FAIRNESS IN AI-BASED RECRUITMENT AND CAREER PATHWAY OPTIMIZATION

By

Dena Freshta Mujtaba

A THESIS

Submitted to Michigan State University in partial fulfillment of the requirements for the degree of

Electrical Engineering – Master of Science

2020

ABSTRACT

FAIRNESS IN AI-BASED RECRUITMENT AND CAREER PATHWAY OPTIMIZATION

By

Dena Freshta Mujtaba

Work has long been a source of human livelihood, financial security, mental and physical well-being, dignity, and meaning. However, advances in computing, big data, artificial intelligence (AI), robotics, and related technologies are expected to usher in unprecedented and widespread changes in the economy and society. It is estimated that by 2030 up to 14% of the global workforce may need to change occupational categories as the world of work is disrupted by technological advances. Many current and future workers that will enter the workforce lack skills that in-demand and future jobs require. In short, the landscape of work is poised for a major and unprecedentedly rapid transformation and this calls for a variety of strategies to meet the needs of workers, employers, the economy, and broader society.

Motivated by these concerns, we investigate two key problems faced by organizations and workers in the future of work. As AI has expanded into human resource applications, organizations are increasingly using AI-based recruitment for sourcing, screening, and selecting talent. We explain how this can lead to biases in decisions and how this bias can be measured, review tools available for bias mitigation, and discuss future challenges for fairness in machine learning specific to recruitment applications. Alongside this, workers are affected not only by biased recruitment, but by the growing automation of tasks in occupations, which will increasingly require job and task transitions. To help workers navigate these transitions effectively, we propose a genetic-algorithm-based optimization engine to search for a worker's optimal career pathway in a network of occupations, given their current knowledge, skills, abilities, and other work-related characteristics. Overall, this thesis presents strategies for organizations to mitigate bias in AI-based recruitment and for workers to plan their career pathway in the face of unprecedented changes in the world of work.

Copyright by DENA FRESHTA MUJTABA 2020

ACKNOWLEDGEMENTS

I would like to express my thanks and appreciation to my advisor Dr. Nihar Mahapatra, who encouraged me to start my graduate studies. This research would not have been possible without his support and guidance. I would also like to thank my committee members Dr. Fathi Salem and Dr. Jiliang Tang for serving on my committee and their time in reviewing this work. I would also like to thank my family, friends, and colleagues for their support and encouragement throughout my research.

TABLE OF CONTENTS

LIST O	F TABLES vii
LIST O	F FIGURES
LIST O	F ALGORITHMS
CHAPT	TER 1 INTRODUCTION 1
1.1	Future of Work
1.2	Motivation and Problem Statement
	1.2.1 Fair Treatment of Workers
	1.2.2 Worker Job Search and Pathway Planning
1.3	Related Work
-	1.3.1 Recruitment Methods
	1.3.2 Worker Career Planning
1.4	Contributions
1.5	Thesis Organization
СЦАРТ	$\mathbf{F} \mathbf{D} \mathbf{O} = \mathbf{F} \mathbf{A} \mathbf{D} \mathbf{N} \mathbf{E} \mathbf{C} \mathbf{C} \mathbf{D} \mathbf{A} \mathbf{C} \mathbf{E} \mathbf{D} \mathbf{D} \mathbf{E} \mathbf{C} \mathbf{D} \mathbf{U} \mathbf{U} \mathbf{T} \mathbf{M} \mathbf{E} \mathbf{N} \mathbf{T} $
ONAF I	LER 2 FAIRNESS IN AI-DASED RECRUITMENT
2.1	$\begin{array}{cccccccccccccccccccccccccccccccccccc$
0.0	$2.1.1 \text{Our Contributions} \dots \dots \dots \dots \dots \dots \dots \dots \dots $
2.2	Background
	$2.2.1 \text{Uauses of Bias} \dots \dots$
0.0	2.2.2 Definitions of Fairness
2.3	Bias in Recruitment
	2.3.1 Identifying and Attracting Candidates
	2.3.2 Candidate Screening/Processing
	2.3.3 Communication and Selection of Candidates
2.4	Bias Detection and Mitigation
	2.4.1 Methods of Bias Mitigation 19
	$2.4.2 \text{Tools} \dots \dots$
2.5	Conclusion and Future Work
CHAPT	TER 3 THE OCCUPATIONAL INFORMATION NETWORK (O*NET) 25
3.1	Background
3.2	Content Model 26
0.2	3.2.1 O*NET Interest Profiler 20
	3.2.2 Green Occupations and Tasks 31
	32.3 Bright Outlook 31
	$3.2.4 \text{Ioh Zones} \qquad \qquad$
	3.2.5 Career Changers and Beginners Matrix
22	Applications
ე.ე ე_/	Highlighting Critical Occupations and Carcor Daths
3.4	ing ing ortical Occupations and Oareer Paths

CHAPT	ER 4	TRANSFERABILITY OF WORKER COMPETENCIES						
4.1	Motiva	vation and Problem Statement						
4.2	Contri	butions $\ldots \ldots 37$						
4.3	Metho	ds						
	4.3.1	GloVe						
	4.3.2	BERT						
4.4	Result	lts and Evaluation						
	4.4.1	Knowledge Transferability						
	4.4.2	Skills Transferability						
	4.4.3	Abilities Transferability						
4.5	Conclu	usion and Future Work						
СНАРТ	EB 5	OPTIMIZATION OF CAREER PATHWAYS 52						
5.1	Motive	ation and Problem Statement 53						
5.2	Contri	butions 54						
5.3	Metho	d 54						
0.0	531	Cenetic Algorithms 55						
	539	Path Representation 56						
	533	Fitness Function 50						
	0.0.0	$5331 \qquad \text{Occupation Distance} \qquad 50$						
		5.3.3.2 Job Zono and Caroor Clusters 60						
		5.3.3.2 JOD Zone and Career Clusters						
		5.3.3.4 Competency Growth and Decay						
		5.3.5.4 Competency Hanslerability						
	591	$\begin{array}{cccccccccccccccccccccccccccccccccccc$						
	0.0.4	$\begin{array}{cccccccccccccccccccccccccccccccccccc$						
		5.3.4.1 Random Occupation Assignment						
	595	Selection 60						
	0.0.0	$5251 \qquad \qquad$						
		$5.5.5.1 \text{Iournament} \qquad \qquad$						
	526	5.5.5.2 Lexicase						
	0.3.0	$Crossover \qquad \dots \qquad $						
		$\begin{array}{cccccccccccccccccccccccccccccccccccc$						
Γ 1		$\begin{array}{cccccccccccccccccccccccccccccccccccc$						
5.4		ation and Results						
	5.4.1	Crossover Experiments						
	5.4.2	Selection Experiments						
~ ~	5.4.3	Discussion \dots 74						
5.5	Conclu	usion and Future Work						
СНАРТ	CONCLUSION83							
BIBLIO	GRAP	НҮ 85						

LIST OF TABLES

Table 2.1:	Sample application selection rate.	13
Table 3.1:	The worker-oriented, occupation-specific, cross-occupation, and job-oriented characteristics in the O*NET content model.	28
Table 3.2:	Holland's occupational themes, or the RAISEC model	29
Table 3.3:	Occupations marked with bright outlook, poor outlook, or green, used for evaluation.	35
Table 4.1:	Example of the GloVe model training method using global word co- occurrence probabilities, adopted from [1]	40
Table 5.1:	Parameters used for the genetic algorithm.	72

LIST OF FIGURES

Figure 1.1:	The different parts that make up a worker, as reflected in their resume	6
Figure 2.1:	Google search trend graph from Aug. 2010 - Aug. 2019 for the phrases "machine learning bias" (blue), "HR AI" (red), and "ethical hiring" (yel- low), showing relative interest over time, with 100 indicating peak pop- ularity.	12
Figure 2.2:	An overview of the steps in a typical recruitment pipeline: (1) identi- fying and attracting candidates, (2) candidate screening or processing, and (3) communication and selection of candidates	17
Figure 2.3:	Overview of user/developer engagement on fairness and interpretability repositories on GitHub, measuring the number of stars(i.e., marked as a favorite by a user), watch count (i.e., user notified of updates), and forks (i.e., user copies project code to contribute or customize).	22
Figure 3.1:	High level overview of the major components of the O*NET content model representation of an occupation.	26
Figure 3.2:	The occupation "Statistical Assistant" and the corresponding KSAOs with the importance provided	27
Figure 3.3:	The different RAISEC codes mapped to an occupation	30
Figure 3.4:	Distribution of RAISEC code values for the first occupation in each cluster.	30
Figure 4.1:	The overall process to get the transferability between two competencies in O*NET, using pre-trained embeddings and a knowledge-based approach.	39
Figure 4.2:	Example GloVe vector distances, adopted from [2]	41
Figure 4.3:	The transferability of knowledge from O*NET using the GloVe-based approach.	43
Figure 4.4:	The transferability of knowledge from O*NET using the BERT-based approach.	44
Figure 4.5:	The transferability of basic skills from O*NET using the GloVe-based approach.	45

Figure 4.6:	The transferability of basic skills from O*NET using the BERT-based approach.	46
Figure 4.7:	The transferability of cross-functional skills from O*NET using the BERT-based approach.	47
Figure 4.8:	The transferability of cross-functional skills from O*NET using the GloVe-based approach	48
Figure 4.9:	The transferability of abilities from $\mathrm{O}^*\mathrm{NET}$ using the GloVe-based approach.	49
Figure 4.10:	The transferability of abilities from O*NET using the BERT-based approach	50
Figure 5.1:	The overall architecture of the career pathway planning system, and the GA (wrapped in the box). The inputs are the end-user's KSAOs (typically extracted from a resume through natural language processing), the desired occupation from O*NET, and the occupation space provided by O*NET. The output is the most fit solution from O*NET, or the best fitting career pathway for the end-user.	56
Figure 5.2:	Distribution of the distances between every occupation to occupation pair in O*NET, using only competencies to calculate the distances (i.e., $w_C = 0$)	61
Figure 5.3:	Distribution of the distances between every occupation to occupation pair in O*NET, using only requirements to calculate the distances (i.e., $w_R = 0$).	61
Figure 5.4:	Distribution of the distances between every occupation to occupation pair in O*NET, using weights $W_R = 0.25, W_C = 0.25$, and $W_Z = 50$.	62
Figure 5.5:	The overall structure used to represent the time in an occupation, and the competencies of the user through each occupation. The user starts with a list of competencies, and these grow or decay depending on the occupation they spent time in. Furthermore, each individual has a list of occupation-time pairs that are used to reference this information	63
Figure 5.6:	The distribution of level and importance for knowledge in O*NET. $\ . \ .$	65
Figure 5.7:	The distribution of level and importance for skills in O*NET. \ldots .	65
Figure 5.8:	The distribution of level and importance for abilities in O*NET	66

Figure 5.9:	The growth and decay functions implemented on the skill "Mathematics" for two highly different occupations.	66
Figure 5.10:	Fitness (i.e., cost) results for the two-point crossover mechanism for bright and green occupations.	73
Figure 5.11:	Fitness (i.e., cost) results for the one-point crossover mechanism for bright and green occupations.	73
Figure 5.12:	Fitness (i.e., cost) results for the uniform crossover mechanism for bright and green occupations.	74
Figure 5.13:	Fitness (i.e., cost) results with a random selection mechanism for bright and green occupations.	75
Figure 5.14:	Fitness (i.e., cost) results for the lexicase selection mechanism for bright and green occupations	75
Figure 5.15:	The distribution of pathway costs from the occupation "41-2011.00" to all others in O*NET	76
Figure 5.16:	The distribution of pathway costs for green and bright occupations in O*NET from "41-2011.00" to all others in O*NET	77
Figure 5.17:	Path costs highlighted from Figure 5.18	78
Figure 5.18:	Sample paths to show different examples of results illustrating differ- ent aspects from the approach. Nodes in the pathway are color coded according to different career clusters. The sun symbol indicates an occu- pation with a bright outlook. The leaf symbol indicates an occupation that is green. Each node between the start and end occupation have the time spent in the occupation in months (m).	79
Figure 5.19:	Comparing the results of the GA with modifications against common shortest path algorithms such as Djikstra's. The green line is the cost of the path found by the shortest path algorithm	80

LIST OF ALGORITHMS

Algorithm 1: Ca	alculating the co	ost of a pathway.						68
-----------------	-------------------	-------------------	--	--	--	--	--	----

CHAPTER 1

INTRODUCTION

Advances in artificial intelligence (AI) and related technologies are expected to usher in unprecedented and widespread changes in the economy and society. It is estimated that by 2030 about 14% of the global workforce may need to change occupational categories as the world of work is disrupted by technological advances [3]. This is a challenge similar in scope to that faced in the early 20th century in North America and Europe, and in China in recent decades, due to a large-scale shift from agriculture to manufacturing, except now occurring over a much more compressed timeframe, and being particularly acute in advanced economics like the US [3]. Addressing this grand challenge requires a holistic understanding of occupational categories, worker characteristics, and a way to connect workers to future jobs and help them transition from one type of work to another and plan their career pathway. This thesis presents several components that together address these needs using artificial intelligence and optimization.

In this chapter, we first provide an overview of the future of work, and how the evolution of technology is affecting the world of work. Within the world of work, we present the key problems to be addressed for different stakeholders being affected by automation, globalization, and climate change. Next, we present the scope of this thesis and the problems addressed by this work, followed by an overview of related work in the area of using artificial intelligence to assist workers.

1.1 Future of Work

Work has long been a source of human livelihood, financial security, mental and physical well-being, dignity, and meaning [4]. However, technological progress, globalization, and demographic change are altering the world of work and are poised to do so even more dramatically in the future [5]. Globalization allows work to cross national borders in pursuit

of appropriately skilled workers, resulting in offshoring (i.e., when there is a comparative advantage in moving work abroad) and reshoring (i.e., when the comparative advantage diminishes) [4, 5]. Aging and the retirement of baby boomers in OECD countries are expected to cause major changes in consumer spending patterns and the type of work that will be in demand [5] – e.g., a likely shift in demand from durable goods (e.g., cars) toward services (e.g., healthcare) [6]. Advances in computing, big data, AI, and robotics are expected to usher in unprecedented and widespread changes in the economy and society [7]. It is estimated that by 2030 up to 14% of the global workforce may need to change occupational categories as the world of work is disrupted by technological advances [3]. In OECD countries, about 14% of jobs are at high risk of automation and 32% face the prospect of radical transformation [5]. Many current workers and those entering the workforce lack skills that in-demand jobs and jobs of the future require [8, 9, 10]. Non-standard work (e.g., art-time, temporary, contract, multiple, on-demand/gig-economy) is no longer a marginal phenomenon [5]. In short, the landscape of work is poised for a major and unprecedently rapid transformation and this calls for a variety of strategies to meet the needs of workers, employers, the economy, and broader society.

Early research in labor economics on the impact of AI and related automation technologies on the future of work relied on analysis at the occupation or job level and projected job losses ranging from 35% to 60% or more in the US and Europe in the next couple of decades [11, 12, 13, 14]. However, jobs are constituted from task groups and when automation technologies are introduced, automation occurs at the task rather than at the job or occupation levels [15, 16, 17]. Automation technologies impact job growth through three channels [18]. First, technology can directly automate certain tasks (i.e., the displacement effect). Second, new jobs may be created that require workers to work alongside and/or manage smarter machines (i.e., the skill-complementarity effect). Finally, higher productivity from the use of these technologies increase demand due to lower prices and overall increase in disposable income (i.e., the productivity effect). These effects do not all occur simultaneously and there can be significant uncertainty and variation in automation across occupations and countries, and due to ethical and legal obstacles to automation [15, 18]. This task-level view of jobs significantly revised downward the estimate of the extent of jobs at high risk of loss to 14% and those expected to rapidly transform to 32% in OECD countries [5, 15]. Thus, at the task level, there can be significant impact on jobs as certain tasks get automated, others changed or augmented, and yet others added. In view of this, a solution is needed to transform decision support for workers and organizations offering future jobs.

1.2 Motivation and Problem Statement

The future of work presents many challenges to be tackled by researchers to assist workers in transitioning jobs. Workers are affected not only by the changing skills and knowledge requirements for future jobs, but by recruitment practices and jobs produced by organizations. Therefore, in this thesis we focus on defining and addressing the key issues from both the organization and worker side. The specific motivation for each area and problem we address are further described next.

