Assessing the Fairness of AI Recruitment systems

BY

Akhil Krishnakumar

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Master of Science

IN

Management of Technology

AT

Delft University of Technology

TO BE DEFENDED PUBLICLY ON January 25TH, 2019

Thesis committee

First Supervisor : Prof. M.V.Dignum

Second Supervisor : Dr. M.R.Alfano

Executive Summary

Businesses have leveraged Artificial Intelligence (AI) into many of their operational activities such as marketing, sales, and finance for its speed and cost-effectiveness. Lately, AI has also found applications in organizational recruitment processes. Unlike the conventional rule-based systems, present-day AI systems learn from data patterns—supported by the growing volumes of (big) data and increasing computing capacity—and make decisions independently without any human interventions. Thus, the perception that AI is fact-oriented and unbiased has led to this change in organizational recruitment practices. Though recent studies have shown that AI decisions could be unfair, scientific research on the fairness of AI recruitment systems is limited. This research fills this gap by designing a conceptual model to assist top-level HR managers in assessing the fairness of AI recruitment tools while drawing from information systems and responsible innovation literature.

Guided by Design Science Research (DSR), the development of the model entailed three cycles of research, i.e., relevance cycle (which focused on design environment), rigor cycle (which focused on the existing knowledge base), and design cycle (which focused on development and evaluation). The design environment was explored by reviewing the literature on fairness in recruitment and algorithmic biases. Understanding both the recruitment fairness and potential causes of unfairness in AI helped to define the goal of the conceptual model.

The design cycle was informed by the design principles for responsible AI, namely Accountability, Responsibility, and Transparency (ART), and General Data Protection Regulation (GDPR). The model presents seven dimensions which translate the principles to design requirements to assess the fairness of AI recruitment system. They are: (1)Justification; (2)Explanation; (3)Anticipation; (4)Reflexiveness; (5)Inclusion; (6)Responsiveness; and (7)Auditablity. The model

also ties these concepts with specific criteria of conventional recruitment fairness such as consistency, interpersonal fairness, job-relatedness, and statistical parity. Finally, the completeness of the model was evaluated by discussing its alignment with other frameworks that had similar objective and utility of the model was validated by collecting feedback from the intended users.

This thesis project makes several scientific and practical contributions. The research discusses the potential risks of using AI in the context of HR recruitment systems thereby contributes to the limited literature available in this respect. By using the DSR methodology for building the assessment model, this research serves as a case for DSR methodology in designing a non-IS artifact. Furthermore, the thesis has unified scattered studies in recruitment justice to provide a comprehensive overview of the characteristics of a fair recruitment system.

Building on the theoretical contributions, the study has developed an assessment model to assist top-level HR managers in assessing the fairness of an AI recruitment tool. Employing this assessment tool can have positive effects on a business organization and society by eradicating the unfairness or bias that AI recruitment tools can bring into the organization. It would also raise awareness regarding the risks of AI. Given that the GDPR (article 35) mandate organizations to take responsibility in assessing the impact while introducing automated processing in new contexts or purposes, the assessment model designed in this study supports these regulations.

Acknowledgment

This master thesis project is the final step of my studies at the Delft University of Technology. When I look back, it was a challenging journey. I could not have made it till here without the guidance and support of many people. So I take this opportunity to thank them.

First and foremost, I would like to thank my first supervisor Virginia Dignum for giving the opportunity to work on this exciting topic and guiding me throughout this research. I would also like to sincerely thank my second supervisor Mark Alfano for patiently reading through my report and giving me valuable feedback.

Secondly, a big thanks to all my friends who made these two and a half years unforgettable. Some special words of gratitude go to Rahul and Nitin for their support and encouragement.

Last but not least, I am grateful to my parents, sister, and brother-in-law for all the love and support.

I dedicate this thesis to the memory of my dearest uncle, Sasi Kunjachan.

Akhil Krishnakumar
Delft, January 2019

This page is intentionally left blank.

Contents

1 Introduction 1

1.1 Research Problem . . . 2
1.2 Research Objective . . . 3
1.3 Research Relevance . . . 3
1.4 Report Outline . . . 4

2 Research Methodology 7

2.1 DSR Methodology . . . 7
2.2 Research Questions . . . 11
2.3 Data Strategies . . . 12
2.4 Research Design . . . 14

3 Organizational Recruitment and Fairness 17

3.1 Organizational Recruitment . . . 17
3.2 Fairness . . . 19
3.3 Outsider's Fairness . . . 22
3.4 Summary . . . 22

4 Technical System and Problems 25

4.1 AI decision-making . . . 25
4.2 AI in Recruitment . . . 28
4.3 Training Data . . . 30
4.4 Algorithmic Focus . . . 30

5 Model Design 33

5.1 Responsible Design of AI . . . 33
5.2 Responsiblity . . . 35
5.3 GDPR and AI . . . 37
5.4 Assessment Model . . . 38

6 Evaluation 41
6.1 Evaluation Strategy . . . 41
6.2 Alignment . . . 42
6.3 Interviews . . . 45
6.4 Final Artifact . . . 47

7 Conclusion and Reflections 51
7.1 Answers to the Research Questions . . . 52
7.2 Implications and Recommendations . . . 53
7.3 Limitation and Future research . . . 54
7.4 MoT Curriculum Alignment . . . 56

References 62

List of figures

2.1 Design Science Research Cycles 9
2.2 Design Science Research - Activities 10
2.3 DSR entry points 10
2.4 Research Framework 15

3.1 Organizational Recruitment Activities 18

4.1 KDD Process 26

6.1 Evaluation Methods 42

This page is intentionally left blank.

List of Tables

3.1 Recruitment fairness 23

4.1 AI in Recruitment 29

5.1 Assessment Model 40

6.1 EAD goals and recommendations 43
6.2 FAT ML principles 44
6.3 Interview Questions 45
6.4 Reflection on the model 47
6.5 Final Assessment Model 50

This page is intentionally left blank.

"Success in creating AI could be the biggest event in the history of our civilization. Or worst. We just don't know."

Stephen Hawking

1 Introduction

Organizations enlist recruitment processes to identify and hire talented, skillful, and competitive workers. With the shift of economies into technologically progressive and knowledge-based ones, organizations are recognizing the strategic significance of human resources. Talented and skillful knowledge workers are essential to respond to the rapid technological changes as well as changing customer demands; hence they are regarded as the most valuable asset of organizations (Barney & Wright, 1998). Koch and McGrath (1996) have adequately discussed the importance of human resources conducive for long-term competitiveness and innovation potential which are essential for organizational growth. In addition, studies on organizational behavior have also highlighted the role of recruitment in boosting the morale of the entire organization (Gilliland, 1993). Therefore, the "effectiveness" of the recruitment process, i.e., to employ the right and relevant knowledge worker for a firm, is critical for organizational success.

Besides the organizational value the recruitment process ought to deliver, it has greater significance embedded in the norms that it should ensure societal welfare. The access to employment is essential for financial independence, professional development, and the autonomy of human beings and hence employment is considered a fundamental right (2012/C326/02,2012, article15).

Also, individual's job choices define his or her social status (Hollingshead et al.,1975); so by facilitating and influencing the job choices of individuals, the recruitment process not only affects the well-being of the individuals themselves but also the group they identify with. Therefore, an efficient recruitment process is in the best interest of society as a whole.

The rise of big data analytics and Artificial Intelligence (AI) decision-making has influenced many of the business operations such as sales forecasting, supply chain management, accounting; and the recruitment process is no exception. Recruitment decision-making is one of the emerging business use cases of AI. According to a recent survey, around 43 percent of the recruiters are using one or the other kind of AI tools in their recruitment process (Human Resources Professionals Association, 2017). The exponential growth of computing capacities has also fed into rapid adoption of AI tools in recruitment. Therefore, the effectiveness and efficiency of the recruitment process can be argued to be determined by the effectiveness and efficiency of Al tools used in the recruitment process.

1.1 Research Problem

The conventional paper-based recruitment process is a time-consuming process. It has undergone several transformations with the developments in information technology. In the initial phase of this transformation, the capabilities of the internet and multimedia were leveraged to standardize and digitize the labor-intensive tasks in recruitment. Tools such as the career portal, online job boards, and professional social networks have enabled applicants to interact with organizations, and vice versa, independent of time and geographical constraints. Data-driven or AI decision-making is the latest addition to this series of changes.

AI in recruitment has many clear benefits like time and cost saving by automating the processes, attracting passive job seekers, and offering improved candidate experience. It is also perceived to reduce unconscious recruiter biases such as in-group bias—favoring candidates from their group—and negative stereotypes thereby promoting workplace diversity. However, many of the recent studies show that AI decisions are not unbiased as it is perceived to be.

Investigative study on COMPAS, an AI system used to predict the defendant's recidivism in the judicial courts of United States (U.S.), concluded that the system mispredicted the recidivism of black defendants at a rate twice than that of white (Angwin et al., 2016). Research on advertisements broadcasted by Google's personalized advertising AI-system found that women were hardly served with high paying job advertisements in comparison with men (Datta et al., 2015). These instances indicate that AI-systems can reintroduce and exaggerate the foregone so-

cial biases. Therefore, the fairness of AI decisions can be of concern.

Given the significance of the recruitment in promoting organizational capabilities and societal welfare, a biased decision in recruitment has far higher stakes than an incorrectly predicted sales forecast; the risks of a biased AI system might be higher than the unintentional human biases since such systems would continuously discriminate certain classes in the society. The price of the discrimination may be that certain classes are not offered the right to employment at all and/or talented and competitive individuals falling in these protected classes are overlooked while selecting the recruits. Therefore, organizations should be cautious while adopting AI technology into the hiring process and must take responsibility in ensuring the AI systems used are fair.

However, only a few scientific researchers have addressed this technology-mediated evolution of the recruitment process. Hence there is a lack of knowledge on how to evaluate the fairness of AI recruitment system. Addressing these will offer a better understanding of the extent to which the AI recruitment tools can be unfair and the implications for the designers of such tools to develop a fair application or service.

1.2 Research Objective

Since the top-level HR managers are the customers in this market ecosystem, by critically examining the fairness of the AI-system and demanding a human-centered system, they can force the developers to design reliably fair systems. Therefore, this research aims to deliver a practical tool that would assist the top-level HR managers to describe and co-design a fair AI-recruitment system. The primary objective of this research is

> To design a conceptual model, to assist top-level HR managers in assessing the fairness of AI recruitment tools, by analyzing the possible value conflicts in the adoption of AI into the hiring process.

1.3 Research Relevance

1.3.1 Practical Relevance

From an organizational point of view, the fairness of recruitment is instrumental in building the brand reputation. Secondly, de jure EU laws, any allegation on hiring discrimination puts the burden of proof on the organization. Above all, the General Data Protection Regulation

(GDPR) article 35 mandates that the organization carries out an impact assessment while introducing automated processing in new contexts or purposes for the safety of the data subjects. So, the assessment model developed in this research is relevant for the organizations that use or intend to use AI recruitment tools. Furthermore, due to the open sourcing of many machine-learning frameworks such as Tensorflow, Keras, and Apache MXnet, many user profiling start-ups are sprouting up. By identifying the critical values that should be incorporated into the AI systems in general and recruitment tools in specific, the study also contributes towards the responsible design and governance of AI tools.

Finally, due to the competitive pressure and concerns on reputation, the organizations around the globe are adopting different AI systems. Even though their intentions to eliminate human mistakes should be appreciated, the lack of knowledge on the possible biases of machines might (unknowingly) negatively impact the society. So, by discussing the AI biases in the backdrop of recruitment, this research would also raise awareness of the risks in AI among the organizational actors. Additionally, the implications and discussions presented in this thesis can contribute to the ongoing political debates of fairness in AI-recruitment.

