Informatics

Artificial intelligence in recruitment: a qualitative analysis and requirements for promoting fairness

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Yi Wang, BSc
Matrikelnummer 01633407

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr.in rer.soc.oec. Sabine Theresia Köszegi

Wien, 16. Dezember 2024

Yi Wang

Sabine Theresia Köszegi

TU WIEN Informatics

Artificial intelligence in recruitment: a qualitative analysis and requirements for promoting fairness

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Yi Wang, BSc
Registration Number 01633407

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr.in rer.soc.oec. Sabine Theresia Köszegi

Vienna, December 16, 2024

Yi Wang

Sabine Theresia Köszegi

Erklärung zur Verfassung der Arbeit

Yi Wang, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 16. Dezember 2024

Yi Wang

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor for her invaluable guidance and trust throughout this journey. Her support and advice have been essential in helping me achieve this milestone.

This thesis was developed while engaged in professional work and managing various personal circumstances, which taught me important life lessons about self-care and resilience.

I extend my deepest thanks to my parents for their unwavering support. Their belief in me has been a constant source of strength. Additionally, I am deeply grateful to my husband for his love, patience, and understanding. I also want to thank everyone who provided me with emotional support.

Thank you all for being a part of this journey and for your generous support.

Kurzfassung

Die Integration von Künstlicher Intelligenz (KI) in Rekrutierungsprozesse wird oft als vorteilhaft angesehen. Trotz behaupteter Objektivität haben KI-Systeme jedoch Anfälligkeiten für Verzerrungen und Fehler gezeigt, die zu Diskriminierung führen können. Die Forschung zur Fairness in KI-gestützter Rekrutierung ist derzeit begrenzt. Zudem stehen Organisationen vor Herausforderungen bei der Auswahl fairer KI-Rekrutierungstools.

Unter Verwendung einer Design-Science-Forschungsmethodik beginnt diese Arbeit mit der Formulierung von Kriterien für einen fairen Rekrutierungsprozess und der Identifizierung potenzieller Verzerrungen in der KI-Rekrutierung. Der Fokus liegt auf den Phasen des KI-Einsatzes, die die Bewertung und Auswahl geeigneter Kandidierenden direkt beeinflussen. Die Definition von Fairness wird aus zwei Perspektiven untersucht: die wahrgenommene Fairness der Kandidierenden und die objektive Fairness in Bezug auf das KI-System. Durch eine systematische Literaturrecherche zu KI-Prinzipien und -Richtlinien werden relevante Dimensionen und Anforderungen für KI-Rekrutierungstools abgeleitet, die Fairness fördern. Rechtliche Perspektiven, wie die EU-KI-Verordnung, werden berücksichtigt. Ein Artefakt mit Leitfragen wird entwickelt, das Organisationen helfen soll, potenzielle Probleme in KI-Rekrutierungstools zu identifizieren. Die Korrektheit und Vollständigkeit des Artefakts werden durch Analysen mit ähnlichen Forschungen validiert. Zur Bewertung der Praktikabilität und Nützlichkeit des Artefakts führt diese Arbeit qualitative Fallstudien zu ausgewählten KI-Rekrutierungsanwendungen durch und identifiziert mehr Aspekte, als in der bestehenden Literatur dokumentiert sind.

Diese Arbeit leistet Beiträge sowohl im akademischen als auch im praktischen Bereich. Sie bietet einen Überblick über KI-gestützte Rekrutierung, einschließlich der damit verbundenen Herausforderungen, und schärft somit das Bewusstsein. Sie adaptiert abstrakte ethische KI-Richtlinien für den Kontext der KI-Rekrutierung und fördert die Entwicklung sowie den Einsatz vertrauenswürdiger KI-Systeme. Sie unterstützt das Verständnis fairer KI-Anwendungen in der Rekrutierung. Die entwickelten Anforderungen und Leitfragen fördern Transparenz und fundierte Entscheidungsfindung bei der Auswahl fairerer KI-Tools. Zudem liefern die Ergebnisse wertvolle Informationen für Anbieter von KI-Rekrutierungstools und regen Verbesserungen im Einklang mit den Anforderungen an. Darüber hinaus betont diese Arbeit die Bedeutung domänenspezifischer KI-Richtlinien und die Notwendigkeit, kritische ethische KI-Prinzipien verbindlich zu machen, und gibt Empfehlungen für zukünftige Verbesserungen.