1.2.1 Fair Treatment of Workers

Past research has shown a correlation between unfair workplace practices and lower job satisfaction [19]. This stems from organizational justice, the concept of justice or fairness in decisions, outcomes, and actions made by an organization, from the perspective of an employee [19, 20]. Procedural justice is fairness of the processes that lead to certain decisions being made [21, 22]. If unfair decisions are repeatedly made by an organization's executives or managers, employees may develop a negative attitude towards these individuals, and eventually lose trust in the organization [20, 23]. One of the first experiences a person will have with their organization is in their hiring process - unfair hiring practices have shown to reflect poorly on a company and make it harder to hire candidates in the future [21]. Therefore, it is crucial each step of the hiring pipeline enforces fair decision making and

transparency with the candidate.

In the hiring pipeline, candidates are first screened and evaluated with their resume, CV, and/or cover letter. With the increase in employment-oriented services such as HireVue and Indeed [24, 25], resume screening tools have become a standard [26]. An *applicant tracking system* (ATS), is an app which HR professionals can store and filter applicants' resumes and information given certain criteria such as education or skills [26]. Past tools, such as BambooHR, Ascendify, and Lever, boost productivity and improve the recruitment process for the candidate and employer by sending automated emails, notifications, and using machine learning techniques for parsing resumes [27]. More recently, AI has been built into ATSs, such as Jobspotting, that match candidates to job openings with AI [28]. Interactive services based on AI, such as question-answer bots, have also become more common, allowing recruiters to save time in interacting with every candidate [28]. ATSs and resume parsing in candidate screening has become a standard in most large companies, and as AI progresses, these services will become more accurate, and will be adapted across the HR industry [29].

By using automation, we are seemingly removing human biases in the screening process - a common problem shown in past research [22, 30]. However, this bias has been found to carry over to machine learning applications involved with people [31, 32]. This stems from biases in the training data used for developing models, such as skewed data towards certain predictions or proxies of protected attributes (i.e. non-protected attributes that may be correlated with protected attributes, resulting in bias) [32]. Therefore, it is crucial to understand the methods to mitigate bias in these applications and provide a fair recruitment process for workers.

1.2.2 Worker Job Search and Pathway Planning

As the jobs landscape is expected to undergo change, it requires workers to make optimal decisions in work transition and career planning. However, searching for the correct job and planning their career pathway present several challenges for workers, including: (a) finding occupations that best fit their worker characteristics, and (b) forecasting the demand of an occupation or skill that they acquire over time.

A substantial amount of research has shown that individuals with certain characteristics will be more successful in some jobs over others [33, 34, 35]. A worker's performance, or person-job fit, can be determined by their knowledge, skills, abilities, and other work-related characteristics (KSAOs) [36], as shown in Figure 1.1. Therefore, it is important to consider these factors in career planning approaches. Organizations face a problem in that they may match individuals to available jobs ineffectively using only a narrow range of characteristics (e.g. personality, interests, interview scores, etc.). This can lead to potential errors in the employee selection process. In addition, it is often difficult for organizations to identify the potential relevance of an applicant's past employment history for a job that he or she may be applying for unless the job is a one-to-one mapping. This may also result in selection errors for individuals who are changing careers or do not have a background in the job the organization is hiring for. This is an important issue for many individuals, such as veterans and the U.S. military where it is often unclear how some military jobs may be relevant to civilian occupations. For example, it is often unclear what KSAOs obtained by Soldiers in the Infantry may be relevant for many civilian occupations. Therefore, use of KSAOs in career pathway planning can help workers and organizations and identify occupations with the best person-job fit.

Furthermore, the demand for a given skill (or KSAO in general) or occupation will vary over time as AI and related technologies cause widespread changes in the world of work. Certain skills will no longer be needed as automation increases, and workers applying to a job will need other skills or new occupations. Though past skills taxonomies, such as O*NET (further discussed in Chapter 3), include information on skills that are frequently found in job postings and occupations that have a bright outlook, this information is for the present user using their database, and does not provide a future prediction of where a skill may be headed, or what the outlook will be further in the future. For workers, when assessing



Figure 1.1: The different parts that make up a worker, as reflected in their resume.

a job transition, the user will need to consider these factors and how it may impact their career pathway. Furthermore, representing these features will greatly lower training costs for workers looking to take courses or get certifications for a certain job that they believe will be needed in their career pathway.

1.3 Related Work

Many AI-based methods and applications have been developed to address the issues found in the future of work. We outline two relevant areas of related work for this thesis which consist of tools and models for recruitment and data-driven approaches to career planning for workers. These and the corresponding related work are further described next.

1.3.1 Recruitment Methods

Much work within AI-based recruitment and person-job fit assessment have been developed in the past decade [37, 38, 39, 40]. Initially, several approaches have focused on parsing resumes [41, 42, 43, 44, 45, 46] and using this for assessing person-job fit [37, 38, 47, 48]. These approaches have often utilized deep neural networks (DNNs) to automatically extract features of the data [37, 47, 48, 49], and show improved results compared to past statistical models or collaborative-filtering methods [50].

There are several deep neural network baselines developed for the person-job fit problem. The Hierarchical RNN Matching (HRNNM) model was presented by Li et al. [48], and the job description and resume to assess fit. Next, the Person-Job Fit Neural Network (PJFNN) is a convolutional neural network (CNN) approach to the person-job fit classification task [37], and similarly, the Ability-Aware Person-Job Fit (AAPJF) network focuses on person-job fit but uses an recurrent neural network (RNN) to encode the information [38]. Lastly, the Interpretable Person-Job Fit (IPJF) model focuses on the person-job fit problem from the employer perspective by learning the intent of the job seeker and employer to assess fit [51]. Overall, these approaches are a step towards automated recruitment and assisting employers to easily screen and assess job candidates for their organization. One area of concern for these models, is whether they are subject to biased decisions, which could in turn hurt a worker's ability to get hired.

1.3.2 Worker Career Planning

Past approaches to optimize career planning and predict job transitions have used occupational data found on sites such as LinkedIn [52] or Indeed [24]. These are also known as *online professional networks* (OPNs), and can be used to look at factors such as future job payoff [53], company influence [54], and ease or likelihood of job transitions [55, 56]. Approaches have used this information to predict or recommend jobs [57, 58], or predict the trajectory of a user's career path [47, 59]. The challenge in these problems is representing the worker, which can lead to a high dimensional space with several attributes such as companies, sequence of job positions, and their skills. However, to alleviate this, past approaches such as NEMO [59] and IPJF-Transfer [51] have used encoder-decoder models to map the worker profile into a fixed-length vector for the neural network. Though these approaches use a large amount of information on real workers, OPNs can be inconsistent and have gaps, depending on what is shared in job postings, user profiles, or salary outlook.

1.4 Contributions

In the face of these unprecedently rapid changes in the landscape of work, we present an overview of fairness in AI-based recruitment that may affect workers transitioning jobs, and seek an effective, data-driven solutions for worker career pathway planning. There are two challenges that are addressed in worker career pathway planning: (a) how should work, workers, and work environments be understood to allow workers to transfer their skills to new occupations, and (b) how do we effectively facilitate work transition and help workers plan their career pathway.

To address the first challenge, we provide a method to semantically relate different KSAOs describing a worker. This method establishes links between the different occupations and KSAOs with a similarity, or transferability measure. With this, worker's skills and knowledge can be moved between occupations, and is then used in addressing the next challenge.

To address the second challenge, we develop a method to capture the training effort needed to transition between two jobs belonging to different occupational categories and formulate an optimization problem to identify career pathways that would best suit the worker's KSAOs. Currently, there is no AI-based method available to analyze workers based on their KSAOs. Therefore, we provide a novel way to facilitate decision making for job transitions and career planning.

1.5 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 provides a detailed overview of the problem of bias in machine learning models and more specifically, AI-based recruitment applications, which have a direct impact on a worker's ability to get hired for a job. Next, Chapter 3 covers the Occupational Information Network (O*NET), a taxonomy of occupations and worker characteristics, and the key dataset used in this thesis. Then, Chapter 4 presents a method to link knowledge and skills from O*NET with their transferability. Next, Chapter 5 describes the problem of career planning and job transitions from the worker perspective, and our solution modeling the job transition cost and searching for the optimal pathway for the worker using O*NET. Last, Chapter 6 concludes this work and discusses some potential future improvements.

CHAPTER 2

FAIRNESS IN AI-BASED RECRUITMENT

Over the past few years, machine learning and AI have become increasingly common in human resources (HR) and worker-based applications, such as candidate screening, resume parsing, and employee attribution and turnover prediction. Though AI assists in making these tasks more efficient, and seemingly less biased through automation, it relies heavily on data created by humans, and consequently can have human biases carry over to decisions made by a model. Several studies have shown biases in machine learning applications such as facial recognition and candidate ranking. This has spurred active research on the topic of fairness in machine learning over the last five years. Several toolkits to mitigate biases and interpret black box models have been developed to promote fair algorithms.

To further fair algorithms in AI-based recruitment applications that affect the ability of workers to fairly go through the recruitment process, this chapter presents an overview of fairness definitions, methods, and tools as they relate to recruitment and establishes ethical considerations in the use of machine learning in the hiring space [60].

The remainder of this chapter is organized as follows. Section 2.1 covers the motivation behind this survey, and Section 2.1.1 presents an overview of our contributions following this. In Section 2.2, we identify the various causes of bias and provide five core definitions of fairness. Next, in Section 2.3, we outline the recruitment process and discuss the ways in which bias may be introduced in AI-based recruitment processes. Section 2.4 details the main categories of bias mitigation methods and the key fairness tools available to address bias in machine learning systems. Last, Section 2.5 covers limitations of the tools and methods for bias detection and mitigation and outlines future challenges in the area of AI-based recruitment.

2.1 Motivation

Recently, AI has been adopted in the human resources (HR) industry, for purposes such as predicting employee attrition, chatbot systems for HR service delivery, and background verification for screening applicants' resumes [61]. Initially, HR technology was an effective approach for tasks such as applicant resume screening and candidate sourcing, but it has recently been recognized as a critical need for organizations to scale and expand their businesses [61]. Applications such as Arya, Google hire, HireVue, and Plum use AI for applicant screening and management [25, 62, 63, 64], allowing businesses to grow efficiently, without the need for human decision making in these hiring processes.

However, as AI adoption becomes more widespread, there is growing concern that there are several ways the decisions made by such systems could be carrying over biases from people in the organization or the model developers, as evidenced by several recently-reported episodes. In 2017, Amazon ended its AI-based candidate evaluation tool because it was shown to discriminate against female candidates, assigning lower scores to resumes of women when ranking applications [65]. The model's bias was a result of under representation of female applicants in the training dataset used to create the model. This is an example of how biases commonly found in the hiring process (e.g., hiring discrimination [66]) can easily carry over to AI-based approaches through the data used to train the algorithm.

Companies using AI-based approaches for recruitment expect a more consistent and ethical approach to decision making and for them to possess the ability to remove biases in contrast to a human decision maker [67]. However, a number of recent research studies have shown biases in machine learning and AI-based applications [31, 32, 65, 68]. This has contributed to an increase in interest in fairness, machine learning bias, and AI in recruitment, as depicted in the Google search trend graph in Figure 2.1.



Worldwide Relative Search Interest in AI-Based Recruitment

Figure 2.1: Google search trend graph from Aug. 2010 - Aug. 2019 for the phrases "machine learning bias" (blue), "HR AI" (red), and "ethical hiring" (yellow), showing relative interest over time, with 100 indicating peak popularity.

2.1.1 Our Contributions

We seek to present an overview of the work in fairness in machine learning and AI-based systems, with a view toward addressing the hiring space. Previous surveys [31, 69, 70, 71, 72, 73, 74] and work in fairness in machine learning have not addressed the types of biases that past studies have shown occur in the recruitment process, such as *confirmation bias* (decisions made in the first few minutes of meeting a candidate) and *expectation anchor bias* (being influenced by only on piece of information) [67], and how these biases that are carried over to the model can be mitigated. Therefore, we present an overview of methods for measuring bias, tools available for bias mitigation, and future challenges for fairness in machine learning specific to recruitment and worker-oriented applications.

2.2 Background

Before measuring bias in algorithms, the concept of "fairness" needs to be defined to consistently compare and evaluate algorithms in AI-based recruitment. Several definitions of fairness have been proposed in the past, originating from anti-discrimination laws such as the Civil Rights Act of 1964, which prohibits unfair treatment of individuals based on their protected attributes (e.g., gender or race) [75, 76]. Furthermore, the US Equal Employment Opportunity Commission (EEOC) established a requirement with similar guidelines for employee selection procedures to ensure fair treatment of employees during the hiring process and to prevent adverse impact on individuals due to hiring decisions made [32, 75]. Adverse impact, per EEOC guidelines, is determined by applying the four-fifths or eighty percent rule, viz., a selection rate for a protected group which is less than four-fifths (or 80%) of the rate for the group with the highest rate [73, 75]. An example of adverse impact is illustrated in Table 2.1, where the selection rate for black applicants is 50% of the white applicant selection rate (which is below the 80% threshold) [75].

Applicants	Hired	Selection Rate	Percent Hired
80 White	48	48/80	60%
40 Black	12	12/40	30%

Table 2.1: Sample application selection rate.

Two key definitions of discrimination were established based on these laws: *disparate treatment*, which is an intentionally discriminatory practice targeted at individuals based on their protected attributes; and *disparate impact*, which includes practices that result in a disproportionately adverse impact on protected groups [68, 75, 76]. While these standards would seem simple to meet by simply removing protected attributes from the training data, there are still many instances where this can still lead to unfair decisions. For instance, when using natural language processing, a model can infer the gender of a candidate by name; this is known as *indirect discrimination*, or discrimination through use of features that are implicitly defined in the dataset [76]. Furthermore, depending on the application,

certain protected attributes will be needed to prevent disparate impact. For instance, the decision support tool COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), used to assess the recidivism rate of a defendant, was shown to discriminate against female defendants because the algorithms were gender-neutral, and women were shown to reoffend at lower rates than men with similar criminal histories [31]. Therefore, to define fairness for a decision-making system, the causes of bias in the data should be addressed. This is discussed next.

2.2.1 Causes of Bias

There are several ways in which human bias can be transferred to the dataset used to train a machine learning system. These are further described next [32, 73, 74].

- 1. Training data: If the training data is biased in some way, the machine learning system trained on it will learn the bias. This may be through a skewed sample, in which proportionately more records are present for a group achieving a particular outcome versus another. The dataset may also be tainted in the labeled outcomes it contains (e.g., if the dataset was manually labeled by a human and any human bias was transferred to the labels).
- 2. Label definitions: Depending on the problem being solved or the decision being made, the target label may contain a vague description of the outcome, and thus result in incorrect predictions and a larger disparate impact. For instance, in the example provided in [32], if a manager were to build a simple binary classification model to label a job candidate as a "good" hire (instead of modeling the different ways in which a candidate could be a "good" hire), many factors may get obscured by the model's prediction on the candidate. Employee motivation, person-job fit, and person-environment fit are a few factors that will typically determine how well a candidate fits an organization and how well they will perform once hired [72, 77, 78].

- 3. *Feature Selection:* Features used to improve the model may result in an unfair prediction [79]. For instance, certain features may not be relevant to the real-world application of the model, resulting in bias against protected groups. In addition, certain features may be derived from unreliable/inaccurate data, which will then result in a lower prediction accuracy for certain groups.
- 4. Proxies: Even with protected attributes removed from the dataset, they may still be found in other attributes, and still result in biased decisions being made. For instance, Amazon's hiring application was biased even without using the gender attribute, because it inferred it from the educational institution listed on the resume of applicants (e.g., all female college or all-male colleges) [65].
- 5. *Masking:* To remove any protected attributes or proxies of these attributes that may lead to disparate impact, new features may be formed to replace these attributes and achieve a new representation of the data (masked features) [32]. However, this may also lead to new biases from the features selected by the human masking the protected features.

2.2.2 Definitions of Fairness

Several definitions of fairness that can be used to assess machine learning systems have been proposed in past literature. Below we cover the five core ones as defined in [73, 74, 79].