1.3.2 Scientific Relevance

Many studies have discussed the organizational recruitment fairness. Similarly, many researchers have addressed the concepts related to the ethical design of AI. However, no studies, to the knowledge of the author, until now has put these concepts together to describe fairness in AI recruitment. This research remedies this shortcoming by offering a synthesized account of the concepts related to the fairness in recruitment and AI.

Secondly, this study follows a Design Science Research (DSR) methodology. Though it is widely used in the design of Information System (IS) artifacts, (as will be discussed in Chapter 2) it is not limited to the IS artifacts. However, only a few studies have used this methodology in designing artifacts other than an IS. Therefore, this study could serve as a use case of DSR for non-IS artifacts.

1.4 Report Outline

The report will be organized as follows:

Chapter 2 discusses the research methodology. At first, the scientific research methodology followed in this research is detailed. Following this, it introduces the research questions and

data gathering methods that shall guide this research to achieve its objective. Finally, the chapter concludes with a detailed design of this research.

Chapter 3 discusses the domain of organizational recruitment. Initially, the chapter discusses the recruitment process and the actors involved. Finally, this chapter surveys the organizational justice literature to describe fairness in recruitment.

Chapter 4 focuses on the problems in adopting AI into recruitment. At first, the chapter provides a background of AI decision-making. Following this, it discusses potential causes of unfairness in AI recruitment systems.

Chapter 5 designs the conceptual model for assessing the fairness of AI recruitment tools. At first, it discusses the values in the responsible design of AI. Further, it discusses the legal side of automated decision-making. Finally incorporating and extending the values and norms, it presents the assessment model.

Chapter 6 focuses on the evaluation of the conceptual model. Initially, the chapter discusses the criteria for evaluating the conceptual model. By discussing its alignment with other frameworks and interviewing the practitioners, this chapter validates the designed model. Finally, from the results of the evaluation, the chapter presents the final assessment model.

Chapter 7 concludes the research by answering the research questions. Further, it reflects on the implications and limitations of this thesis. Finally, it discusses how this research aligns with the Management of Technology curriculum (author's curriculum of study).

This page is intentionally left blank.

2

Research Methodology

The previous chapter discussed the research problem and research objective. A suitable scientific research design would keep the research focused on the objective. This study follows the Design Science Research methodology.

Section 2.1 discusses the DSR methodology and explains the motivation for choosing this methodology. In section 2.2, answerable research questions are introduced to meet the research objective. Section 2.3 discusses the data collection strategies required for this research. Finally, section 2.4 presents the research design.

2.1 DSR Methodology

DSR is a scientific research methodology that focuses on developing new artifacts. Though this methodology is mainly used in the design of Information Systems (IS), it is not limited to IS and computer engineering fields. By definition, DSR is a "research activity that invents or builds new, innovative artifacts for solving problems or achieving improvements"(Iivari & Venable, 2009). Hence, DSR creates new knowledge by developing and/or analyzing a unique artifact that solves a real-world problem. The output of DSR methodology or the artifact can take four

different forms, i.e., it can be concepts, models, methods or instantiations (March & Smith, 1995). Concepts are the vocabulary that describes the problem domain and models explain the relationship between different concepts. Methods provide a guideline to use the models to solve the problem and instantiations create a working instance of the solution.

2.1.1 Why DSR?

This research utilizes the DSR methodological framework for the following reasons.

Firstly, this study aims to solve a practical problem in organizational recruitment by designing a fairness assessment model for AI recruitment systems. By the process of design, this research also intends to create new knowledge on both algorithmic and recruitment fairness. A methodology that would permit to achieve these goals is embedded in the principles of DSR methodology, i.e., "learning through building an artifact".

Secondly, the central theme of this research is to conceptualize an IS, i.e., AI recruitment tool(s) and provide a value orientation for its application. The validity of DSR in IS literature makes this methodology an apt choice. Finally, unlike other design methodology, DSR does not assume a client collaboration and DSR can be used to solve a general problem Iivari & Venable (2009). This flexibility is essential for this research since it attempts to solve a general problem in today's organizational recruitment process.

2.1.2 Practical DSR

According to Hevner (2007), the environment, knowledge base, and design spaces in design research are connected by three cycles of research namely, relevance cycle, rigor cycle, and design cycle. Figure 2.1 illustrates the cycles in DSR.

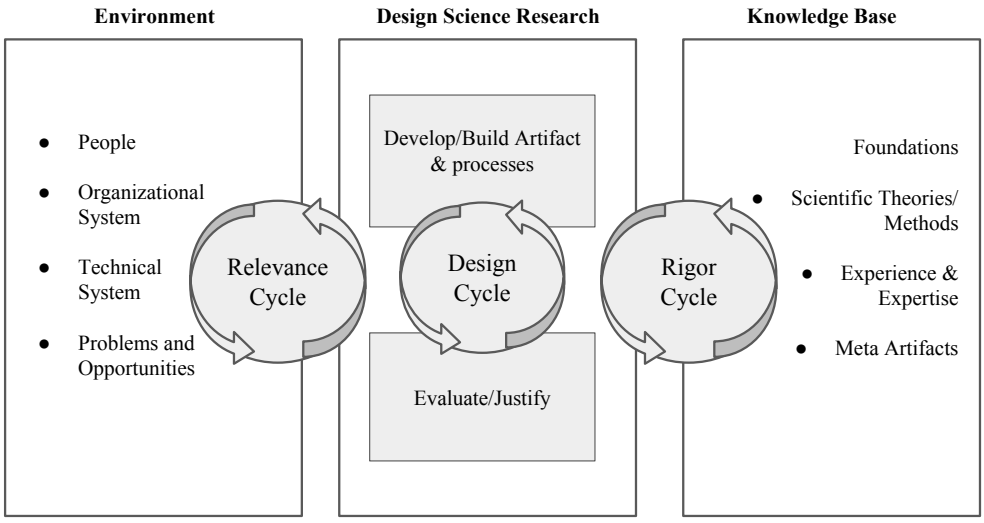

Figure 2.1: Design Science Research Cycles: Different cycles (circles) embodied in the DSR overlaying the different spaces (boxes) of design research - (copied from Hevner (2007))

In the relevance cycle, the researcher observes the application environment and identifies the problems and opportunities in the environment. This cycle defines both the application requirement and the evaluation criteria for the final artifact. The rigor cycle draws the theories and engineering methods from the existing literature (knowledge base), and this serves as a foundation for the design cycle. Finally, based on the relevance and rigor cycle the researcher builds and evaluates the artifact in the design cycle.

Expanding on Heavner's cycles and spaces of DSR, Johannesson & Perjons (2014) outlined five major activities (as shown in figure 2.2) involved in a DSR. The initial phase involves a detailed investigation and analysis of the practical problem that could be addressed by an artifact. Usually, in-depth problem explication requires an extended case study strategy. In the next phase, the requirements of the artifact are collected from the application environment. In-depth literature research and/or stakeholder interviews can be used to collect these requirements. In the following stage, the artifact is developed based on the defined requirements. The artifact development draws ideas from the existing knowledge base and uses creative methods to build new knowledge. In the subsequent phase, the feasibility of the artifact is demonstrated to the stakeholders. Finally, the artifact is evaluated and refined until it meets the requirements.

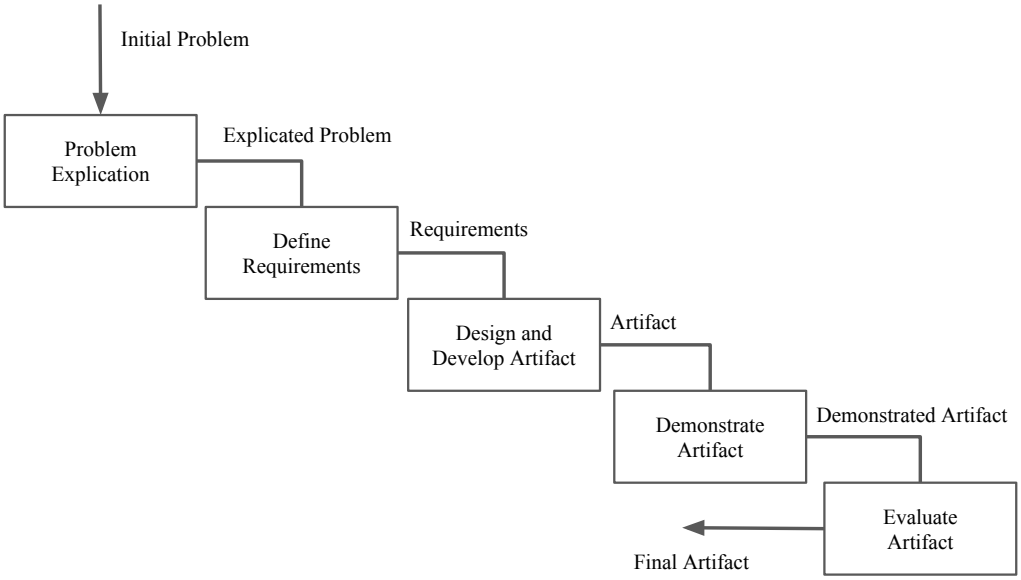

Figure 2.2: Different activities involved in DSR (copied from Johannesson & Perjons (2014))

However, these activities are not always followed in sequential order. Depending on the nature of the project, the researchers can enter DSR from different points (as shown in figure 2.3).

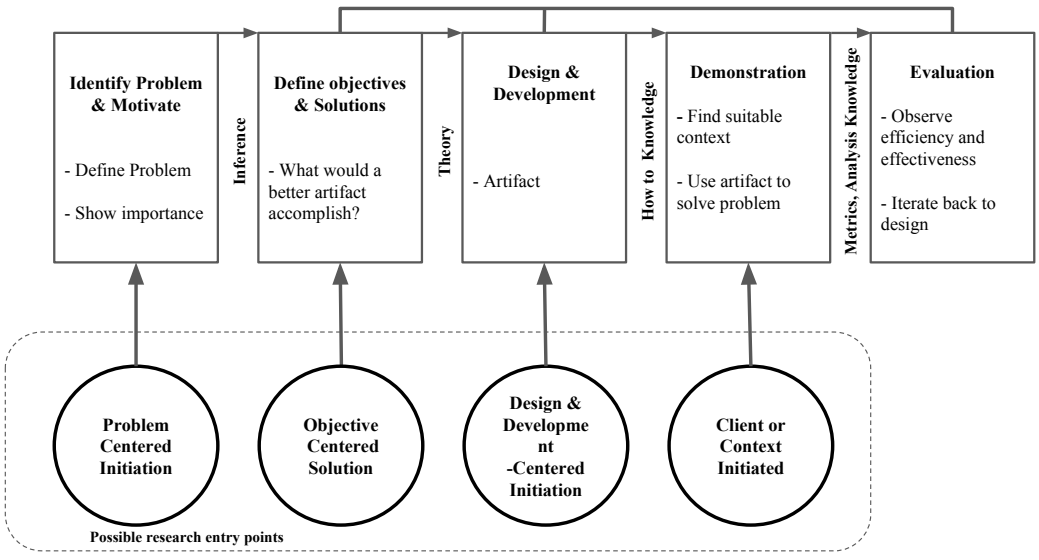

Figure 2.3: Different entry points (circles) in DSR (adapted from Peffers et al. (2007))

Based on the starting point, the DSR approaches are categorized into four; namely, problem-centered, objective-centered, development- centered and context- initiated (Peffers et al., 2007). This research follows an objective-centered approach. That is, this research starts by acknowledging the problem and recognizing that it could be addressed by designing an artifact. Further, the requirements explication and development of the artifact forms the core of the research process. Finally, as discussed in Johannesson & Perjons (2014) feasibility and evaluation of the artifact are done on a smaller scale.

2.2 Research Questions

Returning to the research objective, four answerable research questions are formulated. The research questions are grouped to align with the three cycles in DSR.

2.2.1 Relevance Cycle

RQ1 - What are the different elements of fair recruitment and selection process?

