \beta^S \) is

\[
E[\hat{\beta}^S] = E[E[\hat{\beta}^S|X,A]] = E[E[\beta + (X^TX)^{-1}X^TA\delta + (X^TX)^{-1}X^T \epsilon|X,A]]
\]

\[
\implies E(\hat{\beta}^S) = \beta + (X^TX)^{-1}X^T A\delta + E[(X^TX)^{-1}X^T E[\epsilon|X,A]]
\]

\[
\implies E(\hat{\beta}^S) = \beta + (X^TX)^{-1}X^T A\delta \text{ assuming A3 holds.}
\]

Realize that the parameter estimate that we get by excluding age from the regression model still depends on \( A \) and \( \delta \).

Therefore fairness through unawareness \( A \) does not make the estimated model parameter model independent of \( A \). Also, \( E[\hat{\beta}^S] = \beta + \hat{\pi}\delta \) since \( \hat{\pi} = (X^TX)^{-1}X^T A \). \( \hat{\beta}^S \) captures both (i) \( \beta \), the direct effect of \( X \) on \( Y \), i.e. the direct effect of productivity related characteristics on the callback decisions made by the hiring managers, and (ii) \( \hat{\pi}\delta \), the relationship between \( X \) and \( Y \) that is associated with the changes in \( A \) that affects both \( X \) and \( Y \). The coefficient \( \delta \) captures disparate treatment (including taste-based discrimination and statistical
discrimination), and \( \pi \) captures the covariance between \( X \) and \( A \) which creates a potential avenue for disparate impact in the model.

**Proposition 1** Whenever \( \| \pi \| > 0 \) (sensitive attribute correlates with \( X \)) and \( \| \delta \| > 0 \) (there is disparate treatment in the training data), we will obtain biased model coefficients by satisfying fairness through unawareness.
B Counterfactual Inference

Definition B.1 (Causal Model) A causal model is defined by the triple \((B, V, F)\), where

- \(B\) is a set of background variables that are not caused by any variable in the set \(V\)
- \(V\) is the set of observable variables
- \(F\) is a set of functions, \(\{f_1, \ldots, f_n\}\), such that \(f_i(V_{\text{parents}_i}, B_{\text{parents}_i}) = V_i\), where \(V_{\text{parents}_i} \subseteq V \setminus V_i\) and \(B_{\text{parents}_i} \subseteq B\).

(Kusner et al., 2018; Pearl, 2009)

The model is causal because it is possible to derive the distribution of the \(Z \subseteq V\) after an intervention of \(V \setminus Z\) when the distribution over the background variables \(B\) are given. Specifically, the effect of an intervention to an observable \(V_i\) on another observable \(V_j\) can be calculated by setting the value of \(V_i = v'\), and plugging the value in the structural equation for \(V_j\), i.e. \(V_j = f_j(V'_{\text{parents}_j}, B_{\text{parents}_j})\), where \(V'_{\text{parents}_j}\) is the values of parents \(V_j\) with \(V_i\) is set equal to \(v'\) in the equation. Kusner et al. (2018) highlights that this process captures the idea of an agent, external to the system, modifying the value of \(V_i\) by forcefully assigning it to be \(v'\), “for example as in a randomized experiment.” They note that the structural equations are strong assumptions, but are necessary for calculating the counterfactual quantities (Kusner et al., 2018).

The counterfactual statement “the value of \(Y\) if \(A\) had taken the value \(a\)”, for two observable variables \(Y\) and \(A\) correspond to the question how would the callback values change if the applicant was old. Mathematically, the counterfactual model for \(Y\) given \(B\) where \(A = a\) is denoted as \(Y_{A\leftarrow a}^{a\leftarrow} \) or \(Y_a\).

Definition B.2 (Counterfactual Inference) Counterfactual inference is specified by a causal model \((B, V, F)\) given evidence \(W\) is the computation of probabilities \(P(Y_{A\leftarrow a}^{a\leftarrow}(B)|W = w)\), where \(W, A, Y \subseteq V\). The inference follows three steps:
1. Abduction: for a given prior on $B$, compute the posterior distribution of $B$ given $W = w$

2. Action: replace the initial values of $A = a$ in the equations $F$ with the values $A = a'$ after the intervention

3. Prediction: Using the set of equations $F_{a'}$ and the remaining elements of $V$, compute the posterior $P(B|W = w)$.

(Kusner et al., 2018; Pearl and Mackenzie, 2018; Pearl, 2009)

The posterior can then be used to make counterfactual inference about the values for $Y$, i.e. what would the callback be if $A = a'$.

**Definition B.3 (Counterfactual Fairness)** $\hat{Y}$ is counterfactually fair if under any context where $X = x$ and $A = a$,

$$P(\hat{Y}_{A\leftarrow a} = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'} = y|X = x, A = a')$$

for all values of $y$ and for all values of $a$. (Kusner et al., 2018)
References


C. Schumann, J. Foster, N. Mattei, and J. Dickerson. We need fairness and explainability in algorithmic hiring. 2020.