1. Demographic parity: Also known as statistical parity, this states that acceptance rates of two groups must be equal (e.g., the percent of people accepted for a position in one group must be close to the percent accepted for another). It aligns with laws requiring fair hiring processes (e.g., the four-fifths rule), and is formulated in Equation 2.1 [74], where X represents the features of an individual, A represents the protected attributes, C is a classifier with a binary outcome Y, and C = c(X, A). Then, for groups a and b, the selection rate for both must be equal or within p percent of each other, where

$$\epsilon = \frac{p}{100} \in [0, 1].$$

$$|P_a\{C = 1\} - P_b\{C = 1\}| \le \epsilon$$
(2.1)

2. Accuracy parity: Accuracy parity is the notion of being able to provide a more balanced version of the data, in which we can trade a false positive of one group for a false negative of another, so that it fits the definition in Equation 2.2 [73]; this states we should hire (C = 1) equal proportion of candidates from the qualified ones (Y = 1) in each group (also known as equality of opportunity).

$$P_a\{C=1|Y=1\} = P_b\{C=1|Y=1\}$$
(2.2)

3. Predictive rate parity: Also known as positive predictive parity and negative predictive parity, this requires that the credentials of a candidate be consistent with the model's prediction across different groups. Specifically, both conditions represented in Equation 2.3 (positive predictive parity) and Equation 2.4 (negative predictive parity) [68, 73] for the classifier C must be satisfied to have predictive rate parity.

$$P_a\{C=1|Y=1\} = P_b\{C=1|Y=1\}$$
(2.3)

$$P_a\{C=0|Y=0\} = P_b\{C=0|Y=0\}$$
(2.4)

- 4. Individual fairness (versus group fairness): Individual fairness is different from the above fairness definitions, in that fairness is established on a person-by-person basis instead of on a group basis. It states that similar individuals should have similar outcomes. Using a more fine-grained approach for fairness, this can address limitations of group fairness definitions [80].
- 5. *Counterfactual fairness:* Counterfactual fairness is motivated by the idea of transparency, where even if an algorithm makes a decision that may seem biased, an explanation for this decision is provided. This can assist in finding bias in algorithms, by replacing attributes that may seem protected, and observing the change in decisions.

2.3 Bias in Recruitment

Understanding the recruitment process and where biases can occur will assist in employers applying methods to mitigate bias. The recruitment process consists of the following steps, as identified in [81], and shown in Figure 2.2. These are further expanded on below.



Figure 2.2: An overview of the steps in a typical recruitment pipeline: (1) identifying and attracting candidates, (2) candidate screening or processing, and (3) communication and selection of candidates.

2.3.1 Identifying and Attracting Candidates

After identifying the need for a position, and posting it, an employer may search for candidates or accept resumes sent as a result of the posting. Often, recruiters on job sites such as LinkedIn [52] and Indeed [24] will observe rankings of candidates according to jobfit/similarly, and the employer will have the resumes submitted to the posting ranked. There is much literature on selecting the best candidate from a pool of candidates to balance onboarding costs with a candidate's current experience, KSAOs (knowledge, skills, abilities, and other traits), and person-environment fit [72, 82, 83]. However, with AI implemented in several outsourced recruitment systems and applicant tracking systems (ATSs), the employer may play little to no part in selecting the list of top candidates. How an AI system will rank these candidates depends on past data used to train these ranking models. The biases present in the data may lead to invalid rankings, and can be costly for organizations who will then need to spend more time and resources for training and onboarding, or will experience a lower employee retention and job satisfaction rate [84].

2.3.2 Candidate Screening/Processing

After ranking candidates from a list, an initial pool of resumes or candidates is established, and employers start screening candidates to better assess fit. Methods for screening include resume parsing systems, job experience and knowledge assessments, personality assessments, and in-person phone calls and video streams [72, 80, 83]. As AI is further incorporated in the recruitment process, how assessments and resumes are parsed are of increasing concern since they have been subject to bias in the past, as seen in Amazon's resume parsing/ranking application [65].

2.3.3 Communication and Selection of Candidates

This step is often managed by the employer, and is the last stage of the recruitment process, wherein the employer meets the candidate, either in person, through a call, or video stream.

There are not many AI systems that replace the interview process, largely because research has shown in-person interviews have a better success rate than offline interviews [82]. However, as AI becomes more commonplace, there are several areas, such as image, audio, and video classification, where bias has been studied [31, 85], and methods for remedying such bias will need to be applied.

To assist in a more ethical hiring process, by providing a transparent and fair approach for candidates and employers, we discuss methods for mitigating bias in algorithms in each stage of the recruitment process.

2.4 Bias Detection and Mitigation

Bias mitigation in hiring tools will need to be applied to each step of the recruitment process, depending on the model used. As described in [72], the hiring process should

focus on achieving fairness, among other things, through: (1) accountability (interviewer's obligations to make reasonable decisions and address any mistakes an AI system may make), (2) responsibility (focusing on developing a fair and reliable AI system that also adapts to changes), (3) transparency (explanations of the decisions made by the AI system and employer).

2.4.1 Methods of Bias Mitigation

There are three key classes of bias mitigation algorithms/methods [86]:

- Pre-processing: This involves modifying the dataset before it is used to train the model. Algorithms such as reweighting and optimized preprocessing are used to edit the features and labels in the data according to fairness criteria before classification [86]. This can be used to remove any obvious protected attributes unrelated to the position, or modify any features that would lead to bias. This also benefits instances where the model itself cannot be modified, and can be used across recruitment tasks if multiple models are used for multiple tasks.
- 2. In-processing/optimization: In this approach, the model is optimized to meet any fairness definition described earlier through the setting of constraints on the classification objective (e.g., achieving accuracy parity by declaring we should hire an equal proportion of candidates from qualified individuals in different groups). The benefit of this method is the high performance achieved on fairness measures. However, this also requires modifying the classifier model, which may not be an option when outsourcing the recruitment process. Furthermore, this may modify the accuracy of the classifier, depending on the fairness definition used (e.g., accuracy parity conditions may result in unqualified applicants being hired, to meet equal outcomes).
- 3. *Post-processing/counterfactuals:* This is the process of meeting fairness constraints by modifying the outcome of the model, either by setting a threshold for certain classifica-

tions already made, or providing transparency in the algorithm through counterfactuals. This allows fairness definitions to be met without directly modifying the classifier. In addition, providing counterfactuals will assist not only the employer, but also the candidate in understanding their weaknesses or areas of mismatch for the job, which can then be used to improve performance in the next interview. Furthermore, clear and understandable feedback from an interview or decision has been shown to improve perceived fairness of a system [21].

2.4.2 Tools

Several toolkits and programs have been released to assist developers in embedding fairness algorithms in their machine learning pipeline [86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103]. These toolkits focus on one or more algorithms from the above three categories and are open source for anyone to use and contribute to, leading to more engagement in the machine learning and AI space. In Figure 2.3, we present an overview of user engagement in popular fairness and interpretability resources on GitHub. This is measured using the number of stars (i.e., a user marks the project as a "favorite"), forks (i.e., a user copied the project to either customize for their own needs or contribute to the original project), and the watch count (i.e., number of users which have notifications of updates on the project turned on). Below, we describe some of the larger projects in the area of model fairness and interpretability:

- 1. *AIF360 [86]:* AI Fairness 360 (AIF360) is a toolkit developed by IBM that packages bias mitigation algorithms and fairness metrics for developers to implement in their models. This toolkit, written in Python, focuses on industrial settings that would benefit from fairness algorithms, and allows for bias mitigation in all three of the stages described earlier.
- 2. FairML [89]: FairML focuses on assisting in the post-processing stage of bias mitiga-

tion, by explaining the significance of a model's features.

- 3. FairTest [88]: FairTest is a toolkit to assess the features with the highest impact on a classifier (using an unwarranted association framework). Decision trees are used to divide the dataset into subgroups and assess the strength of features on an outcome, allowing developers to check any protected attributes.
- InterpretML [91]: Though currently in alpha, this project has high engagement on GitHub. Its focus is to explain black box models and develop new, easily interpretable models.
- 5. *Lime [92]:* Lime is a toolkit which focuses on understanding black box machine learning models, by identifying features with the highest contribution to model decisions.
- 6. Lucid [87]: Lucid is a collection of tools and Jupyter notebooks to assist in neural network research, and for interpretability of networks.
- SHAP [93]: Like Lime, SHAP explains the features contributing to a model's decision, but uses game theory for user-understandable explanations and provides many visualization features of the data.
- 8. Themis-ML [90]: Like AIF360, Themis-ML provides tools for all three classes of fairness algorithms, by providing pre-processing methods, pre-built classification models that meet fairness criteria, and post-processing awareness estimators.

2.5 Conclusion and Future Work

Though several approaches for bias mitigation have been provided in toolkits and papers in the field of machine learning fairness, there are problems specific to the hiring space which are still unaddressed, and are challenges for future work in fair AI-based recruitment:

1. Tool evaluation ad selection: Organizations seeking to adopt one of the available fairness tools will need to assess which option best fits their hiring needs, since metrics



Model Fairness and Interpretability Tools vs. GitHub Engagement

Figure 2.3: Overview of user/developer engagement on fairness and interpretability repositories on GitHub, measuring the number of stars(i.e., marked as a favorite by a user), watch count (i.e., user notified of updates), and forks (i.e., user copies project code to contribute or customize).

vary highly depending on the occupation. In a past study comparing SHAP with Lime for model interpretability, results showed that no single interpretability tool achieved the best performance across different types of data [104]. However, one metric which can be used across these tools is engagement and activity of the project repository, as past studies have shown that repositories with higher engagement and project discussion lead to more awareness of the project application and better maintainability over time [105].

- 2. Job-specific requirements: Requirements often vary across occupations, and the representation of candidates and the protected attributes will also need to change accordingly. Developing/training a single model to detect protected attributes in datasets across occupations will not work for occupations that require attributes normally considered protected (e.g., information about religion when applying to be a leader in a religious institute).
- 3. Fairness in job postings: Not only does fairness apply for the candidate in the organization, but also outside as a worker/candidate searches for jobs. How job postings are recommended will vary based on a candidates' uploaded resume or past work experience listed on the job requirement site. However, these may also be subject to unintentional biases, depending on how the employer wrote the job description. Therefore, this will hinder organizations from finding the best candidates.
- 4. *Hiring decision transparency:* After applying, an applicant may not hear back from the employer, depending on the system in place that screens their profile. Often, these responses are not detailed enough to provide the candidate information on the criteria they did not meet. Therefore, counterfactual fairness will be needed for explainability.

Although several advances in AI applications for recruitment have been made and used in current recruitment systems, there has been a lack of fairness studies using these models. This paper presents an overview of the recruitment process, implications of bias in these models, and methods for bias mitigation which can be adopted. Much work has been done in model interpretability and post-processing methods for achieving fairness. However, there is scope for more work in several areas: optimization and pre-processing methods, models transparency in recruitment decisions. Identifying the cause of bias is also a crucial challenge, that would be necessary for current models. These challenges are an opportunity for organizations and machine learning researchers to transform the hiring process for applicants around the world for the better, leading to more effective workplaces and more motivated workers.

CHAPTER 3

THE OCCUPATIONAL INFORMATION NETWORK (O*NET)

In this chapter, we will discuss the key resource used in development of the AI-based applications presented in future chapters, the Occupational Information Network (O*NET) [106]. In order to fully understand the detail required in representing work, workers, and work context, a comprehensive overview of this dataset is needed. First, we will provide the motivation behind development of O*NET in Section 3.1. Then, Section 3.2 provides a high-level overview of the O*NET taxonomy, and the advantages to different key components used in future chapters. Next, Section 3.3 outlines applications and services built using O*NET, as a case study for how it can benefit the future of work and providing an overview of related work for this thesis. Last, Section 3.4 outlines the key occupations and pathways from O*NET based on earlier discussed criteria, that will be used in future chapters to evaluate the validity of our methods.

3.1 Background

With the recent advances in artificial intelligence, big data, and robotics, one of the concerns in the future of work is the changes brought about by automation, and how this may replace human workers and their tasks in their occupation. It is estimated by 2030, up to 14% of the global workforce may need to change occupational categories, as the world of work is changed by technological advances [3]. This has led to research focusing on the jobs landscape, and building tools to assist workers and employers in making informed and optimal decisions in their work transitions and career pathway planning, as further described in Section 3.2.

In 1998, the U.S. Department of Labor – Employment & Training Administration (US-DOL/ETA) collected and organized a wide variety of occupational information (e.g., job requirements, skills and education needed, tasks, etc.) to help students, job seekers, and em-
ployers in their job search and hiring needs. This database was released to the public under the name O*NET (the Occupational Information Network) [107]. Currently, this database has grown to support 1,016 occupations, and contains detailed information about the worker requirements for each occupation, which has enabled several applications, as described next.

3.2 Content Model

A high-level overview of the O*NET taxonomy (i.e., content model) can be seen in Figure 3.1, where each occupation contains the worker characteristics, requirements, experience requirements, occupational requirements, workforce characteristics, and occupation-specific information such as the title, description, and technology skills needed [108]. Each occupation is categorized with a Standard Occupational Classification (SOC) code, that provides a hierarchy of occupations that can be used in clustering similar jobs (e.g., the SOC code for "Computer Programmer" is 15-1131.00, which is of the type 15-0000, or "Computer and Mathematical Occupations") [108]. This is also known as O*NET's career cluster, with a total of 23 groups [108].



Figure 3.1: High level overview of the major components of the O*NET content model representation of an occupation.

Furthermore, O*NET provides an *importance* and *level* label for each KSAO of the occupation, where importance describes how crucial the KSAO is for being a worker in the

occupation, and level describes how competent the worker needs to be in that KSAO. These values were aggregated from surveys in each occupation, and are in the range of one to five, and zero to seven, respectively. An example of this with a sample occupation from O*NET is shown in Figure 3.2. The survey used to acquire this data was gathered by surveying different job incumbents such as workers in the occupation, and human resource professionals or experts on the structure of the occupation. This information was averaged, and the standard deviation, minimum, and maximum values are all provided in the O*NET dataset. Furthermore, if a sample size was small for a given KSAO of an occupation, they were marked for removal, and were not used in our applications to prevent inaccurate representation of workers in the occupation.

However, not all entities in the content model use these labels, such as "technology skills", or "interests", since their purpose and use in the workplace does not require them. Therefore, an overview of all data files and their purpose are presented in Table 3.1 for clarity. Next, we present key characteristics used in the O*NET model by our approach, that are not directly included in their standard content model.



Figure 3.2: The occupation "Statistical Assistant" and the corresponding KSAOs with the importance provided.

Entity	Description				
Worker Charac	teristics				
Abilities	The attributes of the worker that influence their performance.				
Occupational In-	Worker RAISEC interests, further described in Section 3.2.1.				
terests					
Work Values	Aspects of work important to a person's satisfaction based on the				
	Theory of Work Adjustment [109].				
Work Styles	Personality characteristics capturing how a worker performs in their				
	occupation.				
Worker Require	ements				
Basic Skills	Background skills and procedures needed to facilitate learning and				
	work in the occupation.				
Cross-	Fundamental skills used across activities in the occupation.				
Functional					
Skills					
Knowledge	Principles and facts used in the occupation on a daily basis.				
Education	Educational experience or certification required by the occupation.				
Experience Requirements					
Generalized	Activities that frequently occur across many occupations and in-				
Work Activities	dustries.				
Intermediate	Activities that frequently occur across occupations.				
Work Activities					
Detailed Work	Work activities specific to the occupation, or general occupational				
Activities	category.				
Organizational	Organizational characteristics that affect how people work in the				
Context	occupation.				
Work Context	Physical, structural, and interpersonal factors that influence work				
	in the occupation.				
Workforce Characteristics					
Labor Market	Occupational statistics related to the economic conditions and labor				
Information	force characteristics.				
Occupational	The predicted labor force characteristics of the occupation.				
Outlook					
Occupation-Specific Information					
Title	Occupation title.				
Description	A brief statement describing the occupation.				
Alternate Titles	Alternate titles for the occupation.				
Tasks	Common tasks performed by workers in the occupation.				
Technology	The required technology and software used in the occupation.				
Skills					
Tools	The machines or equipment needed or used in the occupation.				

Table 3.1: The worker-oriented, occupation-specific, cross-occupation, and job-orientedcharacteristics in the O*NET content model.

3.2.1 O*NET Interest Profiler

The O*NET interest profiler is a self-assessment taken by workers to explore careers and jobs they might have not otherwise considered, given their interests [110]. The tool is a web-based survey, built on top of Holland's RIASEC codes, a psychology-based model to link occupations to interests [111]. This model characterizes the vocational choice of workers based upon their personality types, which contains dimensions of realistic (R), investigative (I), artistic (A), social (S), enterprising (E), and conventional (C). Past studies evaluating RAISEC codes in recent decades have shown the model was accurately capturing a worker's interests and personality in the occupation [112]. An overview of the different dimensions and their description are presented in Table 3.2 [110].