Describing fairness in recruitment and selection would be the initial phase of exploring the problem environment. By understanding the fairness in organizational recruitment, this phase intends to define the assessment goals of the model.

RQ2 - How do the inherent characteristics of AI decision-making process make the recruitment system unfair?

Once the fairness in organizational recruitment is operationalized, the next step is to understand the problems that entail the new technological transformation. By studying the AI decision-making process in recruitment, the value misalignments could be discerned by comparing it with the goals identified previously.

2.2.2 Rigor Cycle

RQ3 - What are the existing frameworks/principles that focus on responsible design and governance of AI?

Knowledge of the existing frameworks in the responsible design and governance of AI is essential for this research because these frameworks serve as a foundation for the model developed in this research. Further, a better understanding of the existing knowledge would also help to identify the supporting contributions of the model.

2.2.3 Design Cycle

RQ4 - By better understanding the fairness in organizational recruitment, how can the existing frameworks be extended and integrated into a conceptual model that can assess the fairness of AI recruitment tools?

The understanding of fair recruitment process and responsible design and governance of AI can be integrated to design a conceptual model. Further, it is also necessary that the developed model is demonstrated and evaluated. Convincing evaluation of the conceptual model would also help to justify its utility and substantiate the contribution to the knowledge base.

2.3 Data Strategies

By detailing the research phases and providing investigative directions, the DSR methodology essentially guides the general research process. However, this does not provide any information on the data strategies required in building the artifact. Analyzing the research strategies in seven mainstream Management of Information System (MIS) journals over a decade, Palvia et al. (2004) has outlined a list of fourteen rigorous data strategies. From this list, five data strategies align with the research questions.

2.3.1 Library Research

In the library research, the researcher reviews the existing academic literature to get a broader understanding of a topic. This method summarizes a few critical conclusions in a specific area of research. In this research, the operationalization of fairness in the recruitment system requires a broader knowledge on process and actors (RQ 1). Moreover, to identify the potential value conflicts in the design of AI, an in-depth understanding of AI decision-making process is also necessary (RQ 2). Therefore, academic databases such as Google Scholar and Scopus would be used to gather academic literature on these topics.

2.3.2 Literature Analysis

According to Palvia et al. (2004) literature analysis is different from the library research as the former analyzes and extends the existing literature rather than the summarizing it. The objective of this research is to extend the existing frameworks in the value-oriented design of AI to address a critical use case of AI (recruitment). Therefore, literature analysis would aid the process of

analyzing the useful frameworks and extending it to a novel conceptual model (RQ 3). Even though the approach is different from library research, the data collection process would be the same.

2.3.3 Speculation/Commentary

In the speculation or commentary, the researcher collects the data from articles that are hardly based on empirical evidence but reflect author's experience and knowledge in the field. Such speculating articles usually signal the new trends in technology and its management. As this research focuses on managing the risks involved in a rapidly evolving technology (AI), visionary opinions and speculations on AI risks and its management would aid in understanding the possible value conflicts in the technical design (RQ 2) and requirements of the conceptual model (RQ 1). Therefore, this research would use online blogs such as medium, CIO, and a few others to derive some of the practical and far-sighted insights.

2.3.4 Framework/Conceptual Model

Frameworks and conceptual models are useful methods to advance the theory in the disciplines that lack it. Following the DSR approach this research aims to deliver an artifact that would assist the process of assessing the fairness of AI recruitment tools. Here the artifact materializes as a conceptual model (RQ 4).

2.3.5 Interviews

Interviews are a primary data collection method, i.e., the information on the area of interest is directly collected from the respondents (Sekaran & Bougie, 2016). Since the demonstration and evaluation phases in DSR requires direct input from intended users of the artifact, the interviews method would be ideal for this research. Therefore, in this research, the top- level HR managers are interviewed to evaluate the developed conceptual model (RQ 4).

2.3.6 Secondary Data

Secondary data refers to the information that is published by organizations; it includes business case studies, web archives, government, and NGO reports. Many governmental bodies and NGOs have issued official reports on the ethics and governance of AI. This research would utilize

information from such reports to discuss both alignment and contributions of the final artifact (RQ 4).

2.4 Research Design

Figure 2.4 illustrates the design of this research. The research activities are as follows.

1. Initial problem - This study initiates with a general problem definition, i.e., the top-level HR managers lack a scientific tool to assess the fairness of AI recruitment tools. The problem is informed by specific practical cases of AI unfairness in recruitment (Datta et al., 2015; Florentine, 2018).

2. Requirement definition - The requirements for the artifact will be mainly defined by reviewing the literature on recruitment fairness, AI decision, and algorithmic bias. Speculation and commentary on algorithmic risks would also be included in the requirements. Further, the foundational concepts for the assessment model will be derived from the literature analysis of the topic, design, and governance of AI.

3. Design and Development – The foundational concepts will be further extended to meet the defined requirements and design the assessment model.

4. Demonstration and Evaluation- Firstly, the model will be validated by positioning and comparing it with other frameworks that has similar objective. Additionally, the model will be evaluated by potential users.

5. Final Artifact-The shortcomings from the evaluation will be accommodated into the model to produce the final artifact.

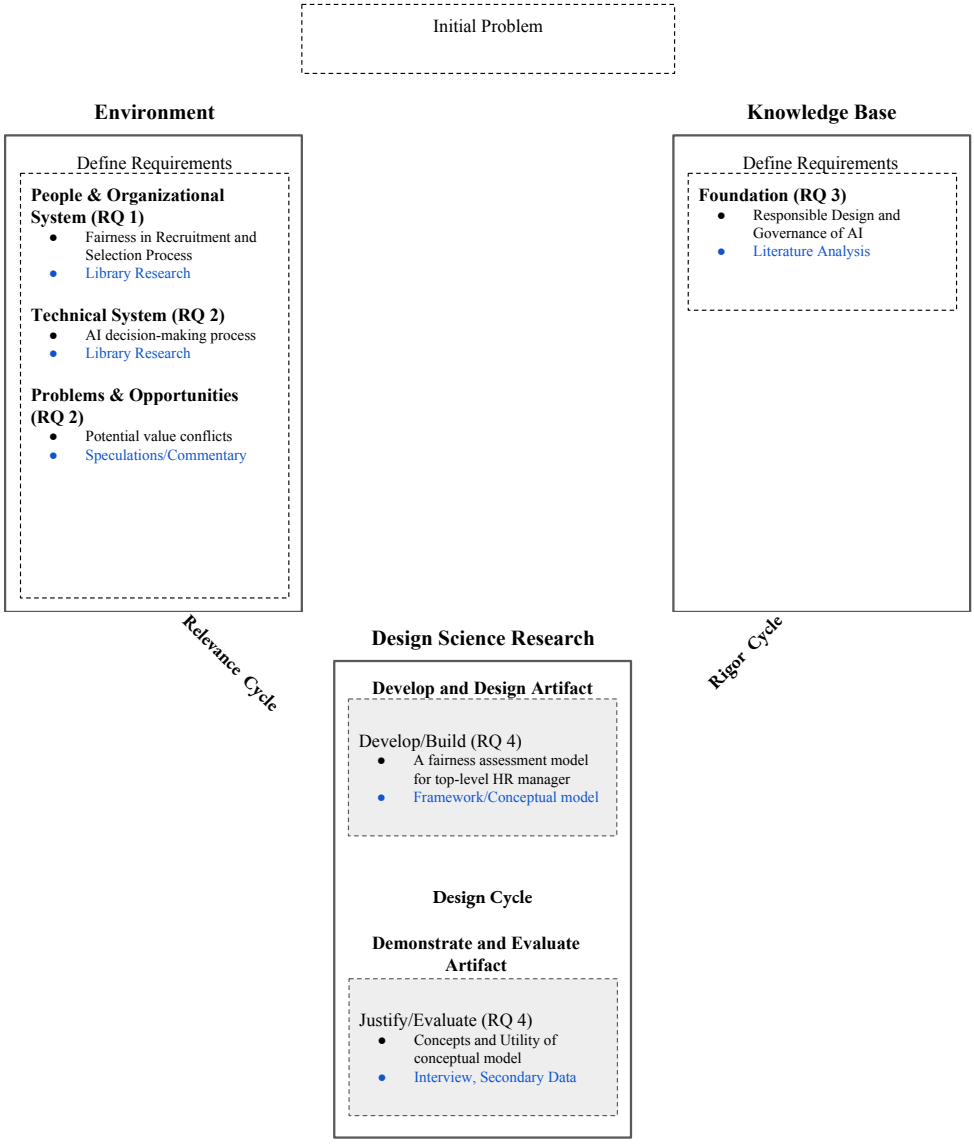

Figure 2.4: Research Framework including design spaces(solid boxes), research activities (dotted boxes), research questions (parentheses) and data strategies (blue font).

This page is intentionally left blank.

3

Organizational Recruitment and Fairness

This chapter discusses the background of organizational recruitment and details fairness in the context of organizational recruitment. Section 3.1 discusses organizational recruitment and the different actors involved in the process. Section 3.2 describes what fairness means for the different actors in an organizational recruitment context. Finally, section 3.3 summarizes the fairness in recruitment.

3.1 Organizational Recruitment

Organizational recruitment encompasses all activities carried out by organizations to identify and, attract potential employees (Barber, 1998). It is a complex process, and different models are used to explain it. According to Barber (1998), the recruitment process has three phases, i.e., generating applicants, maintaining application status, and influencing job choices. These phases highlight the three critical activities in the process, i.e., planning, communicating, and decision-making.

Breaugh & Starke (2000) describe the organizational recruitment and selection process with a five-stage model. Initial stage focuses on setting a recruitment objective by assessing the talent

requirements in the organization. In the following stage, the recruitment strategies are planned. The strategies include decisions on recruitment channels, the recruitment message (job announcement), and selection activities to follow. The third stage is the execution stage where the different recruitment activities are carried out. The recruitment decisions are obtained in the fourth stage, and the final stage evaluates the recruitment process to improve the future processes.

Synthesizing these phases and stages, (Holm, 2012) provided a comprehensive overview of the steps and tasks involved in the process as in Figure 3.1

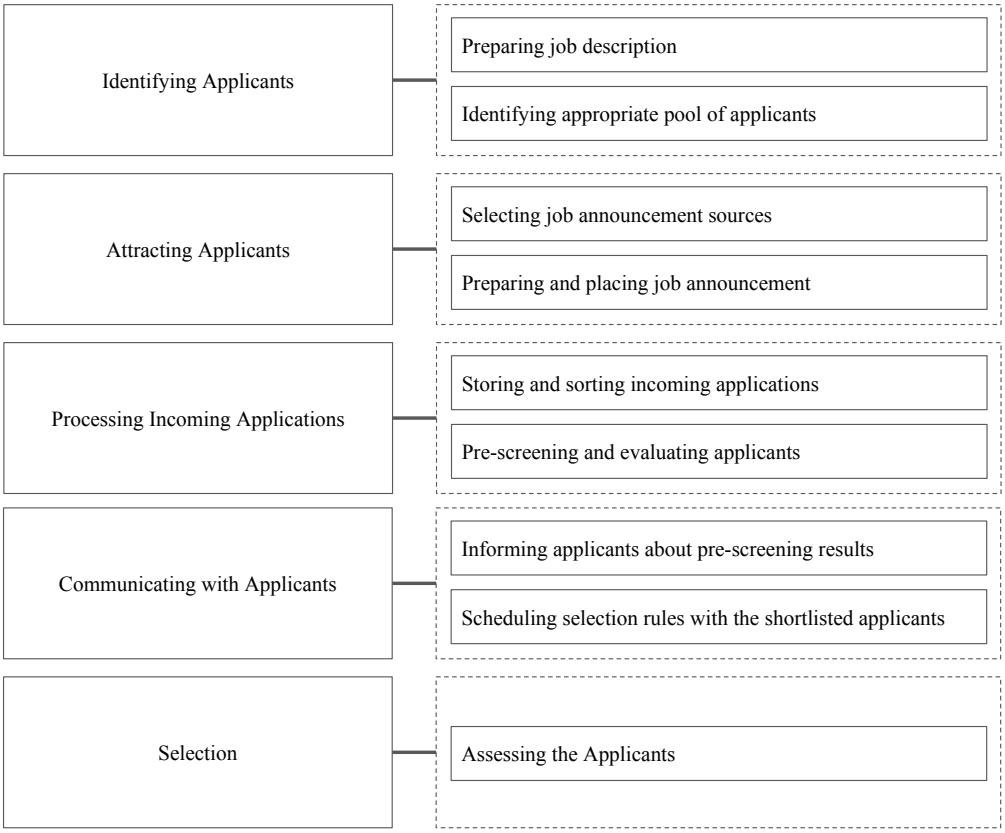

Figure 3.1: Tasks and sub tasks involved in organizational recruitment

Identifying Applicants: It involves, firstly, describing the requirements of the job by systematically collecting and analyzing data about a particular job position. Additionally, based on the job description, the task also includes identifying the pool of applicants to target.

Attracting Applicants: This task focuses on the different elements of the job posting. HRM literature discusses the importance of the content in recruitment messages (job advertisements)

in drawing the attention of the potential candidates. The credibility and comprehensiveness of the message are significant factors of attraction. The comprehensiveness of the message also depends on the recruitment sources and platforms used to convey these messages. The internal sources include employee referrals, considering former employee or previous applicants while external sources include advertisements, job portals, and social networking sites.

Processing incoming applications: The primary tasks involved are storing, sorting, and pre-screening of the incoming applications. Pre-screening is the process of eliminating unqualified applicants by evaluating the knowledge, skills, and abilities of the applicants.

Communicating with the applicants: This task involves communicating the pre-screening results to the applicants and scheduling further selection processes with the shortlisted applicants.

Selection: The primary objective of the selection process is to identify the candidate who could be successful in a specific job position. Hence the candidates are assessed on parameters of person-job and person-organization fit. Depending on the skills and abilities required to thrive in the job position, the organization uses different assessment techniques. The standard assessment techniques are interviews, personality tests, cognitive tests, and work samples.

3.1.1 Actors

Many actors involved in the recruitment process, the primary and direct actors are organizational agents (HR decision-makers or managers) and applicants. Organizational agents perform different activities on behalf of organizations to influence the job choices of the applicants. Applicant's positive reactions to these activities are critical for the success of the process. However, the applicants are not the only ones that react to the process. Since maintaining public relations is another principal aspect of the recruitment and selection, outsiders or society also plays a vital role in the process (Barber, 1998). A negative perception of by the society would impact the behavior of consumers and investors, and it would also affect the organizational attractiveness for future applicants. Therefore, the recruitment practices have to align with the values of both the applicants and the outsiders.

3.2 Fairness

Fairness is a ubiquitous element of society, i.e., in all social exchanges, the social actors are attentive to the fairness of the events (Tabibnia et al., 2008). Many theories have attempted to explain

fairness. For instance, social exchange theory, one of the earliest theories on fairness, argued that the uncertainty in trustworthiness among the societal actors makes the society focus on fairness (Emerson, 1976). Tyler and Linds's relational model 1992 states that individuals continuously seek signals to define their social status and fairness is the most dominant signal for the status. By an experimental study, they also showed trustworthiness as a confounding factor. Finally, uncertainty management theory argues that society values fairness because it helps individual to manage their uncertainties in trustworthiness, morality, status, and goals (Van den Bos & Lind, 2002). So different actors may assess fairness differently (Tyler & Lind, 1992). This section discusses fairness from the perspective of external actors involved in recruitment.

3.2.1 Applicant's Fairness

Applicant's perception of is driven by their desires of equitable treatment. Adam's Equity theory 1965 is one of the prominent theories that describe how individuals perceive fairness from the outcome of the decisions. According to the theory, individuals compare their input with the outcomes; the perceived input-output equity determines the fairness of the decisions. In the context of hiring, the applicants perceive fairness of the outcome by comparing their knowledge, skills, and efforts to the recruitment decision, i.e., if the applicant is hired or not.

However, due to the information asymmetry in recruitment, the applicants often substitute distributional justice with procedural justice. The instrumental view of procedural justice treats fair procedures as a medium of achieving fair outcomes (Thibaut et al., 1973). In this view, Leventhal et al. (1980) discussed six principles of procedural justice, i.e., correctability, consistency, accuracy, ethicality, representativeness, and bias suppression. Adapting these principles into the recruitment context, studies have discussed four characteristics of procedural justice, i.e., job relatedness, consistency, opportunity to perform and reconsider, and the objectivity of (Arvey & Renz, 1992; Arvey & Sackett, 1993; Schuler, 1993; Gilliland, 1993; Van den Bos et al., 1997).

The validity of the recruitment systems is assessed based on job-relatedness (Arvey & Renz, 1992). For content validity, the recruitment procedures should measure the features that are necessary for the job, and criteria related validity the measures should predict an individual's skills and capabilities for the specific job (Messick, 1998). Studies have shown that work sample tests, interview, and cognitive ability tests are perceived to be fair because these tests generally measure job-related features.

Consistency refers to the uniformity of procedures across people and time. The consistent procedures bring the principle of equality into the recruitment procedures and therefore are

considered as (one of) the main criteria for the procedural justice (Gilliland, 1993). In the case of recruitment, the non-uniformity is identified by applicants from the previous experience or discussions with other applicants; non-uniformities in recruitment procedures are perceived to be unfair.

Opportunity to perform addresses the applicant's capability to influence the decisions. Applicants "perceive greater control" over the decisions if they have enough opportunity to demonstrate their knowledge and skills during the selection process. Studies show that personality tests are perceived to be less fair due to the lack of control over the results (Smither et al., 1993). Opportunity to reconsider focuses on the applicant's capacity to control the process after the decisions are made. Freedom to review the scores and repeat the process improves the perceived fairness of the procedures.

Objectivity focuses on the attributes used in the selection process. Arvey & Renz (1992) discusses many different types of information that is perceived to be unfair in the recruitment decision; examples are, information irrelevant for the job, information that intrudes into the privacy of the applicants, and information that can be easily faked. Therefore, an objective recruitment system should avoid considering such information in the decision-making process.

The interactional theory of procedural justice, emphasizes the perception of fairness formed from the interpersonal considerations and information received throughout the procedure. These criteria are different from the instrumental criteria. Colquitt and Chertkoff's empirical study 2001 has also provided sufficient evidence to illustrate the impact of interactional justice on perceived fairness.

Interpersonal justice focuses on the interactions with the organizational entities. According to Bies & Moag (1986), applicants discern fairness based on the respect, dignity, and honesty experienced during the interactions with the organizational agents. The applicants deduce many unknown organizational attributes from their interactions with the recruiters. For instance, recruiter's treatment of minorities and females may signal how the organization values different societal group (Connerley & Rynes, 1997).

Informational justice, on the other hand, focuses on the different aspects of the information shared during recruitment. According to (Shapiro et al., 1994), timeliness of feedbacks, clarity, and understandability of the content are the main factors for informational justice. In the context of recruitment, understandable procedures, honest feedback, and timely information are essential for perceived fairness of the system (Gilliland, 1993).

3.3 Outsider's Fairness

Rawl's theory of justice 1971 suggests that when the society is positioned behind the "veil of ignorance", i.e., without any knowledge of their characteristics, then it would only allow social inequalities under the condition of equality of opportunity. However, according to Van Dyke (1975), the theory of justice does not account for group justice. Groups demand what they consider as fair for themselves as collective entities. In that view, outsider's (groups) perceive fairness differently from applicants (individual).

EU has implemented many regulations regarding such group fairness."The Treaty on the functioning of the European Union" prohibits discrimination on the grounds of gender, race, religion, age, sexual orientation, and physical disabilities. "EU 2000 anti-discrimination" directive rejects any justifications for such direct discrimination. Directives 2006/54, 2000/43 and 2000/78 dealing with the employment, provides an exception on the grounds of Genuine Occupational Requirement (GOR) (Fribergh & Kjaerum, 2011). GOR allows differential treatments in four cases (Speekenbrink, 2012), where two cases support the concept of group faireness i.e, differential treatment are allowed if it would counteract the existing inequalities in the society (affirmative action) and it is a necessary for a democratic society.

The EU directive 2000/43 defines indirect discrimination as "an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim, and the means of achieving that aim are appropriate and necessary". As per the legal statements unjustifiable statistical disparity above a strict threshold is considered as discrimination.

3.4 Summary

Table 3.1 the summarizes the different perspectives of fairness in recruitment system.

Characteristics	Authors
Equity	Gilliland (1993)
Merit-based selection.	Adams (1965)

Job relatedness The selection procedures are related to the job.	Gilliland (1993) Schuler (1993)
Objectivity Attributes used in the decision-making process are objective.	Arvey & Renz (1992)
Consistency The procedures are consistent across people and time.	Leventhal et al. (1980) Arvey & Renz (1992) Arvey & Sackett (1993) Gilliland (1993)
Opportunity to Perform Opportunity to demonstrate his/her knowledge.	Arvey & Renz (1992) Gilliland (1993)
Opportunity to Reconsider Opportunity to review or challenge the results.	Gilliland (1993) Arvey & Sackett (1993)
Interpersonal Effectiveness Respect and warmth in the interactions.	Bies & Moag (1986) Gilliland (1993)
Feedback Timely and informative post process communication.	(Shapiro et al., 1994) Gilliland (1993)
Understandablity The procedures are comprehensible.	Arvey & Sackett (1993) (Shapiro et al., 1994)
Statistical parity Statistical difference between the majority and the protected class getting a particular outcome is small.	Directives 2000/43

Table 3.1: Summarizing recrutiment fairness

This page is intentionally left blank.

4

Technical System and Problems

The previous chapter discussed the different criteria of recruitment fairness. This chapter discusses the potential issues when AI is adopted into the recruitment. Section 4.1 sets a general background for the AI decision-making process. Section 4.2 discusses the use cases of AI decision-making in recruitment and how it might lead to unfairness in recruitment.

4.1 AI decision-making

AI systems are computer systems that can perform tasks that usually require human intelligence; this includes activities such as visual perception, speech recognition, language processing, and decision-making. Though the current AI is far from completely replicating human intelligence, it can perform specific tasks in par with humans. Many techniques from the fields of computer science, linguistics, mathematics, neurosciences, and psychology, are used in present-day AI systems (Russell & Norvig, 2002). By discussing the popular techniques, this section provides a background of AI-decision making process.

4.1.1 Knowledge Discovery in Database

With the overwhelmingly large amount of data generated, extracting useful and reliable information has been a challenging one. Knowledge Discovery in Database (KDD) is the process of discovering hidden information in the data. KDD comprises many steps as shown in figure 4.1. Data mining is the subprocess that deals with data analysis and development of discovery algorithms to extract patterns from the data sets (Fayyad et al., 1996). "Features" and "Class labels" are two important concepts related to data sets. Class labels define the class membership of the data points in the data set while features are the independent properties of each data point that can explain the corresponding class labels. Data mining draws hypotheses from the data sets rather than testing the presumed hypotheses with the data and therefore are significantly different from conventional statistical methods (Calders & Custers, 2013).