Dimension	Description					
Realistic	Occupations with this dimension are highly involved with practi-					
	cal, hands-on work; this often requires working outdoors, without					
	working closely with others and paperwork.					
Artistic	Occupations with this dimension involve design and require sel					
	expression. Often this work can be done without following a clear					
	set of instructions or rules.					
Investigative	Occupations with this dimension involve working with ideas and					
	searching for facts, or problem solving.					
Social	Occupations with this dimension involve working with and teaching					
	others.					
Enterprising	Occupations with this dimension require workers to carry out their					
	own projects and tasks, by leading others. These occupations often					
	require decision-making skills and the ability to take risks.					
Conventional	Occupations with this dimension require individuals to follow a pre-					
	defined set of procedures or instructions. These often involve work					
	with data, rather than with ideas.					

Table 3.2: Holland's occupational themes, or the RAISEC model.

With this, O*NET also includes an *occupational interest* under the worker characteristics associated with each occupation. This provides an occupational-interest score, or a real number between zero and six, each of which represents the RIASEC category rating for workers of that occupation. This is further illustrated in Figure 3.3, which shows the point system representing the RAISEC values, and the different occupations and their worker's most highly rated interest. Furthermore, the distribution of RAISEC codes and the values assigned for each for the first occupation in each career cluster is shown in Figure 3.4.



Figure 3.3: The different RAISEC codes mapped to an occupation.



Figure 3.4: Distribution of RAISEC code values for the first occupation in each cluster.

3.2.2 Green Occupations and Tasks

Another component not directly reflected in the O*NET content model, is the green occupation and green task flags. The term "green" refers to the activity related to decreasing pollution and use of fossil fuels, increasing efficient energy use, and adoption of renewable energy sources [113]. As part of the effort to keep up with the world of work as it changes, this flag was added after an investigation of green economic activities impacting occupations and the individual tasks in them [113].

Among this category, there are three sub-categories an occupation can fall under. These include: green new and emerging occupations, or when the impact of green economic activities and technology is significant and contributed to the creation of the occupation; green enhanced skills, or when the impact of green economic activities and technology has a change in the worker requirements and work done in the occupation; and last green increased demand, or when the impact of green economic activities and technology has led to an increase in the demand for employment for this occupation [113]. Overall, there are 204 occupations that have been classified as a green occupation.

3.2.3 Bright Outlook

O*NET also identifies *bright outlook* occupations, or those that are expected to have a larger number of job openings in the near future, or in which jobs are growing rapidly [114]. Each occupation marked to have a bright outlook, meets one of the following criteria: they are projected to grow rapidly, or have a faster than average growth with an employment increase of 7% or more over the next decade; and/or projected to have a large number of openings, or 100,000 or more openings over the next decade [114]. The current bright outlook occupations were determined using the Bureau of Labor Statistics projections from 2008 to 2018 [114], which might result in outdated results near the end of the current decade, with more rapid changes in the world of work changing employment. However, these provide an initial step to the data-driven job and career search for the future of work, by allowing workers or students to see where their field is headed, and future career goals they might have.

3.2.4 Job Zones

To capture the time and prior work necessary to enter an occupation, O*NET provides a *job zone* label for each occupation, or a one to five value representing the "difficulty" to enter the occupation [115]. For instance, job zone 1 consists of occupations that require little to no preparation, while job zone 5 consists of occupations that require over 4 years of preparation. These values were based on the vocational education (i.e., high school, college training, etc.), apprenticeship training, in-plant training (i.e., a class provided by an employer), on-the-job training (i.e., work under a qualified worker on the job), and necessary experience from other jobs (i.e., entry-level jobs necessary before working jobs that require more experience). This allows workers who are looking to drastically change careers, to observe a starting point and the occupations that they might be better suited for, without experience, on-the-job training, or education in the relevant area.

3.2.5 Career Changers and Beginners Matrix

For each occupation, O*NET contains SOC codes for related occupations, known as the *Career Changers Matrix* (i.e., matrix of occupations which overlap in multiple skills and experience, making it easier for workers to transition between these occupations), and the *Career Beginners Matrix* (i.e., matrix of occupations which overlap in general capabilities and interests, that better fit career explorers) [116].

These were developed using a method termed the *Related Occupations Matrix* (ROM) algorithm [116]. This algorithm uses nine entities from the O*NET content model, based on a balance of job-oriented, worker-oriented, cross-occupation, and occupation-specific features to link to job transfer for workers; these are knowledge, skills, abilities, interests, work styles, work values, generalized work activities, work context, and job zone. These features and their normalized value (e.g., importance for knowledge) are used to calculate the distance between

each occupation and form the similarity matrix for an occupation. The distance measure used is shown in Equation 3.1, where D_{raw} is the final distance between two occupations, X represents the active or target occupation, Y represents the compared occupation, and a, b, and z, represent each entity for the feature they are calculating the distance of [116]. Next, the total distance between two occupations is determined using Equation 3.2 for the starter matrix, and Equation 3.3 for the change matrix, which are the sum of distances of the feature vectors, multiplied by a weight of Z [116].

$$D_{raw} = \sqrt{(X_a - Y_a)^2 + (X_b - Y_b)^2 + \dots + (X_z - Y_z)^2}$$
(3.1)

$$RO_{Starter} = Z(D_{Abilities}) + Z(D_{Interests}) + Z(D_{Styles}) + Z(D_{WorkValues})$$
(3.2)

$$RO_{Change} = 1.3 * Z(D_{JobZone}) + Z(D_{Knowledge}) + Z(D_{Skills}) + Z(D_{GWAs}) + Z(D_{Context})$$

$$(3.3)$$

Though these matrices only look at the overlap of KSAOs between occupations, they are an initial step towards better career pathway planning and search for the future of work.

3.3 Applications

Over the past few decades, many changes and developments in connection to O*NET have been made to assist in career planning and job search. O*NET supplies an application programming interface (API) to use their occupation content model/database, which has enabled a number of web applications building upon it. With this data widely available, several opportunities are present for data-enabled career planning and job search applications. Past studies have shown person-job fit, or alignment of a worker's skills, knowledge, abilities, and other work-related attributes with their occupation contributes to higher job satisfaction [36]. Therefore, it is crucial to have a fine-grained representation that captures each aspect of the worker and each occupation, for job seekers to find the occupation that bests suites them. Among the career planning and job search applications, several have used O*NET and been deployed to the public to search for jobs. These include services such as CareerOneStop (a large-scale career resources and job search database), MySkills MyFuture (a tool to assist in career transitions by looking at new occupations though transferability of a worker's skills), and CareerScope (enhances career planning with transitions, training, and employee retention planning) [117]. Tools and services built on O*NET can be found at their "Products at Work" resource page, which contains over 163 stories on how O*NET information was used to help in worker training, job transiting, and job search. We use this information to create a new approach to career planning and assisting workers and organizations, as described next.

3.4 Highlighting Critical Occupations and Career Paths

To evaluate our methods in future chapters, we outline key occupations and transitions that may occur in the real world with changes in the job market. These are presented in Table 3.3, and reflect the occupations with the highest predicted growth rate, green occupations from O*NET, and occupations with the fastest employment decline rate as determined from the Bureau of Labor Statistics [118, 119, 120]. Occupations were selected such that there was the largest variety of career clusters represented in each category.

Bright Outlook	Poor Outlook	Green	
Administrative Services	Cashiers, 41-2011.00	Green Marketers, 11-	
Managers, 11-3011.00		2011.01	
Financial Examiners, 13-	Telemarketers, 41-9041.00	Financial Analysts, 13-	
2061.00		2051.00	
Software Developers, Appli-	Mine Shuttle Car Opera-	Mechanical Engineers, 17-	
cations, 15-1132.00	tors, 53-7111.00 2141.00		
Environmental Engineering	Telephone Operators, 43-	Robotics Engineers, 17-	
Technicians, 17-3025.00	2021.00	2199.08	
Economists, 19-3011.00	Postmasters, 11-9131.00	Industrial Ecologists, 19-	
		2041.03	
Engineering Teachers, Post-	Postal Service Clerks, 43-	Climate Change Analysts,	
secondary, 25-1032.00	5051.00	19-2041.01	
Coaches and Scouts, 27-	Locomotive Firers, 53-	Environmental Economists,	
2022.00	4012.00	19-3011.01	
Physician Assistants, 29-	Word Processors and Typ-	Sales Representatives,	
1071.00	ists, 43-9022.00 Wholesale and Manufactur		
		ing, Technical and Scientific	
		Products, 41-4011.00	
Medical Assistants, 31-	Watch Repairers, 49-	Forest and Conservation	
9092.00	9064.00	Workers, 45-4011.00	
Forest Fire Inspectors and	Parking Enforcement Work-	Electricians, 47-2111.00	
Prevention Specialists, 33-	ers, 33-3041.00		
2022.00			
Real Estate Brokers, 41-	Legal Secretaries, 43-	Power Plant Operators, 51-	
9021.00	6012.00	8013.00	
Wind Turbine Service Tech-	Computer Operators, 43-	First-Line Supervisors of	
nicians, 49-9081.00	9011.00	Helpers, Laborers, and Ma-	
		terial Movers, Hand, 53-	
		1021.00	

Table 3.3: Occupations marked with bright outlook, poor outlook, or green, used for evaluation.

CHAPTER 4

TRANSFERABILITY OF WORKER COMPETENCIES

To assist in development of more fair and accurate systems, occupations need to have a rich representation for understanding the requirements for screening. Although past research has focused on matching individuals to various aspects of their broad work environments (e.g., their job or team) [121], the increased number of human-machine interactions in the future of work will require examining narrower aspects of jobs. In order to effectively match individuals to occupations, it is first necessary to provide a clear definition of the required competencies of those occupations both now and in the future. Furthermore, understanding how KSAOs in one occupation can transfer to another, is important to capture in work transitions and career planning. Therefore, in this chapter, we present an approach for skill transferability, and extend the O*NET dataset with links between each KSAO providing transferability of the KSAO from one occupation to another.

The remainder of this chapter is organized as follows. Section 4.1 covers the motivation behind this task and related work, and Section 4.2 goes over our contributions to this area. In Section 4.3, we present our approach for learning KSAO transferability between occupations, and the various approaches tested for this task. Next, Section 4.4 goes over results of our method, and evaluates how realistic the transferability scores are. Last, Section 4.5 provides concluding remarks and challenges to be addressed in future work.

4.1 Motivation and Problem Statement

Although past research has focused on matching individuals to various aspects of their broad work environments (e.g., their job or team) [121], the increased number of humanmachine interactions in the future of work will require examining narrower aspects of jobs. In order to effectively match individuals to occupations, it is first necessary to provide a clear definition of the requirements of these occupations both now and in the future. The U.S. Department of Labor - Employment & Training Administration (USDOL/ETA) collects and organizes occupational information via O*NET Program, which maintains the O*NET database. This data is crucial to "understanding the rapidly changing nature of work and how it impacts the workforce and U.S. economy" [122]. The O*NET program uses this database to develop applications that contribute to the development and maintenance of a skilled workforce. The O*NET database covers over 1000 occupations and is periodically updated to reflect changes in the nature of work [107]. Currently, O*NET organizes occupations as a taxonomy, as discussed in Chapter 3, where each occupation contains worker characteristics, worker requirements, experience requirements, occupational requirements, workforce characteristics, and occupation-specific information, such as the title, description, and technology skills needed [107]. This O*NET content model was developed using research on job and organizational analysis, and the descriptions for each occupation were gathered through yearly employee and organization level surveys [107]. However, though O^{*}NET captures the knowledge, skills, abilities, and other characteristics needed to categorize a job, it is limited in its usability for hiring and person-job fit. This stems from limitations it has in capturing individual-specific information for hiring, and the change in skills over time as the nature of work changes [121, 123, 124]. A more fine-grained representation of worker characteristics and job, occupation, and cross-occupation characteristics suitable for processing by datadriven AI systems is needed; this view includes not only occupation, job, and task levels, but between KSAOs.

4.2 Contributions

A substantial amount of research has shown that individuals with certain characteristics will be more successful in some jobs over others [33, 34, 35]. Therefore, it is important to have a representation that captures not only occupational data, but how KSAOs from one job transfer to another. O*NET provides information on skills needed for a given occupation, and the importance of the skills for that occupation (i.e., skill match), which can be used to look at skill overlap. However, in O*NET, there is no connection between different types of skills (or KSAOs) to indicate how a worker who held a previous occupation could easily learn or transfer a skill to another occupation. For instance, past research has indicated many STEM based majors will acquire skills that transfer between occupations in their major [125]. Therefore, we present a method to link *competencies* (i.e., knowledge, skills abilities) in the O*NET content model and provide a better semantic representation of workers in occupations. We focus on competencies, or KSAs, rather than the entire content model, because past research has looked at transferability of competencies in worker training [126], rather than look at values, styles, and other seemingly non-transferability attributes. We also capture additional career pathways that could be taken by individuals in different occupations, which is used in our next method in Chapter 5. Our approach, experiments, and results are further described next.

4.3 Methods

The overall approach used is presented in Figure 4.1. We focus on knowledge, skills, and abilities (i.e., competencies) of each occupation in O*NET, that are the key worker characteristics and requirement considered when transferring between occupations [108]. Each of the competencies contain a description and title, gathered from a survey of job incumbents and workers in given occupations, and researchers in jobs and organizational analysis [108]. This will be used to calculate the similarity (i.e., transferability) between competencies.

To calculate similarity between two texts, there have been several past natural language processing (NLP) approaches developed. A large portion use knowledge-based methods [127, 128, 129] to investigate similarity at a word level. However, individual word-level similarity does not capture the entire context of sentences or documents. Therefore, more recent approaches have used pre-trained word embeddings to go beyond a word level representation, and instead relate the similarity of each word based on a large corpus of text [130, 131]. Furthermore, approaches using neural networks, specifically recurrent neural networks (RNNs), bidirectional long short-term memory (BiLSTM) variants of RNNs, and pre-trained language representations, have greatly advanced NLP applications in the past year [130, 131, 132, 133].



Figure 4.1: The overall process to get the transferability between two competencies in O*NET, using pre-trained embeddings and a knowledge-based approach.

We investigate two different approaches to calculating similarity between KSA descriptors, as described in [134]: (1) a pre-trained approach using context independent word embeddings from GloVe [1], and (2) a pre-trained approach using context dependent word embeddings from BERT [135]. These approaches and the models used in each are further described next.

4.3.1 GloVe

The first method used to connect competencies, is the use of the pre-trained embedding model GloVe. GloVe (Global Vectors) is an unsupervised log-bilinear model for representing words as vectors [1]. Prior natural language processing approaches have numerically represented words with the continuous bag of words (CBOW) approach, that transforms each word (or token) into a feature vector by learning from the co-occurrence of words in the corpus. However, this approach did not consider the semantics of each word, and often led to sparse representations in a corpus with a large vocabulary. Though GloVe also considers word co-occurrence, it does so on a global basis, rather than individual documents, allowing it

represent words according to their semantics [1]. Furthermore, GloVe is advantageous to other pre-trained vector representations such as Word2Vec, that only consider similar words in a local context (i.e., the same sentence), rather than a corpus of documents [1].

Probability & Ratio	k = solid	k = gas	k = water	k = fashion
P(k ice)	$1.9x10^{-4}$	$6.6x10^{-5}$	$3.0x10^{-3}$	$1.7x10^{-5}$
P(k steam)	$2.2x10^{-5}$	$7.8x10^{-4}$	$2.2x10^{-3}$	$1.9x10^{-4}$
P(k ice)/P(k steam)	8.9	$8.5x10^{-2}$	1.36	0.96

Table 4.1: Example of the GloVe model training method using global word co-occurrence probabilities, adopted from [1].

Using Wikipedia and Common Crawl data with billions of tokens, the GloVe model observes the ratios of co-occurrence probabilities of words to encode their semantics in the vector representation. For instance, an example is shown in Table 4.1 [1], in which $P(j|i) = X_{ij}/X_i$ is the probability of a word j co-occurring with word i, and X_{ij} is an entry in the co-occurrence count matrix of j and i from the corpus, and $X_i = \sum_k X_{ik}$ is the occurrence of any word appearing with i [1]. Here, the words "ice" and "steam" are considered out of the word corpus with i = ice and j = steam. The term "ice" has a higher co-occurrence with "solid" than with "gas", which the term "steam" has a higher co-occurrence with [1]. Furthermore, unrelated words such as "fashion", have low co-occurrence with both. Overall the GloVe model results in word vectors whose euclidean distance capture the juxtaposition of the two words, which is further illustrated in Figure 4.2, where the comparative and superlative adjectives are shown [2].

Our approach takes the O*NET competencies, and encodes the descriptions as vectors using GloVe. Then the similarity between the two vectors is calculated using the cosine similarity, as shown in Equation 4.1,

$$S = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$
(4.1)

Where A_i and B_i are competencies from the set $(A_1, A_2, A_3, \ldots, A_{n-1}, A_n)$ of O*NET competencies describing a worker. The reason behind using cosine similarity, over euclidean



Figure 4.2: Example GloVe vector distances, adopted from [2].

distance, is even though two vectors may be far in distance, they may be oriented closer together, making the cosine similarity a more accurate similarity measure. This similarity score represents the transferability between each competency.

4.3.2 BERT

Our second approach, uses Bidirectional Encoder Representations from Transformers (BERT), a pre-trained bidirectional representation of text which has led to many advances and improvements in natural language inference and understanding models [135]. BERT's success can be attributed to its bidirectional training with a Transformer model. A Transformer is an encoder-decoder deep learning model, that uses self-attention (i.e., concentrating on other words in the input sentence as it encodes one) to capture the context of each word in the training phase [136]. The original transformer was proposed as a directional model, or one that trains on text sequentially (i.e., left-to-right or right-to-left) by encoding one word at a time in the sentence. However, BERT users bidirectional training, considering both left-to-right and right-to-left context, to allow the model to properly learn the context of each word and its surrounding [135].

BERT's architecture is composed of multi-layer bidirectional Transformers, with 12 layers in the BASE model, and 24 layers in the LARGE model [135]. BERT was evaluated on several benchmark natural language processing tasks, such as GLUE (General Language Understanding Evaluation) [137], SQuAD (Stanford Question Answering Dataset) [138], and SWAG (Situations With Adversarial Generations) [139], and outperformed baseline OpenAI GPT and BiLSTM models on each [135].

With the advantages BERT provides, we use the pre-trained models provided to extract contextual word vectors for each O*NET descriptor. The distance (or similarity) between these vectors is calculated using the cosine similarity measure, described in Equation 4.1 [134].

4.4 Results and Evaluation

To determine competency transferability, we experiment with both the BERT-based approach and GloVe-based approach. Though BERT has shown to provide improvements in many natural language processing tasks, it only captures the local context of words and GloVe may provide an advantage in being trained on several documents for a global representation.

For the GloVe-based method, we use the 300 dimension vector model trained using 2014 Wikipedia and Gigaword 5 dumps, as provided by spaCy 2.0 [140]. For the BERT-based method, we use the uncased BERT base model, and extract representations from layer 11, which has shown to provide better results than the last layer (i.e., 12), since the last layer is more tailored to individual tasks [141].

We used Python 3.7 and the library spaCy 2.0 [140] to parse the O*NET descriptions into tokens and remove stop-words. O*NET version 24.0 was used to reference the KSAs, and the data files for "knowledge", "skills", "abilities", and "content model reference" were used. The results achieved are further described next.

4.4.1 Knowledge Transferability

In Figure 4.3 we report the transferability results of the knowledge elements from O*NET with the GloVe-based approach, and in Figure 4.4 we report the transferability results of the knowledge elements with the BERT-based approach. A 1.0 indicates elements that are highly similar, and anything lower indicates dissimilarity. We observe several differences between the GloVe-based method for embedding the descriptions, and the BERT-based method. We can see the BERT-based method has a much lower variability in similarity scores (ranging from 0.5 to 1.0), than the GloVe based method (ranging from 0.4 to 1.0). This may be due to the global context GloVe provides, over BERT that only considers local context.



Figure 4.3: The transferability of knowledge from O*NET using the GloVe-based approach.

We can see in the GloVe-based method, the similarity, or transferability, between "Engineering and Technology" and "Design" is 0.9, indicating high transferability, while "Engineering and Technology" and "Foreign Language" has a 0.7 score, indicating low transferability.



Figure 4.4: The transferability of knowledge from O*NET using the BERT-based approach.

With the BERT-based model, we observe a lower transferability score overall throughout the different knowledge elements. For instance, even though it would seem "Mathematics" and "Engineering and Technology" would be highly related, they are given a 0.6 for transferability. Furthermore, the knowledge element "Mathematics" contains a lower transferability score for all other elements, which is not shown in the GloVe-based model. For our future methods, we plan to use the GloVe-based approach results for transferability of knowledge.



4.4.2 Skills Transferability

Figure 4.5: The transferability of basic skills from O*NET using the GloVe-based approach.

For assessing transferability of skills, we look at the two categories of skills O*NET provides, *basic skills* (i.e., skills that can facilitate learning across occupations), and *cross-functional skills* (i.e., skills that can improve performance across occupations). The results for the GloVe-based approach for basic skills is shown in Figure 4.5, and cross-functional skills in Figure 4.8. The results for the BERT-based approach for basic skills is shown in Figure 4.6, and cross-functional skills in Figure 4.7.

For basic skills, we can see that the GloVe model provides more realistic results than the BERT-based model. Skills such as "Mathematics" are closer to skills such as "Science" than



Figure 4.6: The transferability of basic skills from O*NET using the BERT-based approach.

"Reading Comprehension". However, in the BERT-based approach, we can see that though "Mathematics" also received a similar score of 0.9 transferability, it has a higher score for other skills like "Reading Comprehension", that received a 0.77.

Furthermore, skills such as "Speaking" are shown to be highly transferable with the GloVe-based model in other skills like "Writing" and "Active Listening". However, with the BERT-based model, this received a much lower transferability score, and instead was closer to the transferability between "Mathematics" and "Speaking". Overall, basic skills are more general and are meant to be used across different occupations to determine how an individual can acquire new knowledge [108], so having a high transferability between these makes sense with their real-world application.



Figure 4.7: The transferability of cross-functional skills from O*NET using the BERT-based approach.

However, though cross-functional skills are also meant to be applied to a variety of domains, their focus is in performance, which will vary based on occupation. Looking at the results for cross-functional skills, we can see a wider distribution of transferability scores for the GloVe-based approach, than the BERT-based approach, that has most of the scores ranging between 0.7 and 0.85.

With the BERT-based approach, we can see the skills "Time Management" and "Instructing" have low transferability scores with all other skills. However, the GloVe-based approach shows different results, and those that are more realistic. It shows "Time Management" and skills such as "Management of Financial Resources' and "Management of Personnel Resources" to be highly transferable, with a score of 0.8. Furthermore, with the GloVe-based



Figure 4.8: The transferability of cross-functional skills from O*NET using the GloVe-based approach.

approach, skills that would not be transferable in a real-world application, such as "Installation" and "Persuasion" have a low transferability of 0.5. However, with the BERT-based approach, this received a 0.7 score, the same as "Programming" and "Technology Design". For our future methods, we plan to use the GloVe-based approach results for transferability of basic and cross-functional skills.

4.4.3 Abilities Transferability

The results for the GloVe-based approach for transferability of abilities in O*NET is presented in Figure 4.9, and the results for the BERT-based approach for abilities is presented in Figure 4.10. Unlike skills and knowledge, that were worker requirements for occupations in



Figure 4.9: The transferability of abilities from O*NET using the GloVe-based approach.

O*NET, abilities are a type of worker characteristic, meaning attributes that will influence the performance of a worker [108]. The transferability between abilities, will modify how worker performance is predicted when hiring.

From the results, we can see a worker with the ability "Written Comprehension" can transfer this to other abilities such as "Oral Comprehension", "Speech Recognition", and "Speech Clarity". The GloVe-based approach shows a more prominent similarity between these elements than the BERT-based approach. Though the BERT-based approach reflects similar similarities/dissimilarities, there is less variation in the scores assigned between abilities.

4.5 Conclusion and Future Work

In this chapter, we linked knowledge, skills, and abilities from the O*NET dataset using natural language processing and their descriptors to determine transferability between them.



Figure 4.10: The transferability of abilities from O*NET using the BERT-based approach.

For our future methods, we plan to use the GloVe-based approach results for transferability of competencies. From the results, we can see that the GloVe-based approach provided more realistic results that better fit transferability of competencies in real-world scenarios. However, this may be in part due to the need of fine-tuning BERT to our domain, which with additional data could be done. Furthermore, the sentence level representation used was the mean of all tokens in the sentence, which could blur the representation for both approaches.

For future work in this problem, we seek to use job descriptions to verify these results and evaluate the accuracy of our approach. By looking at overlap between competencies described in job postings, the transferability of skills can be assessed between jobs, or occupations. We could also improve the approach itself, by fine-tuning BERT, and using a different sentence representation to capture the variance of tokens in the text. Furthermore, a knowledgebased approach using ConceptNet [142] or WordNet [128] will be considered, along with an approach combining BERT or GloVe with these knowledge bases for better global context, and overall accuracy.

CHAPTER 5

OPTIMIZATION OF CAREER PATHWAYS

As the world of work undergoes changes caused by automation and artificial intelligence, workers and employers will need to make informed and optimal decisions regarding work transitions and career plans. Past databases and career transfer algorithms (such as ROM presented in Chapter 3), do not account for the intricate details found in a career pathway and the future of work. Furthermore, transferability of competencies has not been considered in the past. The goal of the research presented in this chapter, is to assist workers in their career pathway planning and job transitions using the O*NET database. We create a genetic algorithm to search for the optimal career pathway for a given worker and their desired career outcome using O^{*}NET. A career pathway is defined as the route from one occupation to another, wherein past approaches have optimized for career pathway planning using OPNs and looked at future payoff, company influence, or likelihood of job transition, as described in Section 1.3.2. However, for this thesis, we search for the optimal career path based on criteria of person-job fit, a measurement used in industrial-organizational psychology to determine an individual's fit in an occupation, which is determined using the individual's knowledge, skills, abilities, and other occupation-related attributes. This will assist in workers achieving a higher job satisfaction and employee retention for the organization they join. Overall, the genetic algorithm provides an initial step towards the future of career planning and assistance for job seekers in the quickly changing world of work.

The rest of this chapter is organized as follows. Section 5.1 defines the motivation behind this problem and the scope. Section 5.2 presents an overview of our contributions. Next, Section 5.3 presents an overview of our approach to solving this problem and the structure of the genetic algorithm. Then, Section 5.4 presents our results and discussion of these results from running the genetic algorithm against real-world scenarios, and evaluates our approach by comparing to common graph traversal methods. Last, Section 5.5, discusses future work and challenges to be addressed with this approach.

5.1 Motivation and Problem Statement

As the jobs landscape is expected to undergo rapid transformations, it requires workers and employers to be able to make informed and optimal decisions regarding work transition and career pathway planning. However, there is a lack of a tool for individuals to anticipate and plan for their career and future transitions between occupations. Despite a large number of studies on career adaptability and related constructs [143, 144], these studies almost exclusively focus on individuals' motivation and tendency to adapt. For individuals interested in transitioning from an occupation to another, assuming they qualify for the KSAOs to get hired, they still lack the information on the cost and benefit of such a transition. Cost, broadly defined, can entail the time individuals have to spend on learning activities, the expenses related to learning, the additional amount of practice to achieve proficiency, and the possible need for certification. Benefit includes occupational income and prestige. Conducting a cost-benefit analysis prior to embarking on career transition can enable individuals to make an informed decision in their career plans.

Past approaches to career pathway optimization have looked at using future payoff [53], company influence [54], or likelihood of job transitions [55, 56]. However, these approaches are limited in their ability to properly capture the cost of transition from one occupation to another relative to each individual worker. A worker can be characterized in terms of their knowledge, skills, abilities, and other occupation-related attributes, which can be used to assess their fit for an occupation. This is also known as person-job fit, a measurement used in industrial-organizational psychology based on the overlap of a worker's KSAOs and the occupation KSAOs [36]. The alignment of KSAOs, or better person-job fit, has shown to lead to a higher job satisfaction and employee retention rate for the organization they are hired into [36]. Therefore, it is crucial to use this information when looking at a worker's career pathway and optimizing for the best sequence of occupations. Furthermore, not only is person-job fit a factor in assessing a better career transition, but accounting for the changing world of work is necessary. Certain occupations may have a lower demand in the future, making it an inadequate choice for the user in the future. Furthermore, with changes in environmental protection and the push towards energy efficiency, some occupations will be lower in job demand, and new occupations may arise that have a brighter outlook.

Currently, there is a lack of a tool for workers to plan their career pathway considering all these elements. Therefore, we propose an approach to search for the optimal career pathway, while considering the several factors needed to accurately provide the best options for the worker. Our contributions and approach to solving this problem are described next.

5.2 Contributions

In this work, we seek to provide a worker-based solution for career pathway search and planning, by using the criteria of person-job fit and occupation outlook. The approach to solving this problem consists of a genetic algorithm (GA) to search for the best career pathway for the individual worker, where an optimal path is characterized as the lowest cost path. The representation used for each individual solution captures the worker's KSAOs, the transferability of competencies as determined in Chapter 4, and a wide range of occupations using O*NET. The details of our method are further described next.

5.3 Method

We present a genetic algorithm to search for the optimal career pathway for an end-user given their KSAOs. An overview of the structure and benefits of genetic algorithms for optimization problems, and its use in this Chapter is provided next.

5.3.1 Genetic Algorithms

Genetic algorithms are a metaheuristic optimization or search method inspired by Darwin's theory of evolution [145]. First introduced in 1960 by John Holland, genetic algorithms provide the benefits of evolution and natural selection, wherein a population of solutions is generated, and go through generations of selection and crossover of their genotypes (i.e., traits describing the solution). Natural selection contributes in this method, with more fit individuals (i.e., solutions closer to the optimal solution) in the population surviving into the next generation [145].

There are five phases in a genetic algorithm: population initialization, fitness evaluation, selection, crossover, and mutation. Population initialization is the creation of an initial set of solutions for the problem to solve, characterized by a set of parameters, known as a gene. Next, fitness evaluation consists of determining how fit each individual in the population is using a pre-defined fitness function (i.e., objective function). Then, each genetic operator is run including selection (i.e., selecting individuals from the population based on their fitness), crossover (i.e., recombination of genes to create offspring for the next generation), and mutation (i.e., probability-based change of one or more genes in an individual). These steps are repeated until this stop criteria is reached, or when the population converges to a fit solution. We use a genetic algorithm for this problem to efficiently search a large solution and as seen in previous performance of a GA in graph-based problems [146, 147].

The overall workflow for the GA is shown in Figure 5.1, where the input user KSAOs would typically be extracted from a resume or assessment data, and the output is the most fit career pathway for their desired career outcome (i.e., the pathway with the shortest distance, as further described below). In creating this GA, custom operators were created to mimic a realistic career pathway traversal and the cost of transitioning between occupations. This allows the end-user to achieve the best results that can be applied in the real world.

The first path of the career planning GA is the creation of the genotype using O*NET, that has been used to create a large graph of occupations. Then, an initial population



Figure 5.1: The overall architecture of the career pathway planning system, and the GA (wrapped in the box). The inputs are the end-user's KSAOs (typically extracted from a resume through natural language processing), the desired occupation from O*NET, and the occupation space provided by O*NET. The output is the most fit solution from O*NET, or the best fitting career pathway for the end-user.

of individual solutions (i.e., career pathways) is generated for the GA. The population is evaluated using a unique fitness function which closely models a realistic career pathway. Next, the selection and crossover operators are applied to create the next generation. Two mutation operations also occur, and are specific to the career pathway representation used. Finally, the GA will run until the stop condition is reached, which is by number of generations until there is no further increase in fitness. Each of these is further described next.

5.3.2 Path Representation

Each individual solution (i.e., career pathway), is represented as a sequence of occupations traversed through the occupation graph, that was created from the O*NET 24.0 content model. The user starts at a given occupation based on their KSAOs, and a desired occupation is provided and marks the end of the career path. Each occupation and KSAO is encoded as an integer, or unique identifier, in the range of occupations and KSAOs from O*NET.

One key difference between this career pathway GA and typical graph traversal problems, is accounting for the time spent in each node in the graph. A worker will typically spend anywhere from months to years working in a job, and their knowledge, skills, abilities, and other worker characteristics will change in that time. Therefore, this was a crucial part for the pathway representation.