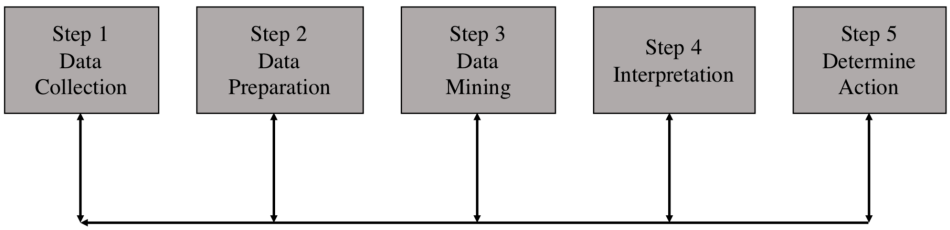

Figure 4.1: Steps in KDD copied from Fayyad et al. (1996)

Data mining is particularly useful for creating accurate data profiles. Profiling is "the process of discovering correlations between data in databases that can be used to identify and represent a human or non-human subject (individual or group) and/or the application of profiles (sets of correlated data) to individuate and represent a subject or to identify a subject as a member of a group or category" (Hildebrandt, 2008). Classification, clustering, and regression are data mining techniques that are relevant to profiling (Calders & Custers, 2013).

> In the classification technique, the data is mapped to predefined groups or classes based on the similarity in data features. In such tasks, the classes are exhaustive and mutually exclusive, i.e., each entity maps only one of the class, and there is no entity which cannot be mapped to any class.
>
> The clustering technique is similar to classification, but the difference is that it does not require a set of predefined classes and entities can have membership in more than one class. Since clustering technique is used to identify a finite number of clusters in the data, therefore it does not require a labeled data set.

Regression technique, on the other hand, is used to predict numerical values from a given dataset. In this technique, output values are presented as a function of the data features.

4.1.2 Machine Learning

Machine learning (ML) method is used to optimize the hypotheses by automating the knowledge acquisition (Langley & Simon, 1995). Though both data mining and ML have the same origins and often referred to as the same, there is a fundamental difference, i.e., data mining focuses on recognizing relationships in data while ML focuses on modeling data and optimizing it (Kononenko & Kukar, 2007). Machine learning uses many of the data mining techniques to automate and streamline knowledge discovery.

A machine's, learning process depends on two factors: feedback and knowledge representation (Russell & Norvig, 2002). Based on the types of feedback, learning is classified into four: Supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. In supervised learning, the system is exposed with some input-output pairs (labeled), and from these examples, the errors are fed back to the system to develop a mathematical function that would map the outputs for any new input. This type of learning is used in predictive tasks such as classification. In unsupervised learning, the system is fed with unlabelled data, and hence there is no feedback. This type of learning is used to find patterns in the data. Semi-supervised learning is a combination of supervised and unsupervised learning. In this type of learning, the labeled data is used to learn a mathematical function (supervised learning), the function is used to label the unlabelled data (unsupervised learning), and finally, the system learns from the entire data (supervised learning). Lastly, in reinforcement learning the feedback are implicit, i.e., either reward or punishment.

Based on the knowledge representation, there are five main paradigms in ML - namely Neural networks, Instance or case-based learning, Genetic algorithm, Rule Induction, and Analytic learning (Langley & Simon, 1995). Neural network functions similar to the neurons in the human brain, while case-based learning draws its inspiration from human memory. Genetic algorithms have their roots in evolution and rule induction is based on heuristic search while analytic learning is grounded on the formal logic. Each community has their traditions, methods, and popular algorithms. Feature extraction plays a vital role in representing knowledge internally. General ML procedures suffer from a significant limitation in this arena. Designing a feature extractor in general ML requires sophisticated engineering skills as well as a high level of domain expertise. Deep Learning solves this issue by the method of representation learning (LeCun

et al., 2015). It is a sub-field of machine learning which cuts out the human from the loop by automating the data feature extraction. Therefore, this technique plays a vital role in developing fully autonomous systems.

4.2 AI in Recruitment

As most, (if not all) recruitment process requires extracting useful information from text, speech or images, Natural Language Processing (NLP) and Image Processing (IP) is the central theme of AI in recruitment. Many studies on NLP and IP have proposed different combination learning and profiling techniques to automate the recruitment process. For instance, Faliagka et al. (2012) introduced a tool that infers applicant's personality from their Linkedin profile and ranks them for a specific job description. They used regression technique to rank the applicants and supervised learning to optimize the task. Similarly, studies in semantic web search suggest a combination of supervised and unsupervised learning techniques to improve the conceptual and contextual relevance of the web search. Table 4.1 presents some of the common use cases AI in recruitment.

Tasks	AI solution
Identifying the candidates	Semantic Search Rather than a word to word search, the AI search engines seek to improve the search accuracy by identifying the possible relations and including that in the search results. For example, if the majority of candidates currently working in customer relations previously worked at non-profit organizations, AI will add non-profit organizations into the search, broadening the search results, and the likelihood of finding the right talent*.

*https://focus.kornferry.com/talent-acquisition/3-ways-artificial-intelligence-can-improve-recruitment/

Attracting Applicants	Augmented Writing Augmented writing tools assist in writing job advertisements by suggesting the likeliness of responses and appeal of the message. For instance, *Textio* an augmented writing tool can provide insights into the gender appeal of a recruitment message[†].
Processing incoming applications	Resume screening beyond keywords Besides matching, keywords for binary criteria like experience or education AI-based resume screening software can provide other insights on other criteria. For instance, *CVviz* AI-based CV screening software can predict the cultural-fit of the candidate from his/her CV[‡].
Communicating with applicants	AI-chat bots Recruitment chatbots can provide the applicants a personalized experience during the interaction with the organization. For instance, *Mya*-a popular recruitment chat-bot- provides a personalized experience for the candidates by understanding the candidate better from their profiles[§].
Selection	Automated Video interview AI video interviewing software assess the candidates autonomously. For instance, *HireVue* an AI video interviewing software, finds the candidate that resembles the keywords, facial-expression, and tones with high-performing employees in the company[¶].

Table 4.1: Application of AI in the different stages of recruitment

Though AI recruitment tools would significantly reduce the time and effort, but due to

[†]https://textio.ai/diversity-and-inclusion-in-your-writing-4caf7c8657f
[‡]https://cvviz.com/blog/ai-for-resume-screening/
[§]https://hiremya.com/blog/6-reasons-your-candidate-experience-is-lousy
[¶]https://www.businessinsider.nl/hirevue-ai-powered-job-interview-platform-2017-8/

the inherent characteristics discussed above, AI recruitment tools may make decisions that are unfair. The following section details the argues why it is so.

4.3 Training Data

AI systems learn inductively from the training data; hence, the validity of decision-models is based on an implicit assumption that the training data accurately represents the populations. However, this assumption might not be true for various reasons (Calders & Žliobaitė, 2013; Barocas & Selbst, 2016). Firstly, in supervised learning or semi-supervised learning, incorrectly labeled training data may have a significant effect on the model. Since hiring is a subjective decision, the decision models trained on historical training data would reflect the historical discriminations that existed in recruitment. For instance, in the past, leadership was associated with masculinity, and hence men were preferred over the women for leadership roles. Learning from such data AI systems would directly and continuously discriminate women. Therefore, such data bias would conflict with the values of objectivity, consistency, opportunity to perform and would create a gender gap in the society.

Secondly, training data might be correctly labeled but might not rightly represent the society. That is, due to data sampling biases certain social groups can be over-represented (or under-represented) in the training data, making the decision model skewed (Calders & Žliobaitė, 2013). For instance, the facial recognition system that failed to detect individuals from a particular race is deemed to such sampling errors (Buolamwini & Gebru, 2018). Therefore, such biases in training data would also affect the interpersonal fairness.

4.4 Algorithmic Focus

The use of information such as gender, race, and sexual orientation are considered to be morally inappropriate and are illegal in organizational recruitments. So generally, algorithms are designed to exclude such special category data while training the decision model. However, since algorithms can discover statistically significant patterns from the data, it might find features that are correlated to special category data (proxies) and use these proxies while training the decision-model. For example, even if it is explicitly coded not to discriminate based on gender in the decision model, the algorithms might pick up the correlation between gender and the majoring subjects in high school (due to the biased focus) and discriminate individuals based on a seemingly neutral feature(Ruggieri et al., 2010; Danks & London, 2017). Hence algorithmic focus

bias would also lead to unfair and even illegal decision-making. ’

4.4.1 Opaque Decision-making

AI systems are often referred to as black-box systems due to its opaqueness (Diakopoulos, 2015). Machine learning techniques especially neural networks are highly complicated that even the designers find it hard to back-trace the output logic. Moreover, high dimensionality of input data also makes the decision-models incomprehensible. The lack of transparency of the system makes it impossible to provide honest feedback to the applicants, and it also affects the understandability of the applicants.

This page is intentionally left blank.

5 Model Design

By extending the existing knowledge on responsible design and governance of AI, this chapter develops a conceptual model for assessing the fairness of AI recruitment system. Section 5.1 discusses the design principles of responsible AI and how it translates into the design of AI recruitment systems. Section 5.2 discusses GDPR and its implications on AI recruitment systems. Finally, section 5.3 summarizes the different concepts identified to build an assessment model.

5.1 Responsible Design of AI

Most philosophers of technology agree that technology can shape and limit the actions of the users, so design responsibility has always been one of the core themes in technology ethics. Discussing the ethics of AI, Dignum (2017) proposed the three design principles for responsible AI, i.e., accountability, responsibility, and transparency. Accountability refers to the systems obligation to justify and explain their actions; responsibility refers to the capacity of the entire system to reduce the risks in decisions, and transparency aims to provide clarity to the users by describing and reproducing the decision-making process — this section, details how these principles can be used to assess the fairness AI recruitment system.

5.1.1 Accountability

Accountability refers to the agent's obligation to reason its actions and redress any harms caused by the actions. In the context of AI system, reasoning would be both, justifying the assumptions embodied in the design and explaining the decisions made by the system.

AI systems undeniably embody inherent values that root from the assumptions in the design. Seeking justification for the assumptions used in the decision model will elicit the implicit values in the design. The assumptions in the design can be both epistemic and normative (Binns et al., 2018). Epistemological assumptions refer to the assumptions made by the developer or designer based on their observations of the world(McCarthy, 1981). The concerns in the epistemological assumptions correspond to inconclusiveness, inscrutability, and misguidance that may exist in the evidence that the developer observes and feeds into the AI (Mittelstadt et al., 2016). For instance, in some recruitment systems, the developers may assume a correlation between a particular writing style and performance. Such an AI system would select candidates based on criteria unrelated to the job and make the system unfair. To justify this epistemological assumption, the developers should provide relevant scientific guarantees and statistical evidence (Doshi-Velez et al., 2017).

The normative assumptions, on the other hand, are based on the ethical principles of the developer. Suppose if the developer observes a bias in estimation (or data) premised on his/her ethical orientation, then the developer may use a biased estimator deliberately to counter this bias. Therefore, the concerns are related to the developer's understanding of fairness and its transformation into the system and their effects on the society (Mittelstadt et al., 2016). For instance, if affirmative actions coded into the system, it might yield better results in some organizations which lack diversity however in other organizations, it might lead to reverse discrimination of the applicants and statistical disparity. It would also impact the consistency and objectivity of the recruitment process. Therefore the normative justifications should focus on the bias remediation embedded in the design.

Next dimension of accountability focuses on the system outputs, i.e., explaining the decisions made by the system. Explanations or feedbacks are essential to ascertain the appropriateness of the recruitment decisions. Therefore the system should provide information on the primary features used to in the decisions, the impact of the personal characteristics of the applicant on the decision, and conditions for treating similar cases differently. However, AI in its basic form does not provide any explanation. The complexities in algorithmic decision modeling, make the system unpredictable and uninterpretable by humans. Researchers in the field of

Human-Computer Interaction (HCI) have discussed many methods for explaining the decision to the end user which can be broadly classified into two.

- Rule-extraction method: The method uses rule extraction algorithms to comprehend the decision-making model and formulate human-interpretable rules from it (Ribeiro et al., 2016). It is model agnostic (can explain any classifier regardless of the techniques used to train it) because it explains the output by approximating the decision-making model.
- Attribution method or counterfactual explanation: The method employs a contrastive explanation technique, i.e., explanations are based on the required changes in input features to achieve a different outcome. (Wachter et al., 2017b).

Since the explanations are intended to improve the understandability of the decision-making, it is also essential that the information provided is relevant for the specific group or person (Miller, 2017). For instance, explanations targeted to the recruiters might not be convincing to the applicants and vice-versa. Hence assessing explanations would also depend on the targeted user of the explanations.

5.2 Responsiblity

Accountability is a backward-looking responsibility in the sense that it applies to the consequence of action rather than actions itself. The principle of responsibility discusses the active role of both humans and AI in eliminating its adverse implications for society. Many scholars have pointed out the need for forward-looking responsibility in the development and governance of AI (Stahl et al., 2013; Russell & Norvig, 2002; Bostrom & Yudkowsky, 2014). Technology assessment literature has discussed four forward-looking dimensions of responsible innovation namely anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al., 2013). These dimensions are useful while assessing the responsibility of the AI system.

The dimension of anticipation highlights the developer's explicit attempts to foresee the negative consequences of technology and control the risks in the technology trajectory. Anticipation in technology development is about reducing the uncertainties of the system by answering 'what if' questions and considering the entire range of possibilities (Ravetz, 1997). In the case of AI, security attacks that can impact the privacy of the users and unintentional behavior are the main potential risks that have been discussed in the literature (West, 2018). Therefore, the responsible design of AI recruitment systems should at least have mechanisms to handle the potential security attacks and reconsider the individuals who have been discriminated by the system.

The reflexivity refers to self-referential critiquing (Lynch, 2000). It means reflecting on the current status of the technology which involves describing the purpose, risks and knowns and unknowns about the technology to interpret the quality (Owen et al., 2013). In the case of AI, it has dimensions in both data and the decision model. Reflexivity in data entails examining the data and recognizing the strengths and limitations of the system. By understanding the weaknesses of the model and the intended use, the organizations using the system could avoid the potential interactional unfairness.

Inclusion discusses the accommodation of diverse viewpoints into the system design, and it advocates the importance of representativeness in the development and governance of technology (Stilgoe et al., 2013). This dimension relates to group fairness, and therefore in the design of AI recruitment systems, it has two focus. First, it considers whether all the social groups are represented equally in the decisions. The representativeness of the decision models is assessed by the false positives and false negatives rates across different demographic groups. Secondly, it focuses on the diversity in the development team that would enable inclusion in the design of AI. The development team requires both human and cultural diversity to achieve an inclusive system. Here human diversity refers to the diversity based on immutable characteristics of people such as race, age, and gender, while cultural diversity refers to ethics, ideologies and working styles. Many of the reported cases of AI bias points out that such team diversity is crucial for the responsible design of AI (Guillory, 2017).

Finally, responsible innovation must also respond to the changing circumstances and stakeholder values (Stilgoe et al., 2013). Responsiveness in responsible innovation refers to the adaptability of the technology to the changing values of the society. In the field of recruitment, the demographics, nature of work and inclusion policies are continually changing. Therefore, AI systems used in the recruitment process has to adapt to such changes and requires periodic data revisions and retraining. This would ensure consistency of the recruitment system.

5.2.1 Transparency

Transparency addresses the opaqueness of the AI system. Transparency is an instrumental value, i.e., it is instrumental in achieving algorithmic accountability and responsibility accountabilityananny2018seeing. According to Burrell (2016), there are three types of opacity in algorithmic decision making, i.e., Intentional opaqueness introduced by the companies, opaqueness due to the technical illiteracy, and opaqueness due to the complexities of machine learning algorithms.

Companies that develop AI systems appropriate their competitive advantage from their pos-

session of data and algorithmic efficiency. As algorithms are just mathematical methods, protecting such methods by law is too complicated and often impossible. Therefore, the companies secure their algorithms and data by making it inaccessible to the public. Here the transparency means providing public the access to training dataset or algorithms. The effectiveness of algorithmic audits depends on the understandability of the algorithms and data. There are many best practices in the development and testing of AI that needs to be followed to enable understandability. For instance, meaningful names for the algorithmic functions, and documenting the code are some of the practices involved in software development. However, often due to the lack of experience these practices are not followed, making the system opaque to an extent. In this case, transparency refers to the legibility of codes, data, and its purpose.

The machine learning methods such as the multi-layered neural networks and convolutional networks are so sophisticated that even the developers cannot interpret the decision models or trace back the decisions to the input. Therefore, such opaqueness relates to the explanations as discussed in the accountability of AI. However, here the explanations are about providing logical explanations of the internal working of the system, it does not necessarily mean explaining each decision.

5.3 GDPR and AI

Since a fair AI recruitment should also be compliant to the legal norms, this section reviews the legal landscape of AI.

GDPR * is a binding rule that gives effect to the fundamental freedom for personal data protection and privacy to all individuals within EU. The regulation is grounded on seven primary principles: i) lawfulness, fairness, and transparency; ii) purpose limitation; iii) data minimization; iv) accuracy; vi) storage limitation; vi) integrity and confidentiality; vii) accountability. Though this regulation is intended to govern the personal data, it has provisions for automated decision-making and profiling. This section discusses these provisions.

GDPR defines profiling as "automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements"(The European Parliament and the Council of the European Union, 2016, p.33). Article 22 discusses

*https://gdpr-info.eu/

the regulations regarding profiling. The article gives data subjects the right to reject any significant decisions (like legal trials, online credit applications, and e-recruitment) that are entirely based on profiling and obtain a certain level of human intervention in the autonomous decision-making systems. So responsible design of AI recruitment tools should ensure that there are provisions for reconsideration and human interventions.

Article 13-15 addresses the opaqueness of automated decision making and profiling techniques. These articles give data subjects the right to get "meaningful explanations of the logic involved" in the automated decision making. According to Wachter et al. (2017a) this regulation would only mandate ex-ante explanations, i.e., explanation regarding the decision models rather than the rationale of a particular decision (ex-post explanations). Furthermore, GDPR also mandates explicit consent from the data subjects while processing the special category personal data such as race, sexual orientation, religion, biometric information and so on (defined in article 9). So, legally the systems are required to provide information on the types of information collected from the applicants and its processing methods.

5.4 Assessment Model

Section 5.1 and 5.2 discussed different concepts that require attention in assessing fairness AI recruitment system. This section builds an assessment model by incorporating these concepts.

	Concepts	Guiding Questions	Recruitment Fairness Characteristics
Accountability	Justification	Is there any scientific basis for the design? Does the system implement any bias remediation?	Job-relatedness

	Explanation	Does the system explain its decisions? Are the explanations directed to the recruiters or the applicants? Are the explanations tested for the interpretability of the end-users? Does the system explicitly communicate the data requirements and the data processing methods with the applicants?	Understandability, Feedback
Responsibility	Anticipation	Are there provisions for reporting a biased selection decision? What are the recourse measures in case of the biased decisions and who is liable for it? What are the recourse measures in case of a data breach?	Opportunity to reconsider, Legal compliance
	Reflexiveness	How was the data collected? Are there any known limitations of the training data? How was the model tested?	Interpersonal Fairness

	Inclusion	What is the difference between the false positive rates and false negatives rates across the different social groups? What was the diversity in the development team? (socio-cultural and discipline expertise)	Statistical parity
	Responsiveness	What is the frequency of retraining the model?	Consistency
Transparency	Auditability	Is the training data open to the public or verified by any third-party? Are the software and the data documented? Is the algorithm open for verification?	Statistical parity

Table 5.1: Conceptual model for assessing fairness of AI recruitment system

6 Evaluation

The previous chapter presented the conceptual model for assessing the fairness of AI recruitment systems. However, the affirmations made on concepts and utility of the model needs to be validated. Therefore, this chapter discusses the evaluation of the model. Section 6.1 describes the evaluation strategy. Sections 6.2 and 6.3 presents the evaluation of the artifact. Finally, section 6.4, presents the final version of the assessment model.

6.1 Evaluation Strategy

DSR literature has discussed a wide range of strategies for evaluating the research output. Based on the type of artifact, March & Smith (1995) has discussed different criteria for evaluation. The concepts are evaluated on completeness, simplicity, elegance, understandability, and ease of use, while the models are evaluated on their fidelity with real-world phenomena, completeness, level of detail, robustness, and internal consistency. Finally, the research significance is evaluated based on the appropriateness of the artifact in achieving its purpose (Venable et al., 2012).

As the final artifact is an assessment model for fairness, two qualitative methods are used to evaluate the artifact (see figure 6.1). In the first method, the assessment model is positioned

and compared with other frameworks that has similar objective. By discussing the similarities and differences in concepts, this method aims to validate the completeness of the model. In the second method, the practitioners are interviewed to get their feedback on the model. This method primarily validates the utility.

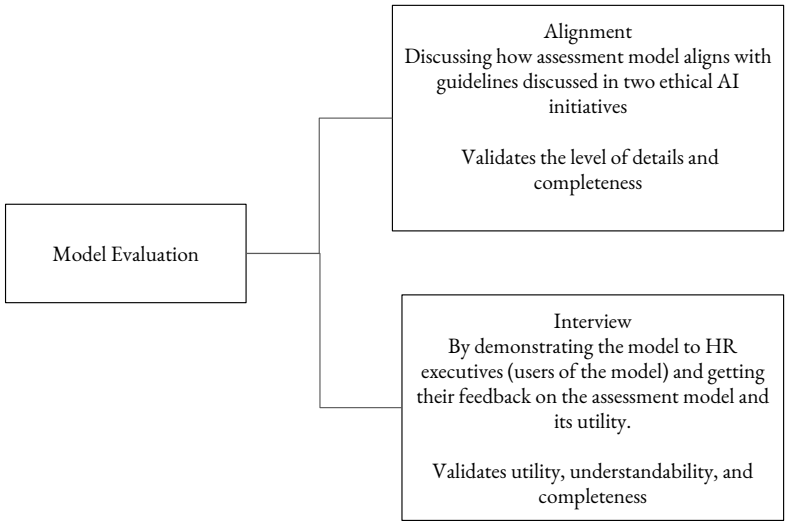

Figure 6.1: Evaluation Methods

6.2 Alignment

Due to the growing use of AI in social decision-making scenarios, AI fairness has been a topic of interest for many stakeholders. Therefore, many initiatives like IEEE* Ethically Aligned Design(EAD), Asilomar Principles, Open AI, and Fairness Accountability, and Transparency in Machine Learning (FAT ML) have put forward many guidelines for the ethical design and implementation of AI. This evaluation reviews EAD and FATML to check if the assessment model aligns with their goals and principles. EAD and FATML are chosen because these initiatives provide two different perspectives, i.e., EAD is intended to develop a standard and codes of conduct for the designer while FAT ML initiative addresses the technical challenges of machine learning.

*Institute of Electrical and Electronics Engineers

6.2.1 EAD

IEEE EAD initiative was launched in 2016. In 2017, the community published a document titled 'Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems' that put forward many recommendations for the ethical design of AI (IEEE, 2017). Table 6.1 provides an overview of these goals and recommendations.

Goals	Recommendations
Human Rights	Regulatory bodies and policies should ensure that the autonomous systems do not infringe upon human rights
Well being	Prioritize human well being on widely accepted metrics.
Accountability	Designers and developers should be aware of the diversity and cultural norms. Documenting the intended use, training data, data sources, algorithmic performance, and optimization goals of the system.
Transparency	Developing systems that are explainable. Forming agencies that can certify algorithms.
Awareness of misuse	Educate society about the potential risks of the autonomous systems.

Table 6.1: EAD goals and recommendations

The assessment model aligns with all the goals of EAD. For instance, the *reflexiveness, inclusion, auditability* discussed in the assessment model evaluates if the system infringes upon human rights like non-discrimination and equality of opportunity. The model also assesses human values such as safety, and security. Finally, the model also evaluates how the design communicates the data requirements and data processing methods.