The representation for the *nth* occupation O_n from O*NET is given by the following tuple,

$$O_n = (S_n, C_n, R_n, E_n, B_n, Z_n)$$
 (5.1)

where $\{n \in \mathbb{Z} : 1 \le n \le 974\}$, S_n is the SOC encoded career cluster number, C_n is a list of the competencies associated with the occupation, R_n is a list of the occupational requirements associated with the occupation, E_n and B_n are the green and bright occupation flags, respectively, and Z_n is the job zone number.

Competencies C_n for an occupation O_n is represented as a set, where each competency is either a knowledge, skill, or ability from the O*NET content model, as represented in Equation 5.2 below,

$$C_n = \{c_1, c_2, \dots, c_k\}$$
(5.2)

where c_k is an individual competency and its level associated with the occupation, and $\{k \in \mathbb{Z} : 1 \le k \le 119\}.$

Occupational requirements R_n for an occupation O_n is represented as a list where each requirement is either an interest, work style, or work value from the O*NET content model, as represented in Equation 5.3,

$$R_n = \{r_1, r_2, \dots, r_k\}$$
(5.3)

where r_k is an individual requirement and its importance associated with the occupation, and $\{k \in \mathbb{Z} : 1 \le k \le 42\}$. The reason for separating the competencies and occupational requirements, though together in the O*NET content model, is to account for their differences in calculating occupational distance in the fitness function, as later described in Section 5.3.3. Next, the green occupation flag E_n from O*NET is used to indicate if an occupation will be changed by the green economy [113], and is represented in the solution as,

$$E_n \to \{0, 1\} \tag{5.4}$$

where 1 indicates the flag as present, and 0 is missing. Similarly, the bright outlook flag O_n from O*NET is used to indicate if an occupation is expected to have a large number of job openings in the near future [114], and is represented as,

$$B_n \to \{0, 1\}\tag{5.5}$$

where 1 indicates the flag as present, and 0 is missing. These can be used by the worker to observe occupations in their final path that may be better to apply to in the future.

Furthermore, the occupation representation contains the job zone category Z_n to represent the amount of experience and time required to enter the occupation, that is represented by,

$$Z_n \to \{1, 2, 3, 4, 5\} \tag{5.6}$$

where job zone 1 indicates little to no preparation required, and job zone 5 indicates extensive preparation of four or more years. This is used in the path when calculating the distance between two occupations.

Each individual solution in the population is then represented as a sequence of occupationtime pairs as each gene G, given by the tuple,

$$G_i = (O_{ni}, t_i) \tag{5.7}$$

where t_i is the time (in months) the worker stayed in the occupation, and $\{t \in \mathbb{Z} : 1 \le t \le 60\}$. Then the entire path P is given by,

$$P = (G_1, G_2, \dots, G_x)$$
(5.8)

wherein x is the number of occupations in the path where $\{x \in \mathbb{Z} : 1 \le x \le 13\}$, and 13 is the rounded average number of jobs an individual will hold in their lifetime [148].

5.3.3 Fitness Function

The fitness of each individual was determined using the total distance of the career pathway. The overall distance is the sum of the costs from one occupation and time to another. Therefore, a more fit individual would be a lower costing path (i.e., shorter distance). However, unlike typical graph traversal problems, transitioning from one job to another is not the same for all workers. There are several factors that need to be considered such as past work experience, skills, time spent in past work, and their future goals. Therefore, there are several factors that influence how the fitness function was designed for this problem, as described next.

5.3.3.1 Occupation Distance

The distance between occupations is calculated using the euclidean distance between the end-user's current state and the occupation. As the end-user traverses through the path, they acquire the competencies needed for that occupation, as further described in Section 5.3.3.3. This allows the distance between each occupation to be custom to each end-user and their career. Furthermore, competencies and requirements that are of higher importance for the occupation, and those which the user has a higher level for (i.e., more experience), are accounted for in the distance calculation.

Consider a worker W_x in state x of the pathway,

$$W_x = (S_x, C_x, R_x, Z_x) \tag{5.9}$$

where S_x is the SOC encoded career cluster number of their last occupation, C_x is a list of the competencies associated with the worker, R_x is a list of the worker's requirements, and Z_x is the job zone number of their last occupation. The distance between the worker's profile and the occupation O_x is given by $D(W_x, O_x)$, where the distance D is defined as the following function,

$$D(q,p) = w_Z(Z_p - J(q,p)) + w_C \sqrt{\sum_{k=1}^{||C_p||} (q_{ck} - p_{ck})^2 + w_R} \sqrt{\sum_{k=1}^{||R_p||} (q_{rk} - p_{rk})^2}$$
(5.10)

where a weight w_Z is given to the job zone difference between q and p, and a weight of w_C is given to the distance for competencies and w_R for requirements. The value of the weights for requirements and competencies need to be of equal size to properly consider the worker's characteristics and the required competencies for being hired. The value for the job zone factor w_Z was determined looking at the distribution of the occupation distances. Figure 5.2 shows the distribution of occupation to occupation distances using only competencies (i.e., $w_R = 0, w_Z = 0, w_C = 1$), and Figure 5.3 shows the distribution of occupation to occupation distances using only requirements (i.e., $w_C = 0, w_Z = 0, w_R = 1$). We can see a normal distribution for both competencies and requirements, where a maximum value of 400 and 300 are found, respectively. Therefore, for the weight for job zone, we set $W_Z = 50$, where $w_C = .25$, and $w_R = .25$. This will allow the result of the job zone function to be properly accounted for in the overall distance function (i.e., objective function). The function J(q, p)used to represent the job zone of the worker, is further described next.

5.3.3.2 Job Zone and Career Clusters

To properly represent the time and experience needed to enter an occupation, we use the job zone and career cluster when calculating distance between two occupations, or the worker transitioning from an occupation to another. When a worker transitions to a new field (i.e., career cluster), they should enter in an entry-level position (i.e., job zone 1). Therefore, the job zone is "reset" if the career cluster number changes when transitioning between two occupations.

The equation to assign job zone J is given by,

$$J(q,p) = \begin{cases} Z_q & ||S_q - S_p|| \le 0\\ 1 & ||S_q - S_p|| > 0 \end{cases}$$
(5.11)



Figure 5.2: Distribution of the distances between every occupation to occupation pair in O*NET, using only competencies to calculate the distances (i.e., $w_C = 0$).



Distribution of Requirement Distances Between Occupations

Figure 5.3: Distribution of the distances between every occupation to occupation pair in O*NET, using only requirements to calculate the distances (i.e., $w_R = 0$).

where Z_q is the job zone of the input user q and is the career cluster code for q is different than the career cluster code for p, the job zone is reset to 1, otherwise the job zone Z_q , or the job zone of the last occupation the worker was in, is used. With this value, we can see
the overall distribution of distances with the weights defined earlier and the result of the job zone function in Figure 5.4. We can see the change in occupation cost from occupations in different career clusters being in a majority of higher-costing occupation to occupation distances, than occupations in the same cluster.



Distribution of Distances Between Occupations

Figure 5.4: Distribution of the distances between every occupation to occupation pair in O*NET, using weights $W_R = 0.25$, $W_C = 0.25$, and $W_Z = 50$.

5.3.3.3 Competency Growth and Decay

In addition to job zone, we also consider how competencies of a worker will change as they transition between occupations. Depending on how long a worker stays in a job, the cost of the overall path should change. At the beginning of the career path, the user has a set of competencies and their level. Next, the user will transition to an occupation in the path, and stay in that occupation for a period of time specified in the representation. This will in turn modify the end-user's competencies for the next occupation hop. Figure 5.5 shows this overall process.

In a real-world scenario, workers adapt to new environments and gain new competencies when transitioning to jobs [149, 150]. If they do not use old competencies in their new



Figure 5.5: The overall structure used to represent the time in an occupation, and the competencies of the user through each occupation. The user starts with a list of competencies, and these grow or decay depending on the occupation they spent time in.Furthermore, each individual has a list of occupation-time pairs that are used to reference this information.

job, they may forget or become less proficient in them [149, 150]. Therefore, a growth a decay function are implemented in our approach when traversing the career path. When the user reaches a new occupation, their competencies will change based on growth and decay functions as adapted from work in training [151] and learning [152] where growth is the following sigmoid-based function,

$$f_{growth}: c_{kx} \to L_0 + \frac{L_{kn}}{1 + e^{-\frac{I_{kn}}{100}(\lambda + t_x - Z_x)}}$$
 (5.12)

and decay is defined as the following,

$$f_{decay}: c_{kx} \to L_0 - \frac{L_{kn}}{1 + e^{-\frac{I_{kn}}{100}(\lambda + t_x - \frac{1}{2}Z_x)}}$$
(5.13)

where L_0 is the initial level of the user for the *kth* competency *c* in state *x* of the path, L_{kn} is the level associated with the competency for the *nth* occupation, t_x is the time in the occupation-time pair from the pathway in state *x*, and I_{kn} is the importance level for the competency. We weigh the decay as half of the SVP time in months, and growth as the entire SVP time in months, to account for on-the-job training to allow the worker to learn at a quicker rate than decay. Furthermore, λ is provided to allow growth when the initial value is 0, the value of which is provided in Table 5.1.

However, to prevent the level of the user from growing beyond the occupation's level for the competency, we use the following to saturate the growth,

$$\begin{cases}
L_{kx} & L_{kx} \le L_{kn} \\
L_{kn} & L_{kx} > L_{kn}
\end{cases}$$
(5.14)

and similarly for the decay we use the following to prevent a skill from decaying below zero,

$$\begin{cases} 0 & L_{kx} \le 0 \\ L_{kx} & L_{kx} > 0 \end{cases}$$
(5.15)

where L_{kx} is the level associated with the competency c_k for the user, and L_{kn} is the level associated with the competency for the *nth* occupation.

Previous work have observed growth and decay at an occupation level [149, 150, 151]. However, for this approach, we use a similar method but with O*NET's importance and level scales for each competency [108]. First, looking at the distribution of the importance and skill for each competency (i.e., knowledge, skill, and ability) for the first occupation in each career cluster, we can see that as the importance of a competency changes, the level changes in a similar fashion. This can be observed from the results in Figure 5.6 for knowledge, Figure 5.7 for skills, and Figure 5.8 for abilities.

This allows a more realistic growth and decay of each competency for the occupation, such that a less important competency that is not exercised as often in the job will grow at a slower pace. Furthermore, if the user starts at a higher level, the decay will be slower if they move to an occupation that does not use that competency. This can be observed in a sample growth and decay of the skill "Mathematics" from O*NET for the occupation "Singer" and "Software Developer" in Figure 5.9. In this demonstration, the two occupations are largely different, and require different levels for the skill Mathematics. We can observe how a user with a starting level of 100 in Mathematics, and enters the Singer position, will



Level and Importance for Knowledge of Occupations

Figure 5.6: The distribution of level and importance for knowledge in O*NET.





Figure 5.7: The distribution of level and importance for skills in O*NET.

decay at a faster rate than when entering the position of Software Developer since the skill is more important for that occupation. Furthermore, a user with a level of 0 entering the position of Software Developer grows much faster than for entering the occupation Singer, that requires a lower level for the skill.



Level and Importance for Abilities of Occupations

Figure 5.8: The distribution of level and importance for abilities in O*NET.



Figure 5.9: The growth and decay functions implemented on the skill "Mathematics" for two highly different occupations.

5.3.3.4 Competency Transferability

Next, we include the results from Chapter 4 to capture how other competencies the user has can transfer between occupations. This will capture a more accurate representation of the worker and the distance in the occupation transfer in the pathway. Even if one competency is not present in the user's profile, others may contribute if they have a high transferability score.

For a given occupation and worker, if the worker's competency c_k in step x of the pathway is 0, then the transferred score is determined with the following,

$$L_{kx} = \frac{1}{k} (\sum_{i=0}^{k} c_{ki})$$
(5.16)

where the new level for the kth competency is the mean of the transferable competencies $c_k i$, where a competency is considered transferable if it is larger than the threshold $c_{ki} > 0.85$. A score of 85% or higher would indicate the competency could transfer to the other.

5.3.3.5 Final Representation

The overall fitness function can be represented as the following,

$$F(P) = \sum_{i=1}^{x} D(W_{i-1}, O_i) + D(W_x, A)$$
(5.17)

where the fitness, or cost of the pathway, is equal to the total distance between each occupation hop, and the final end-goal occupation A, where the time provided to the growth and decay function is 0. The pseudocode for the algorithm to calculate the distance incorporating the earlier discussed elements is presented in Algorithm 1. Algorithm 1: Calculating the cost of a pathway

```
Input : The worker W
         Individual P = [(O_1, t_1), (O_2, t_2), \dots, (O_i, t_i)]
         End-Goal Occupation A
Output: Pathway Cost
distance = 0
// Set the last occupation to compare to as the worker's past
prevOccupation = W
// loop through all occupation-time pairs in path
for x \leftarrow 0 to i - 1 do
   distance += \text{occupationDistance}(prevOccupation, P[x][0])
   for k \leftarrow 0 to W.competencies do
      if W.competencies[k] \in P[x][0].competencies \land W.competencies[k] \leq
        P[x][0].competencies[k] then
         growF(W.competencies[k])
                                                        // Grow the competency
       else
         decayF(W.competencies[k])
                                                       // Decay the competency
       end
   end
end
distance += \text{occupationDistance}(W, A)
return distance
```

5.3.4 Mutation

Two mutation operators were included in the genetic algorithm to introduce randomness in the population. These are further described next.

5.3.4.1 Random Occupation Assignment

The first mutation operator will switch the occupation ID in one occupation-time pair in the career pathway with another. This will allow variability of occupations and modified competencies and requirements in the career pathway. The index of the occupation-time pair in the pathway is randomly selected.

5.3.4.2 Time Perturbation

The second mutation operator will modify the time spent in an occupation in the pathway. The time will be increased or decreased by 1 month each mutation. In developing this operator, we found that it provided a more fine-grained representation of the time for each occupation. Furthermore, if the solution mutates past the time limited for an occupation, or below the job zone minimum time requirement, we prevent the mutation from going past these values. This way, the solution cannot mutate to become infeasible in the real world (e.g., working in an occupation for less than a month, or entering a job that requires a four-year degree without the required education and/or experience).

5.3.5 Selection

To create the new generation, tournament selection was used. However, we experiment with two different selection approaches to find the best result, and compare against random selection. These are further described next.

5.3.5.1 Tournament

Selection operators are designed such that the probability of selecting an individual for the new population is based on their fitness [153]. In tournament selection, a random predefined number of individuals T is selected from the population of size K, and from this the probability s of selecting an individual r, as defined in [154], can be represented by the following,

$$s(P_r) = \frac{F(P_r)}{\sum_{w=1}^{K} F(P_w)}$$
(5.18)

where the probability is inversely proportional to an individual's cost, or the fitness function [153, 154].

5.3.5.2 Lexicase

Though tournament selection is a commonly applied selection operator for genetic algorithms, a more recent development is Lexicase selection, proposed in 2014 for genetic programming problems [155]. This selection method is used to select individuals from the population, based on the test cases [155] present in the solution space. First, when selecting a parent, the algorithm will randomly arrange the test cases, and remove individuals in the population with poor performance on the first case. Next, if there is more than one individual, it iterates through the next test case and repeats. Once there is only one individual in the pool or all test cases have been used and it randomly selects an individual from the pool.

The advantage of Lexicase is the diversity it brings to the population by considering different cases. For our application, we consider cases of: (1) job zone distance, (2) competency difference, and (3) requirements difference.

5.3.6 Crossover

To create the next part of the population, crossover was applied from the selected parents. We experimented with several different crossover mechanisms including one-point, two-point, and uniform. In these crossover mechanisms, the occupation-time pair was exchanged between the two parents, depending on the randmly selected crossover points.

5.3.6.1 One-Point and Two-Point

The first two experiments with crossover were with one-point and two-point crossover operators. A one-point crossover strategy consists of randomly selecting an index in the pathway, and the pairs G to the right of the index are swapped between the two parents. A two-point crossover strategy consists of randomly selecting two indices in the pathway, and swapping the pairs between both indices. Furthermore, due to the different lengths in pathways, if one selected parent is longer than the other, the individual with the smaller length is used as the first parent for deciding the indices for crossover. This is an approach similar to the one used in [156]. The probability of crossover for this experiment is further covered in Table 5.1.