6.2.2 FAT ML

FAT ML initiative was launched in 2014 with an objective to bring researchers and practitioners in machine learning to address the novel concerns on fairness, accountability, and transparency. The initiative has discussed these issues from a technical perspective and has outlined five prin-

ciples for algorithmic accountability [†]. These principles are discussed in table 6.2.

Principles	Description
Responsibility	Responsibility ensures that there a legal person to take the blame and redress any harms caused by the algorithm.
Explainability	Explainability is the obligation to explain the decision to the end users.
Auditability	Auditablity focuses on the ability to inspect, understand and criticize the system by any interested third party.
Accuracy	Accuracy is about identifying the sources of error in both algorithm and data to mitigate the risks involved.
Fairness	Fairness refers to the mitigation of unjust impacts across the different subgroups of the society.

Table 6.2: FAT ML principles)

The assessment model discussed in this report evaluates each of these principles. The model has discussed a few guiding questions to evaluate the explanation and auditability of the system. Anticipation discussed in the model inquiries the liability and redressal methods. Furthermore, reflection on the limitations of the data and performance of the model would discern the accuracy of the system. Finally, the inclusion discussed in the model evaluates the demographic fairness.

Apart from aligning with all the principles of FAT ML the assessment model also highlights two other dimensions, i.e., justifying the design assumptions and responding to changing demographics. Firstly, as misguided and inconclusive evidence could lead to the discrimination of applicants, the statistical or scientific guarantees for assumptions in design is significant in assessing the validity of recruitment tools. Secondly, with globalization the workplace demographics is continuously changing, systems inability to adapt would inhibit the diversity in the workplace (section 5.2.1). Therefore both the concepts are essential in assessing AI recruitment tools.

[†] http://www.fatml.org/resources/principles-for-accountable-algorithms

6.3 Interviews

The level of details and completeness of the model were validated in the previous section. However, the artifact is relevant only if it has utility in the real world. Therefore, the targeted users of the model - top-level HR executives that use AI systems in their recruitment process- were interviewed to discern the utility of the model. A total of eight (30-45 minutes long) practitioner interviews conducted in this evaluation.

The interviews were structured into three sections. The initial section focused on understanding the current process of assessing the AI-recruitment tool. The second section involved explaining the assessment model developed to the interviewees. In the final section, the interviewees were asked to reflect on different aspects of the model. Table 6.3 presents the interview questions.

Domain	Questions
Current process and tools	What were the primary considerations while selecting an AI recruitment tool? Do you use any frameworks to assess the fairness of these AI recruitment tools?
	Explaining the model
Fidelity with the real-world	In your opinion, are the concepts and the guiding questions comprehensible? In your opinion, Does this model match with the fairness you may seek in a recruiting tool? If not, what are the elements that you would add/remove from this model?

Table 6.3: Interview Questions

6.3.1 Discussion

The initial section of the interview aims to understand whether the interviewees and their companies used any formal procedures or models to assess the recruitment tools before deploying it. Instead of an in-depth understanding of the existing process, the questions were focused on indicating the utility of the assessment model.

The responses to these questions were similar. Data security and data protection were the major considerations while evaluating the AI tools. According to a respondent
Interviewee 1

"Basically there are three types of checking, of course, the functionality, IT department checks the GDPR, data security compliance [and] financial department [checks] the financial matters"

The functional assessment was focused on data presentation, and the insights tools could provide. Though many interviewees mentioned candidate experience as one of their primary motives for using AI recruitment tools, the fairness was not considered in the assessment. The interview responses also revealed that many lacked an understanding of the risks of AI. Interviewee 4

"we use [a machine learning tool] for searching candidates.... with requirement keywords, we select the candidates. How it can be unfair?"

Though a couple of interviewees were aware of AI unfairness (given the recent case of bias in amazon's recruitment system) but their understanding of the problem was also limited.
Interviewee 8

"Yes, AI fairness is a hot topic, but I think it is about the data."

Another response also indicated the utility of the model.
Interviewee 1

"No, we don't use any frameworks, but with problems of amazon AI, I think we have to consider fairness."

In the second section, the assessment model was explained to the interviewees. The general impression of the interviewees was positive. The third section of the interview focused on getting feedback on the model. The interviewees were asked to comment on the comprehensibility and utility. Table 6.4 gives an overview of the responses.

Job Titile	Understandability	Utility
Talent Acquisition Strategist	✓	✓

Talent Acquisition Manager	✓	✓
Global HR	✓	✕
Head of Talent Acquisition	✓	✓
Global HR	✓	✓
Talent Acquisition Specialist	✓	✕
Talent Acquisition Consultant EMEA	✓	✕
Head of Talent Acquisition	✓	✓

Table 6.4: Feedback on the model- The tick indicates positive response and cross indicates negative or neutral response

The negative responses were related to the ease of use of the model. Therefore the final artifact has to addresses this shortcoming.

Interviewee 6

"model is understandable, but [it is] too broad, and I think its complex. Actually, if you want the managers to use, it should be simple."

Interviewee 7

"... you can make it as a checklist"

6.4 Final Artifact

By creating a checklist with the concepts and guiding questions discussed in the previous model, the final model ensures ease of use. Table 6.5 presents the final version of the assessment model.

	Concepts	Questions	Yes	No	Comments
Accountability	Justification	Is there any scientific basis for the design of the system?			No implies that the assessment recruitment system lacks validity and hence may not be assessing on job-related features.
		Does the system implement any remediation for societal bias?			Yes implies that there is a possibility of both improving and reducing diversity. Check if it matches with the company's diversity goals.
	Explanation	Does the system provide any explanations?			No implies that the system is not transparent and it would affect fairness dimensions like understandability and feedback.
		Does the system explain individual decisions?			Yes implies that the system is transparent (understandable and provides feedback) provided the explanations are interpretable.

		Are the explanations tested for interpretability?			
		Does the system explicitly communicate data requirements and data processing methods with the applicants?			No implies that the system is not legally compliant and understandable.
Responsibility	Anticipation	Are there provisions for reporting biased selection decisions?			No implies that the system is not legally compliant and affects the fairness criteria - Opportunity to reconsider.
		Is there a legal person liable for the harms of the system?			No implies the system is not legally compliant.
		Are there measures for redressing a biased decisions?			No implies the system does not actively promote Opportunity to reconsider.
		Are there measures to redress the data breaches?			No implies that there is a concern about the privacy of the applicants
	Reflexiveness	Was the data collected for the purpose?			No implies that the data might not be representative and would affect the statistical parity and interpersonal fairness.

		Are there any known limitations for the training data/decision-making model?			If yes, check if the limitations are acceptable for the context of use.
	Inclusion	Were the recruitment/psychology experts part of the development team?			No implies that the system might not assess the applicants on job-related features.
		Were the ethical experts part of the development team?			
		Is there any test results that show parity in false negatives across different social groups?			No implies that the system might run a risk of discriminating certain social groups.
	Responsiveness	Is the decision-model retrained frequently?			No implies that the system may be inconsistent in the future.
Transparency	Auditability	Is the training data and/or public or verified by any third-party?			No implies that the system is not validated and has a possibility of being inconsistent.
		Are the software and data documented?			No implies that the system cannot be validated in the future.

Table 6.5: Fairness assessment model AI recruitment system

7

Conclusion and Reflections

Artificial intelligence is undeniably one of the dominant technology to date, and the rapid developments in the technology also confirm that it will continue to permeate our society. Currently, the AI-driven transformation of the hiring process raises many ethical issues (Barocas & Selbst, 2016; Calders & Žliobaitė, 2013; Florentine, 2018). The improper use of such systems would degrade the society and undermine trust in the technology and the organizations deploying it. Therefore, organizations adopting AI recruitment systems have to take responsibility in ensuring its fairness.

This study started with an objective that emphasizes the role of organizations (adopting AI recruitment systems) in ensuring fairness in recruitment processes. The underlying hypothesis of this research was that a fairness assessment tool would enable the organizational decision-makers to filter out the unfair AI recruitment systems and consequently, this would urge the developers to focus on societal values. Therefore the study designed and evaluated a conceptual model for assessing the fairness of AI recruitment tools by following the DSR framework.

In this final chapter, Section 7.1 will conclude the research by answering the research questions, and the rest of the report reflects on different aspects of this research. The implications

of this study are discussed in section 7.2. Section 7.3 details the research limitations and recommendations for future research. Finally, section 7.4 explains how this research aligns with the Management of Technology (author's curriculum of study) curriculum.

7.1 Answers to the Research Questions

Section 2.2 formulated four research questions which will be answered in this section

RQ1 - What are the different elements of fair recruitment and selection process?

Different actors assess recruitment fairness differently. The report discussed recruitment fairness from the two prospective applicants, i.e., applicants and outsiders (see section 3.2). The applicants discern fairness from the outcome equity, job-relatedness, consistency, objectivity, opportunity to perform and reconsider the decisions. Further, applicants also value understandability, feedback, respect, and warmth as pointed out by the interactional theory of fairness.

The outsider's focus on group fairness rather than individual fairness. Therefore representativeness of different social groups is fundamental for the fairness from society's perspective. EU laws address the representativeness by mandating a certain level of statistical parity among different social groups in the recruitment decisions. Moreover, the laws also prohibit the use of sensitive attributes such as gender, race, sexual orientation and physical disabilities in the hiring process to ensure social inclusiveness in the job market.

RQ2 - How do the inherent characteristics of AI decision-making process make the recruitment system unfair?

The discussion on KDD and ML highlighted some of the inherent characteristics of AI such as pattern recognition and inductive learning. The different paradigms in knowledge representation also underlined the complexities involved in AI algorithms(see section 4.1). These characteristics may lead to unfairness in AI recruitment systems.

Inductive learning from historical or non-representative data would lead to inconsistent recruitment decisions, and this may also affect the interpersonal effectiveness of recruitment. Secondly, though the use of specific sensitive attributes (like gender and race) are prohibited by law, by pattern recognition the AI tools might find proxies for such sensitive attributes. So, if algorithms lack focus, it would potentially discriminate certain groups in the society. Finally, the algorithmic complexities make the AI decisions opaque. It reduces the understandability of recruitment procedures and decisions.

RQ3 - What are the existing frameworks or principle that focuses on responsible design and governance of AI

Focusing on the responsible design, the ART design principle of AI discussed accountability, responsibility, and transparency as the primary design consideration of AI(see section 5.1). Here accountability refers to the obligation to justify and explain the decisions to the end users; responsibility focuses on reducing the risks in decision-making, and by describing and reproducing the decision-making process. GDPR informed the norms in automated data processing. According to the regulation human interventions and meaningful explanations are necessary while processing recruitment data. (see section 5.2).

RQ4 - By better understanding the fairness in organizational recruitment, how can the existing frameworks be extended and integrated into a conceptual model that can assess the fairness of AI recruitment tools?

By exploring the design principles for responsible AI, and General Data Protection Regulation (GDPR). The model presents seven dimensions which translate the principles to design requirements to assess the fairness of AI-based recruitment system. They are: (1)Justification; (2)Explanation; (3)Anticipation; (4)Reflexiveness; (5)Inclusion; (6)Responsiveness; and (7)Auditablity. The model also ties these concepts with specific criteria of organizational recruitment fairness such as consistency, interpersonal fairness, job-relatedness, and statistical parity (for a detailed overview see section 5.3).

Finally, the completeness of the model was evaluated by discussing its alignment with other frameworks with similar objective and utility of the model was validated by collecting feedback from the intended users (HR executives).

7.2 Implications and Recommendations

This research has contributed to a better theoretical understanding of recruitment and AI fairness. Furthermore, the research findings also have other implications.

From the practitioner interviews conducted during the evaluation phase, it was found that the people championing AI transformations in HR are not fully aware of the risks of AI. A few participants seemed to have overestimated their control over the system- which was evident from their responses- and claimed that humans made the final selection of candidates. However, the pool of candidates from which the recruiters make their choice is pre-selected by a (biased)

AI algorithm. This apparent superior control can, therefore, be deemed to be ”an illusion of control bias” as discussed by Langer (1975). With this illusion of control, the managers might unknowingly feed the bias in AI into their recruitment process. In this context, it can be expected that when the managers would start using the assessment tool, they will develop algorithmic literacy and be more aware of the risks of AI.

Corporate Social Responsibility (CSR) is an emerging field in business research and management of corporates; CSR, in general, entails that firms engage in ” actions that appear to further some social good, beyond the interests of the firm and that which is required by law” (Choi & Wang, 2009, p.117). Entertaining participation of diverse social groups and eradicating discrimination is a social good. Since firms are the largest employer of human resources in a society, actions internal to the firm, such as recruitment process, would have a broader societal impact. In particular, if firms discriminate their employees or (potential) future employees based on their background, this process undermines the goals of CSR. We have seen that the AI algorithms tend to discriminate candidates based on their diversity and can reject people from underrepresented classes of the society. Therefore, by employing the model developed in this research, which will support firms to ensure that their recruitment processes are fair, the potential discrimination that the AI recruitment tools can bring can be avoided. Therefore, there is a possibility that the firms can attract and employ more qualified candidates from diverse social groups contributing to the goals of CSR.

Finally, the artifact designed in this research is not a standalone tool to ensure fairness, a few regulations should complement it. As discussed in section 4.1, opening the data and algorithms for audit would conflict with the developers' notion of protecting their intellectual property and proprietary technology. If auditing the data and algorithms are made mandatory by governmental policy, and if the audit was to be carried out by a trusted third-party, it can be expected that the developers may be less hesitant to do so. Therefore, a recommendation would be to implement such a policy and facilitate auditing by the government or other relevant parties.

7.3 Limitation and Future research

There are a few issues in this research that can be discussed as limitations of this research.

Firstly, this study explored the recruitment fairness by reviewing the behavioral science literature; empirical research was not performed to confirm this understanding. Though it can be argued that the behavioral science literature have already discussed the perception of recruitment fairness quite widely, but these literature has addressed the fairness of the traditional re-

cruitment process with human decision-makers. Since, studies have pointed out that the people judge human agents and artificial agents differently (Malle et al., 2015), the recruitment fairness discussed in this report might digress from what people expect from an artificial agent. Future research on fairness in AI recruitment could consider crowdsourcing the perception of fairness from the actors by projecting different scenarios in AI recruitment. Such a study could extend the assessment model as well as account for the cultural differences in assessing fairness.

The second limitation relates to the evaluation of the artifact. The evaluation of DSR artifact focuses on iterative improvement. However, due to the lack of awareness on risks in AI, the interviewees (i.e., executive HR managers that use AI tools in their recruitment process) were unable to contribute to any significant improvement in the model. Though it can be attributed to the early stages of the technological use-case, it can also be seen as a limitation of the sampling. More respondents could have given a better model. An additional possibility could have been scenario workshops focusing on the unfairness of AI recruitment tools, including both the HR Manager and IT managers. Since the workshop may reconcile the distinctive perceptions about the topic, it may potentially elicit more creative input for the design, and modification of the conceptual model.

The third limitation relates to DSR methodology. The DSR methodology has widely discussed different strategy for evaluating the artifact. However, it does not provide any guideline on evaluating the design process. Though the practitioner interview confirmed the utility of the model, it was not able to provide any creative suggestions for the model. If there had been guidelines for evaluating the design process, this research could also have strengthened its claims by evaluating it. Therefore the future research on DSR methodology can address this problem by exploring strategies to evaluate the design process.

Finally, the evaluation in this research was limited to the perceived utility of the assessment model. Though this is acceptable and common in the DSR, for an extensive understanding of the usability and effectiveness of the model requires real-world testing. Depending on the tools used in different stages of the recruitment and selection process, the assessment would have to stress more on specific concepts in the model. So including these contextual factors to the model and evaluating the actual performance of the model by a case study approach and could be future research.

7.4 MoT Curriculum Alignment

The MoT curriculum is an exploration of how technology can be leveraged to advance the processes in organizations. The ethics and responsibility in the technological innovation remain a core theme of the curriculum. Aligning with the curriculum this thesis analyzed the impact of AI technology on the recruitment process, and the areas it could be improved for the benefit of the society and organizations. Encompassing the scientific methods of design, discussed in the Research Methodology course of MoT curriculum, this thesis ensures the scientific rigor. The study touches upon different topics in Social values, Inter- and Intra- organizational decision-making, and Data and Information management (ICT management specialization) courses in the curriculum and also has policy implications for it. Thus this research leveraged both knowledge and values the MoT curriculum.

References

Adams, J. S. (1965). Inequity in social exchange. In *Advances in experimental social psychology*, volume 2 (pp. 267–299). Elsevier.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias.

Arvey, R. D. & Renz, G. L. (1992). Fairness in the selection of employees. *Journal of Business Ethics*, 11(5-6), 331–340.

Arvey, R. D. & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In *Personnel selection*. Jossey-Bass.

Barber, A. E. (1998). *Recruiting employees: Individual and organizational perspectives*, volume 8. Sage Publications.

Barney, J. & Wright, P. M. (1998). On becoming a strategic partner: The Role of Human Resources in Gaining Competitive Advantage. *Center for Advanced Human Resource Studies*, 37(1), 1–25.

Barocas, S. & Selbst, A. (2016). Big Data ' s Disparate Impact. *California law review*, 104(1), 671–729.

Bies, R. J. & Moag, J. S. (1986). Interactional Justice: Communication Criteria of Fairness. In *Research on Negotiation in Organization, Vol. 1* (pp. 43–55). Greenwich.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 377).: ACM.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 316, 334.

Breaugh, J. A. & Starke, M. (2000). Research on employee recruitment: So many studies, so many remaining questions. *Journal of management*, 26(3), 405–434.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

Calders, T. & Custers, B. (2013). What is data mining and how does it work? In *Discrimination and Privacy in the Information Society* (pp. 27–42). Springer.

Calders, T. & Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society* (pp. 43–57). Springer.

Choi, J. & Wang, H. (2009). Stakeholder relations and the persistence of corporate financial performance. *Strategic management journal*, 30(8), 895–907.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of applied psychology*, 86(3), 386.

Connerley, M. L. & Rynes, S. L. (1997). The influence of recruiter characteristics and organizational recruitment support on perceived recruiter effectiveness: Views from applicants and recruiters. *Human Relations*, 50(12), 1563–1586.

Danks, D. & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691–4697).

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 1–26.

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398–415.

Dignum, V. (2017). Responsible autonomy. *arXiv preprint arXiv:1706.02513*.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Emerson, R. M. (1976). Social exchange theory. *Annual review of sociology*, 2(1), 335–362.

Faliagka, E., Ramantas, K., Tsakalidis, A., & Tzimas, G. (2012). Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*: Citeseer.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Int Conf on Knowledge Discovery and Data Mining*, (pp. 82–88).

Florentine, S. (2018). Amazon's biased AI recruiting tool gets scrapped | CIO.

Fribergh, E. & Kjaerum, M. (2011). *Handbook on European non-discrimination law*, volume 33. Publications Office of the European Union.

Gilliland, S. W. (1993). The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. *The Academy of Management Review*, 18(4), 694.

Guillory, D. (2017). Why Diversity in Artificial Intelligence (AI) Is Non-negotiable.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2), 4.

Hildebrandt, M. (2008). Defining profiling: A new type of knowledge? *Profiling the European Citizen: Cross-Disciplinary Perspectives*, (pp. 17–45).

Holm, A. B. (2012). E-recruitment: Towards an Ubiquitous Recruitment Process and Candidate Relationship Management. *German Journal of Human Resource Management: Zeitschrift für Personalforschung*, 26(3), 241–259.

Human Resources Professionals Association (2017). A New Age of Opportunities What does Artificial Intelligence mean for HR Professionals?

IEEE (2017). Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems.

Iivari, J. & Venable, J. (2009). Action research and design science research-seemingly similar but decisively dissimilar. In *ECIS* (pp. 1642–1653).

Johannesson, P. & Perjons, E. (2014). *An introduction to design science*. Springer.

Kononenko, I. & Kukar, M. (2007). *Machine learning and data mining: introduction to principles and algorithms*. Horwood Publishing.

Langer, E. J. (1975). The illusion of control. *Journal of personality and social psychology*, 32(2), 311.

Langley, P. & Simon, H. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11), 54–64.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Leventhal, G., Karuza, J., & Fry, W. (1980). Beyond fairness: a theory of allocation preferences. In *Justice in social interaction* (pp. 167–218). New York: Springer.

Lynch, M. (2000). Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture & Society*, 17(3), 26–54.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124).: ACM.

March, S. T. & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251–266.

McCarthy, J. (1981). Epistemological problems of artificial intelligence. In *Readings in artificial intelligence* (pp. 459–465). Elsevier.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1-3), 35–44.

Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.

Owen, R., Bessant, J., & Heintz, M. (2013). *Responsible innovation: Managing the responsible emergence of science and innovation in society*. John Wiley & Sons.

Palvia, P., Leary, D., Mao, E., Midha, V., Pinjani, P., & Salam, A. (2004). Research methodologies in mis: an update. *The Communications of the Association for Information Systems*, 14(1), 58.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45–77.

Ravetz, J. R. (1997). The science of 'what-if?'. *Futures*, 29(6), 533–539.

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 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* (pp. 1135–1144).: ACM.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2), 9.

Russell, S. J. & Norvig, P. (2002). *Artificial intelligence: a modern approach (International Edition)*. {Pearson US Imports & PHIPEs}.

Schuler, H. (1993). Social validity of selection situations: A concept and some empirical results.

Sekaran, U. & Bougie, R. (2016). *Research methods for business: A skill building approach*. John Wiley & Sons.

Shapiro, D. L., Buttner, E. H., & Barry, B. (1994). Explanations: What factors enhance their perceived adequacy? *Organizational behavior and human decision processes*, 58(3), 346–368.

Smither, J. W., Reilly, R. R., Millsap, R. E., AT&T, K. P., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel psychology*, 46(1), 49–76.

Speekenbrink, S. (2012). *European Non-discrimination Law: A Comparison of EU Law and the ECHR in the Field of Non-discrimination and Freedom of Religion in Public Employment with an Emphasis on the Islamic Headscarf Issue*.

Stahl, B. C., Eden, G., & Jirotka, M. (2013). Responsible research and innovation in information and communication technology: Identifying and engaging with the ethical implications of icts. *Responsible innovation*, (pp. 199–218).

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568–1580.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, 19(4), 339–347.

The European Parliament and the Council of the European Union (2016). Regulation (EU) 2016/679 (GDPR). *Official Journal of the European Union*, (pp. 1–88).

Thibaut, J., Walker, L., LaTour, S., & Houlden, P. (1973). Procedural justice as fairness. *Stan. L. Rev.*, 26, 1271.

Tyler, T. R. & Lind, E. A. (1992). A relational model of authority in groups. In *Advances in experimental social psychology*, volume 25 (pp. 115–191). Academic Press.

Van den Bos, K. & Lind, E. A. (2002). Uncertainty management by means of fairness judgments. In *Advances in experimental social psychology*, volume 34 (pp. 1–60). Elsevier.

Van den Bos, K., Lind, E. A., Vermunt, R., & Wilke, H. A. M. (1997). How do I judge my outcome when I do not know the outcome of others? The psychology of the fair process effect. *Journal of Personality and Social Psychology*.

Van Dyke, V. (1975). Justice as fairness: for groups? *American Political Science Review*, 69(2), 607–614.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *International Conference on Design Science Research in Information Systems* (pp. 423–438).: Springer.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017a). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99.

Wachter, S., Mittelstadt, B., & Russell, C. (2017b). Counterfactual explanations without opening the black box: Automated decisions and the gdpr.

West, D. M. (2018). *The Future of Work: Robots, AI, and Automation*. Brookings Institution Press.