5.3.6.2 Uniform

In uniform crossover, each individual occupation-time pair is swapped between the parents with equal probability. The probability used for this experiment is further covered in Table 5.1. The advantage this approach presents compared to one-point and two-point crossover is the ability for more diverse offspring in comparison to the parents.

5.4 Evaluation and Results

The overall configuration for the GA is shown in Table 5.1. These parameters were determined after several runs and experiments with the GA. We start with a population size of 100 individuals, and run about 100 generations, or until there is no further change in the minimum fitness in the population after 10 generations. The mutation rates for the occupation mutation and time perturbation are 25% and 50%, respectively. Crossover was performed in each generation, and the selection size for tournament selection was 5 individuals. Lastly, the maximum length of a path is set to 13, as described earlier in 5.8, and the maximum time to be spent in a job is 60 months, as described in 5.7.

We used Python 3.7 and the genetic algorithm library DEAP [157] to implement the career pathway GA. O*NET version 24.0 was used to reference the different occupation data, and the data files "knowledge", "skills", "abilities", "interests", "work styles", "work values", "job zones", and "green occupations". We also use the bright occupations exported from the O*NET site [114]. Furthermore, sample configuration files representing worker profiles were setup with the competencies, requirements, job zone, and SOC code for each worker being represented in the problem. The worker profiles used were those discussed in Chapter 3,

and the worker profiles are the characteristics described in O*NET. In a real application, these characteristics would be extracted through an assessment by the worker, but for the experiments we use the data from O*NET.

Property	Value
Starting Population Size	100
Number of Generations	100
Occupation Mutation Rate	0.25
Time Mutation Rate	0.5
Crossover Rate	1
Uniform Crossover Rate	0.5
Tournament Size	5
Growth & Decay λ	0.001
w_C	0.25
w_R	0.25
w_Z	50
Transferability Threshold	0.85
Maximum Path Length	13
Maximum Job Time	60

Table 5.1: Parameters used for the genetic algorithm.

5.4.1 Crossover Experiments

We first run three experiments with different crossover operators: two-point, one-point, and uniform crossover. For uniform crossover, we used a 50% chance of crossover occurring for each occupation-time pair. Furthermore, a tournament selection was used consistently for all three experiments. The worker profile for occupation "41-2011.00", or "Cashier" was used for all these experiments, since this was classified as an occupation with a poor outlook, and we look at the results to the bright and green outlook occupations. The results for twopoint crossover is presented in Figure 5.10, one-point crossover in Figure 5.11, and uniform crossover in Figure 5.12.

We can see from these results the results were similar across the three crossover operations, though the one-point crossover provided better results (i.e., reached a more fit solution



GA Run with Two-Point Crossover

Figure 5.10: Fitness (i.e., cost) results for the two-point crossover mechanism for bright and green occupations.



GA Run with One-Point Crossover

Figure 5.11: Fitness (i.e., cost) results for the one-point crossover mechanism for bright and green occupations.

quicker) for more of the occupations. Therefore, one-point crossover was used in each of the selection experiments, as described next.



GA Run with Uniform Crossover

Figure 5.12: Fitness (i.e., cost) results for the uniform crossover mechanism for bright and green occupations.

5.4.2 Selection Experiments

Using the one-point crossover operator described above, we ran experiments with two selection methods: lexicase selection and tournament selection. For tournament selection, we used a selection we used a tournament size of 5 individuals. The results can be compared to random selection presented in Figure 5.13. The results for lexicase selection can be seen in Figure 5.14, and the results for tournament selection can be seen in the previous section in Figure 5.11.

We can see from these results, both selection mechanisms outperformed random selection. Between these, the tournament selection mechanism provide slightly better results than lexicase, reaching the stop criteria faster. Therefore, for the final results, we use tournament selection with one-point crossover.

5.4.3 Discussion

Figure 5.11 shows the overall final results. Each generation with the GA took 1.8 seconds, with about 145 seconds for achieving a solution, using a PC with an 8th generation Intel Core



GA Run with Random Selection

Figure 5.13: Fitness (i.e., cost) results with a random selection mechanism for bright and green occupations.



GA Run with Lexicase Selection

Figure 5.14: Fitness (i.e., cost) results for the lexicase selection mechanism for bright and green occupations.

i7 with 16 GB RAM. Looking at the final results, the population converges to a better solution within the range of 60 to 100 generations. We select the lowest costing path from the final population as the final pathway solution. For a better understanding of the solutions provided by our approach, using the occupation "41-2011.00" we determine pathways between all other occupations in O*NET. The results, or distribution of the solution costs are presented in Figure 5.15, where the cost of pathways to occupations outside of the cluster of the starting occupation and those in the same cluster and highlighted. The distribution of the solution costs separating green and bright outlook occupations are shown in Figure 5.16.

From these results, we can see the resulting solutions where the goal occupation was in the same cluster (i.e., "41") had overall lower costs, than those in other clusters. Since there are more occupations in clusters outside of cluster 41, there are more solutions from these clusters in the distribution.



Figure 5.15: The distribution of pathway costs from the occupation "41-2011.00" to all others in O*NET.

From the results in Figure 5.16, we can see that the distribution between green and bright outlook occupations in O*NET are similar. The goal occupation is not labeled as green or bright outlook, and instead has a poor outlook, and so from this we can see no affect of green or bright outlook occupations to this occupation. However, we can see that there are a slightly higher number of pathway costs between 150 and 300 for green and bright occupations, in part because occupations with these flags typically have a higher job zone. Job zone weight can also be seen in Figure 5.15, where pathways with a cost greater than 200 have job zones higher than that of "41-2011.00", that has a job zone of 1.



Figure 5.16: The distribution of pathway costs for green and bright occupations in O*NET from "41-2011.00" to all others in O*NET.

To outline typical and atypical pathways, and compare advantages and limitations of the current approach, we present sample pathways from the results in Figure 5.18. This shows pathways for the starting occupation "41-2011.00" Cashier, and we also provide solutions for two starting occupations with a bright outlook and a green flag. The cost of these pathways for the Cashier "41-2011.00" occupations are highlighted on the distribution graph in Figure 5.17.

From Figure 5.17, we can see a variety of costs for pathways. However, we can see the occupations with a higher job zone such as "17-2199.00" and "27-3031.00" with a job zone of 4, have a slightly higher cost than "35-3041.00" with a job zone of 1. However, there is still an impact on the cost through the fit of requirements and competencies for the worker, which can be seen in occupation "51-9199.01", that has a job zone of 2, but largely different competencies and requirements than the starting occupation.

Looking at the pathways in Figure 5.18, the pathways seem feasible for a real world user,



Figure 5.17: Path costs highlighted from Figure 5.18.

with occupations with similar skills and knowledge between each other in the pathway. we can see that there are three common patterns in the solutions: (1) shorter pathways overall, (2) repeats in the occupation-time pairs in the path, (3) and multiple occupations in a similar career cluster. Overall, the GA tended to return shorter pathways for each worker profile. Even though the permissible range for pathways was up to 13 occupations, a majority of the most fit solutions at the end of the run was anywhere from two to four occupations. This is in part due to the distance measurement used between occupations. Since O*NET has such a wide range of competencies and requirements, there will be a large amount of variation between each competency and occupation, resulting in a non-zero value for each occupation hop in the career pathway. Furthermore, we can see repeats in the pathway where the user can return a previously held occupation, as seen in path 6 in Figure 5.18, which may occur in a real-world scenario. Last, we can see a pattern in all pathways where the transition is to an occupation in a different cluster, where a majority of the pathway would be in the same career cluster. Again, this is due to the cost of transferring outside to a different cluster.

We also look at two paths, one with a starting green occupation "17-2141.00" Mechanical Engineers with a job zone of 4, and one with a bright outlook occupation "15-1132.00"



Figure 5.18: Sample paths to show different examples of results illustrating different aspects from the approach. Nodes in the pathway are color coded according to different career clusters. The sun symbol indicates an occupation with a bright outlook. The leaf symbol indicates an occupation that is green. Each node between the start and end occupation have the time spent in the occupation in months (m).

Application Software Developers with a job zone of 4. Both of these resulted in reasonable pathways. However, we can see for occupations with higher job zone, the experience and time spent in intermediate occupations in the path are not properly captured. For instance, for the sixth pathway in Figure 5.18, the transition back to Software Developers occurred, after Database Architectures "15-1199.06", and last to Biostasticians "15-2041.01", that is a job zone of 5. However, this transfer may not be needed in a real world scenario. This may be in part due to the division of job zones, where job zone 4 and above have a very wide time range (i.e., 4 years of preparation, experience, education, or more). Therefore, for future work, we can see that a more fine-grained approach is necessary, with additional data sources for education and experience.



GA Run Compared Against Djikstra's Algorithm Results

Figure 5.19: Comparing the results of the GA with modifications against common shortest path algorithms such as Djikstra's. The green line is the cost of the path found by the shortest path algorithm.

To further evaluate our approach, we compare the distance measurement used against common shortest path algorithms, such as Djikstra's algorithm. To do this, we remove any time-based aspect of our approach including growth and decay of competencies, and the time in the occupation-time pair. The results for this can be seen in Figure 5.19, where the green dotted line is the cost of the shortest path found by Djikstra's. It can be seen that the cost of the path found by Djikstra's is near 20 - 70 for each goal occupation, and that the algorithm has converged to that minimum cost path.

Overall, we can see our approach provides promising pathway solutions between a wide variety of occupations using the pathway representation and GA operators. Users could use this approach to determine paths that may fit best with their current profile. However, there are areas of improvement where the solutions found could be more realistic, as further described next.

5.5 Conclusion and Future Work

In this chapter, we present a genetic algorithm for career pathway planning. The algorithm uses new operators for mutation and a fitness function to mimic a real-world career pathway traversal. Furthermore, several factors such as time spent in different occupations, fit of competencies, and worker values and styles, were accounted for in the pathway representation used. The results are promising, with the GA obtaining better results after experimenting with different selection and crossover operators. However, there are still areas where this approach can be improved and where there were limitations.

We plan to try additional worker profiles (including real example resumes to evaluate our approach) to observe the different outcomes. Due to the time taken for one run of the GA, it is infeasible to attempt every possible worker profile/occupation to occupation pathway. Furthermore, the representation of the problem, where job zone was included in calculating the cost, may need to be reconsidered after including additional data sources. Better accounting for credentials and experience, such as degrees, prior to entering an occupation, would change the cost of transition from one occupation to another and the time spent in the occupation itself, making the pathway more realistic. Furthermore, another area of improvement would be to include other data sources outside of O*NET and courses, such as BLS statistics on salary, unemployment rates, job satisfaction, and job location, along with NAICS codes, could be taken into account for a more personalized career planner. Last, the growth and decay functions could be further improved by using models specific to each occupation in O*NET, and accounting for how individuals in those occupations are able to learn or retain information.

Overall, a career pathway planning tool will assist workers in achieving a higher job satisfaction and employee retention for companies they work in. A GA to optimize the career pathways for individuals, is a step forward to data-driven approaches for career planning and helping students, job seekers, and career advisors.

CHAPTER 6

CONCLUSION

The aim of this thesis is to address two key problems faced by organizations and workers in the future of work. First, an overview of how advances in AI applications for recruitment have been made is presented, along with the implications of bias in these models, methods for bias mitigation, and the work that has been done in model interpretability and postprocessing methods for achieving fairness. Next, an approach is presented to address the worker problem of searching and optimizing their career pathway. For this, the dataset used by our approach, the Occupational Information Network (O*NET), is covered along with its different components. Next, O*NET is extended by creating links between the worker's knowledge, skills, and abilities (i.e., competencies), using a natural language processing-based approach to career pathway optimization, using a genetic algorithm to mimic a realistic worker career traversal. Results for this approach were shown to provide its feasibility for real workers.

Future work in this area, in addition to those described by the specific problems addressed in previous chapters, include other applications for workers and organizations for the future of work. For organizations a solution to fair recruitment is needed, by providing a hiring approach that explains decisions made in the process and identifying bias. This would mean using fairness toolkits described in Chapter 2 to explain features fed to the hiring model, and providing a method to determine the cause of bias given the metrics presented. For workers, it is necessary to not only assist in optimization of their career pathway, but assist in reaching each step in the pathway. Therefore, future work also includes a more detailed pathway representation and better recommendations with an enhanced occupation network, using other data sources to extend O*NET. We believe there is also room for further improvement in the problems addressed in this thesis, as discussed in earlier chapters. This includes improvement of calculating competency transferability, through fine-tuning of BERT and a sentence level representation to capture variance of terms. Implementing these could assist people in future decision-making in regards to their career and organization.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical meth*ods in natural language processing (EMNLP), pages 1532–1543, 2014.
- [2] Glove: Global vectors for word representation. https://nlp.stanford.edu/projects/glove/.
- [3] James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi. Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Institute*, 2017.
- [4] Foundation for Young Australians. The new work order: ensuring young australians have skills and experience for the jobs of the future, not the past, 2015.
- [5] OECD. The future of work OECD employment outlook 2019. 2019.
- [6] Organization for Economic Cooperation and Development (OECD). Future of work and skills. 2nd meeting of the G20 employment working group, 2017.
- [7] James Manyika. A future that works: AI, automation, employment, and productivity. McKinsey Global Institute Research, Tech. Rep, 2017.
- [8] The council of economic advisers. Addressing america's reskilling challenge. 2018.
- [9] Engineering National Academies of Sciences and Medicine. Information Technology and the US Workforce: Where Are We and Where Do We Go from Here? The National Academies Press, 2017.
- [10] Steven Miller and Debbie Hughes. The quant crunch: How the demand for data science skills is disrupting the job market. *Burning Glass Technologies*, 2017.
- [11] Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.
- [12] Mika Pajarinen and Petri Rouvinen. Computerization threatens one third of finnish employment. *Etla Brief*, 22(13.1):2014, 2014.
- [13] C Brzeski and I Burk. The robots come. consequences of automation for the german labour market. *ING DiBa Economic Research*, 2015.
- [14] J Bowles. The computerisation of european jobs. Bruegel, Retrieved from www.bruegel.org/nc/blog/detail/article/1394-the-computerisation-of-european-jobs/, 01 2014.

- [15] Melanie Arntz, Terry Gregory, and Ulrich Zierahn. The risk of automation for jobs in oecd countries. 2016.
- [16] David H Autor and Michael J Handel. Putting tasks to the test: Human capital, job tasks, and wages. *Journal of Labor Economics*, 31(S1):S59–S96, 2013.
- [17] Alexandra Spitz-Oener. Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of labor economics*, 24(2):235–270, 2006.
- [18] Ekkehard Ernst, Rossana Merola, and Daniel Samaan. The economics of artificial intelligence: Implications for the future of work. ILO Future of Work Research Paper Series No, 5, 2018.
- [19] Yochi Cohen-Charash and Paul E Spector. The role of justice in organizations: A meta-analysis. Organizational behavior and human decision processes, 86(2):278–321, 2001.
- [20] Jerald Greenberg. Organizational justice: The dynamics of fairness in the workplace. 2011.
- [21] Stephen W Gilliland. The perceived fairness of selection systems: An organizational justice perspective. Academy of management review, 18(4):694–734, 1993.
- [22] Brian R Dineen, Raymond A Noe, and Chongwei Wang. Perceived fairness of webbased applicant screening procedures: Weighing the rules of justice and the role of individual differences. Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 43(2-3):127-145, 2004.
- [23] Philip M Podsakoff, Scott B MacKenzie, Julie Beth Paine, and Daniel G Bachrach. Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of management*, 26(3):513–563, 2000.
- [24] Indeed job search. https://www.indeed.com/.
- [25] HireVue about us. https://www.hirevue.com/company/about-us.
- [26] SHRM. Screening and evaluating candidates, Jan 2019. https: //www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/ screeningandevaluatingcandidates.aspx.
- [27] SHRM. Your guide to applicant tracking systems, Oct 2018. https://www.shrm.org/ resourcesandtools/hr-topics/talent-acquisition/pages/ats-table.aspx.
- [28] Dave Zielinski. Today's ATS solutions go well beyond resume storage, Oct 2018. https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/ats-solutions-buyers-guide-shrm.aspx.

- [29] Baby steps in HR technology: What is resume parsing? https://recruiterbox.com/ blog/baby-steps-in-hr-technology-what-is-resume-parsing-2.
- [30] Kathryn M Neckerman and Joleen Kirschenman. Hiring strategies, racial bias, and inner-city workers. *Social problems*, 38(4):433–447, 1991.
- [31] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018.
- [32] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [33] Timothy A Judge and Cindy P Zapata. The person-situation debate revisited: Effect of situation strength and trait activation on the validity of the big five personality traits in predicting job performance. Academy of Management Journal, 58(4):1149– 1179, 2015.
- [34] Christopher D Nye, Rong Su, James Rounds, and Fritz Drasgow. Interest congruence and performance: Revisiting recent meta-analytic findings. *Journal of Vocational Behavior*, 98:138–151, 2017.
- [35] Frank Parsons. Choosing a vocation. Houghton Mifflin, 1909.
- [36] Jeffrey R Edwards. Person-job fit: A conceptual integration, literature review, and methodological critique. John Wiley & Sons, 1991.
- [37] Chen Zhu, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. Person-job fit: Adapting the right talent for the right job with joint representation learning. ACM Transactions on Management Information Systems (TMIS), 9(3):1–17, 2018.
- [38] Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Liang Jiang, Enhong Chen, and Hui Xiong. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 25–34, 2018.
- [39] Michael S Cole, Hubert S Feild, William F Giles, and Stanley G Harris. Recruiters' inferences of applicant personality based on resume screening: do paper people have a personality? *Journal of Business and Psychology*, 24(1):5–18, 2009.
- [40] Barbara K Brown and Michael A Campion. Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology*, 79(6):897, 1994.
- [41] Satyaki Sanyal, Souvik Hazra, Soumyashree Adhikary, and Neelanjan Ghosh. Resume parser with natural language processing. *International Journal of Engineering Science*, 4484, 2017.

- [42] Pooja Shivratri, Preeti Kshirsagar, Rashmi Mishra, Ronit Damania, and Nandana Prabhu. Resume parsing and standardization. International Journal of Computer Sciences and Engineering, 3(3):129–131, 2015.
- [43] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In Proc. International Conference on Internet and Web Applications and Services. Citeseer, 2012.
- [44] CH Ayishathahira, C Sreejith, and C Raseek. Combination of neural networks and conditional random fields for efficient resume parsing. In 2018 International CET Conference on Control, Communication, and Computing (IC4), pages 388–393. IEEE, 2018.
- [45] Chen Zhang and Hao Wang. Resumevis: A visual analytics system to discover semantic information in semi-structured resume data. ACM Transactions on Intelligent Systems and Technology (TIST), 10(1):8, 2018.
- [46] Martin Cronje and Abejide Ade-Ibijola. Automatic slicing and comprehension of cvs. In 2018 5th International Conference on Soft Computing & Machine Intelligence (IS-CMI), pages 99–103. IEEE, 2018.
- [47] Qingxin Meng, Hengshu Zhu, Keli Xiao, Le Zhang, and Hui Xiong. A hierarchical career-path-aware neural network for job mobility prediction. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 14–24, 2019.
- [48] Rui Yan, Ran Le, Yang Song, Tao Zhang, Xiangliang Zhang, and Dongyan Zhao. Interview choice reveals your preference on the market: To improve job-resume matching through profiling memories. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 914–922, 2019.
- [49] Junhua Liu, Chu Guo, Yung Chuen Ng, Kristin L Wood, and Kwan Hui Lim. IPOD: Corpus of 190,000 industrial occupations. arXiv preprint arXiv:1910.10495, 2019.
- [50] Yingya Zhang, Cheng Yang, and Zhixiang Niu. A research of job recommendation system based on collaborative filtering. In 2014 Seventh International Symposium on Computational Intelligence and Design, volume 1, pages 533–538. IEEE, 2014.
- [51] Ran Le, Wenpeng Hu, Yang Song, Tao Zhang, Dongyan Zhao, and Rui Yan. Towards effective and interpretable person-job fitting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1883–1892, 2019.
- [52] LinkedIn. https://www.linkedin.com/.
- [53] Richard J Oentaryo, Xavier Jayaraj Siddarth Ashok, Ee-Peng Lim, and Philips Kokoh Prasetyo. JobComposer: Career path optimization via multicriteria utility learning. arXiv preprint arXiv:1809.01062, 2018.

- [54] Yu Cheng, Yusheng Xie, Zhengzhang Chen, Ankit Agrawal, Alok Choudhary, and Songtao Guo. Jobminer: A real-time system for mining job-related patterns from social media. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1450–1453. ACM, 2013.
- [55] Huang Xu, Zhiwen Yu, Hui Xiong, Bin Guo, and Hengshu Zhu. Learning career mobility and human activity patterns for job change analysis. In 2015 IEEE International Conference on Data Mining, pages 1057–1062. IEEE, 2015.
- [56] Navneet Kapur, Nikita Lytkin, Bee-Chung Chen, Deepak Agarwal, and Igor Perisic. Ranking universities based on career outcomes of graduates. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 137–144. ACM, 2016.
- [57] Vachik S Dave, Baichuan Zhang, Mohammad Al Hasan, Khalifeh AlJadda, and Mohammed Korayem. A combined representation learning approach for better job and skill recommendation. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 1997–2005, 2018.
- [58] Shiqiang Guo, Folami Alamudun, and Tracy Hammond. Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60:169–182, 2016.
- [59] Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. NEMO: Next career move prediction with contextual embedding. In *Proceedings of the* 26th International Conference on World Wide Web Companion, pages 505–513, 2017.
- [60] Dena F Mujtaba and Nihar R Mahapatra. Ethical considerations in ai-based recruitment. In 2019 IEEE International Symposium on Technology and Society (ISTAS), pages 1–7. IEEE, 2019.
- [61] S Biswas. The beginner's guide to AI in HR. HR Technologist, 5, 2019.
- [62] GoArya. https://goarya.com/.
- [63] Google Hire. https://hire.google.com/.
- [64] Plum. https://www.plum.io/how-it-works.
- [65] David Meyer. Amazon reportedly killed an AI recruitment system because it couldn't stop the tool from discriminating against women. Fortune. Tillgänglig online: https://fortune. com/2018/10/10/amazon-ai-recruitment-bias-womensexist/(2019-09-27), 2018.
- [66] N Lewis. Will AI remove hiring bias. SHRM, 2018.
- [67] A Johnson. 13 common hiring biases to watch out for. *Harver*, 2018. https://harver. com/blog/hiring-biases/.

- [68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference* on world wide web, pages 1171–1180, 2017.
- [69] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–16, 2019.
- [70] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In NIPS Symposium on Machine Learning and the Law, volume 1, page 2, 2016.
- [71] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148, 2015.
- [72] Akhil Krishnakumar. Assessing the fairness of AI recruitment systems. 2019.
- [73] Ziyuan Zhong. A tutorial on fairness in machine learning. *Medium*, 2018.
- [74] CS294: Fairness in machine learning. https://fairmlclass.github.io/.
- [75] Equal Employment Opportunity Commission et al. Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register*, 44(43):11996–12009, 1979.
- [76] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.
- [77] John E Hunter. Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of vocational behavior*, 29(3):340–362, 1986.
- [78] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.
- [79] Sahil Verma and Julia Rubin. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pages 1–7. IEEE, 2018.
- [80] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.
- [81] Anna B Holm. E-recruitment: towards an ubiquitous recruitment process and candidate relationship management. German Journal of Human Resource Management, 26(3):241–259, 2012.

- [82] Jennifer D Shapka, Jose F Domene, Shereen Khan, and Leigh Mijin Yang. Online versus in-person interviews with adolescents: An exploration of data equivalence. Computers in Human Behavior, 58:361–367, 2016.
- [83] Derek S Chapman and Jane Webster. The use of technologies in the recruiting, screening, and selection processes for job candidates. *International journal of selection and assessment*, 11(2-3):113–120, 2003.
- [84] Alice Snell. Researching onboarding best practice: Using research to connect onboarding processes with employee satisfaction. *Strategic HR Review*, 2006.
- [85] Tianlu Wang, Jieyu Zhao, Kai-Wei Chang, Mark Yatskar, and Vicente Ordonez. Adversarial removal of gender from deep image representations. arXiv preprint arXiv:1811.08489, 2018.
- [86] Rachel Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Ramazon Kush, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943, 2018.
- [87] Lucid. https://github.com/tensorflow/lucid.
- [88] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. FairTest: Discovering unwarranted associations in data-driven applications. In 2017 IEEE European Symposium on Security and Privacy (EuroS&P), pages 401–416. IEEE, 2017.
- [89] Julius A. Adebayo. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.
- [90] Niels Bantilan. Themis-ML: A fairness-aware machine learning interface for end-toend discrimination discovery and mitigation. Journal of Technology in Human Services, 36(1):15–30, 2018.
- [91] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223, 2019.
- [92] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* international conference on knowledge discovery and data mining, pages 1135–1144, 2016.
- [93] S.M. Lundberg and S. Lee. SHAP. https://www.plum.io/how-it-works.
- [94] Benjamin Bengfort and Rebecca Bilbro. Yellowbrick: Visualizing the scikit-learn model selection process. *Journal of Open Source Software*, 4(35):1075, 2019.

- [95] Eli5. https://github.com/TeamHG-Memex/eli5.
- [96] Skater. https://github.com/oracle/Skater.
- [97] StellarGraph machine learning library. https://github.com/stellargraph/stellargraph.
- [98] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelli*gence, 2018.
- [99] Przemysław Biecek. DALEX: explainers for complex predictive models in r. The Journal of Machine Learning Research, 19(1):3245–3249, 2018.
- [100] Machine learning interpretability (MLI). https://github.com/h2oai/mli-resources.
- [101] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577, 2018.
- [102] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.
- [103] Reductions for fair machine learning. https://github.com/microsoft/fairlearn.
- [104] Radwa Elshawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pages 275–280. IEEE, 2019.
- [105] Alexey Zagalsky, Joseph Feliciano, Margaret-Anne Storey, Yiyun Zhao, and Weiliang Wang. The emergence of github as a collaborative platform for education. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pages 1906–1917, 2015.
- [106] Dena Mujtaba and Nihar Mahapatra. Towards data-enabled career planning with the occupational information network (o* net). In 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pages 1547–1549. IEEE, 2019.
- [107] Norman G Peterson, Michael D Mumford, Walter C Borman, P Jeanneret, and Edwin A Fleishman. An occupational information system for the 21st century: The development of O*NET. American Psychological Association, 1999.
- [108] The O*NET® content model, 2019. https://www.onetcenter.org/content.html.
- [109] Rene V Dawis and Lloyd H Lofquist. A psychological theory of work adjustment: An individual-differences model and its applications. University of Minnesota press, 1984.
- [110] O*NET® interest profiler. https://www.mynextmove.org/explore/ip.

- [111] Gary D Gottfredson and John L Holland. Dictionary of Holland occupational codes. Psychological Assessment Resources Inc, 1996.
- [112] Terence J Tracey and James Rounds. Evaluating the riasec circumplex using high-point codes. Journal of Vocational Behavior, 41(3):295–311, 1992.
- [113] Erich C Dierdorff, Jennifer J Norton, Donald W Drewes, Christina M Kroustalis, David Rivkin, and Phil Lewis. Greening of the world of work: Implications for o* net®-soc and new and emerging occupations. O* NET, February, 2009.
- [114] Additional initiatives, 2020. https://www.onetcenter.org/initiatives.html# bright-outlook.
- [115] O*NET online help: Job zones., 2019. https://www.onetonline.org/help/online/zones.
- [116] MT Allen, G Waugh, M Shaw, S Tsacoumis, D Rivkin, P Lewis, M Brendle, D Craven, C Gregory, and D Connell. The development and evaluation of a new o* net related occupations matrix. National Center for O* NET Development. URL https://www. onetcenter. org/dl_files/Related. pdf, 2012.
- [117] O*NET® products at work. https://www.onetcenter.org/paw.html.
- [118] Fastest growing occupations. https://www.bls.gov/ooh/fastest-growing.htm.
- [119] All green economy sectors. www.onetonline.org/find/green?n=0&g=Go.
- [120] Fastest declining occupations. https://www.bls.gov/emp/tables/ fastest-declining-occupations.htm.
- [121] Amy L Kristof-Brown, Ryan D Zimmerman, and Erin C Johnson. Consequences of individuals'fit at work: A meta-analysis of person–job, person–organization, person– group, and person–supervisor fit. *Personnel psychology*, 58(2):281–342, 2005.
- [122] O*NET. O*NET online. https://www.onetonline.org/. https://www.onetonline.org/.
- [123] Benjamin Schneider. The people make the place. Personnel psychology, 40(3):437–453, 1987.
- [124] John L Holland. Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources, 1997.
- [125] David J Deming and Kadeem L Noray. STEM careers and technological change. Technical report, National Bureau of Economic Research, 2018.
- [126] Michael T DeGrosky. Transfer of knowledge, skills, and abilities from leadership development training. Northcentral University, 2013.
- [127] Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi, et al. Wordnet:: Similaritymeasuring the relatedness of concepts. In AAAI, volume 4, pages 25–29, 2004.

- [128] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.
- [129] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
- [130] Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1576–1586, 2015.
- [131] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *thirtieth AAAI conference on artificial intelligence*, 2016.
- [132] Yang Shao. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate semantic textual similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 130–133, 2017.
- [133] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.
- [134] Adrien Sieg. Text similarities : Estimate the degree of similarity between two texts, Nov 2019.
- [135] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [136] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
- [137] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461, 2018.
- [138] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQUAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.
- [139] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A largescale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326, 2018.
- [140] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
- [141] Han Xiao. BERT-as-service. https://github.com/hanxiao/bert-as-service, 2018.

- [142] Robert Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [143] Sherry E Sullivan and Yehuda Baruch. Advances in career theory and research: A critical review and agenda for future exploration. *Journal of management*, 35(6):1542– 1571, 2009.
- [144] Mark L Savickas and Erik J Porfeli. Career adapt-abilities scale: Construction, reliability, and measurement equivalence across 13 countries. *Journal of vocational behavior*, 80(3):661–673, 2012.
- [145] John H Holland. Genetic algorithms. Scientific american, 267(1):66–73, 1992.
- [146] John Grefenstette, Rajeev Gopal, Brian Rosmaita, and Dirk Van Gucht. Genetic algorithms for the traveling salesman problem. In *Proceedings of the first International Conference on Genetic Algorithms and their Applications*, volume 160, pages 160–168. Lawrence Erlbaum, 1985.
- [147] Edwin SH Hou, Nirwan Ansari, and Hong Ren. A genetic algorithm for multiprocessor scheduling. *IEEE Transactions on Parallel and Distributed systems*, 5(2):113–120, 1994.
- [148] NLS FAQs. https://www.bls.gov/nls/nlsfaqs.htm#anch41, Jan 2020.
- [149] Mohamad Y Jaber. Learning and forgetting models and their applications. *Handbook* of industrial and systems engineering, 30(1):30–127, 2006.
- [150] Douglas John MacKinnon. How individual skill growth and decay affect the performance of project organizations. PhD thesis, Stanford University, 2007.
- [151] Alexandra N Trani, Clayton J Hutto, Cara B Fausset, Samuel Cheng, Chris R Hale, Thomas McDermott, and Dennis J Folds. Modeling and simulation of skill decay at the organizational team level. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 740–744. SAGE Publications Sage CA: Los Angeles, CA, 2017.
- [152] A Newell and PS Rosenbloom. Mechanisms of skill acquisition and the law of practice cognitive skills and their acquisition. *Cognitive Skills and Their Acquisition*, pages 1–56, 1980.
- [153] David E Goldberg and Kalyanmoy Deb. A comparative analysis of selection schemes used in genetic algorithms. In *Foundations of genetic algorithms*, volume 1, pages 69–93. Elsevier, 1991.
- [154] Mauro Dragoni. An evolutionary strategy for concept-based multi-domain sentiment analysis. IEEE Computational Intelligence Magazine, 14(2):18–27, 2019.

- [155] Thomas Helmuth, Lee Spector, and James Matheson. Solving uncompromising problems with lexicase selection. *IEEE Transactions on Evolutionary Computation*, 19(5):630–643, 2014.
- [156] Sunil Nilkanth Pawar and Rajankumar Sadashivrao Bichkar. Genetic algorithm with variable length chromosomes for network intrusion detection. *International Journal of Automation and Computing*, 12(3):337–342, 2015.
- [157] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. DEAP: Evolutionary algorithms made easy. Journal of Machine Learning Research, 13:2171–2175, jul 2012.
- [158] Icons8 LLC. Icons8. http://icons8.com.