Abstract

Integrating Artificial Intelligence (AI) in recruitment processes is often seen as advantageous. However, despite claims of objectivity, AI systems have demonstrated vulnerabilities to biases and errors that can lead to discrimination. Research on fairness in AI-assisted recruitment (AI recruitment) is currently limited. Additionally, organisations face challenges in selecting AI recruitment tools concerning fairness.

Using a design science research methodology, this thesis begins by formulating criteria for a fair recruitment process and identifying potential biases in AI recruitment. The focus is on the phases where AI tools are used, directly influencing the evaluation and selection of suitable candidates. The definition of fairness is examined from two perspectives: the perceived fairness by the candidates and the objective fairness concerning the AI system. Through a systematic literature review of AI principles and guidelines, relevant dimensions and requirements for AI recruitment tools that promote fairness are derived. Legal perspectives, such as the EU AI Act, are also considered. An artefact with guiding questions is developed to help organisations identify potential issues in AI recruitment tools. The correctness and completeness of the artefact are validated through comparative analyses with similar research. To evaluate the practicability and usefulness of the artefact, this thesis conducts qualitative case studies on selected AI recruitment applications and identifies more aspects than those documented in existing literature.

This thesis contributes to both academic and practical fields. It provides an overview of AI recruitment, including its associated challenges, thereby raising awareness. It adapts abstract ethical AI guidelines to the context of AI recruitment and promotes the development and adoption of trustworthy AI systems. It supports the understanding of fair AI applications in recruitment. The developed requirements and guiding questions foster transparency and informed decision-making in the selection of fairer AI tools. Additionally, the results offer valuable insights for providers of AI recruitment tools and encourage improvements in alignment with the requirements. Furthermore, this study emphasises the importance of domain-specific AI guidelines and the necessity of making critical ethical AI principles binding and provides recommendations for future enhancements.

Contents

CHAPTER 1

Introduction

1.1 Problem Statement

Artificial Intelligence (AI) is increasingly being adopted in recruitment processes. This is seen as beneficial due to various perceptions such as time savings through task automation and hiring quality improvement through standardised job matching [1]. Moreover, it is often claimed to be objective, although studies have shown that AI is vulnerable to biases and errors [2]. For example, Amazon shut down its AI recruiting tool after discovering that it had learned to prefer male applicants and discriminate against female applicants due to the unbalanced training dataset used to build the model. The model was originally intended to help review applicants' CVs and automate the search for top talent. This incident illustrates how an AI algorithm can introduce bias leading to discrimination and an unfair result [3, 4, 5].

Currently, scientific research on the topic of fairness concerning AI in the recruitment process is limited [6]. In practice, this topic is not only important for data scientists and software engineers but also for the recruiters and decision-makers of the organisation that intends to select and use an AI-assisted tool for recruitment. Due to the knowledge gap, it is difficult for stakeholders to evaluate the various tools offered on the market, especially regarding fairness criteria. AI is integrated into various stages of the recruitment process. In this thesis, the recruitment process is defined as the entire sequence from identifying candidates to making final decisions (see Chapter 2). This thesis focuses more on the phases in which AI tools are used and directly impact the evaluation and selection of suitable candidates.

1.2 Motivation

Fairness definition varies depending on the context. Fairness is of great importance in the recruitment. An example of a prominent theory is organisational theory, which focuses on perceptions of fairness in organisational processes by considering distributive, procedural, and interactional justice. In the context of AI systems, the High-Level Expert Group on AI (HLEG) appointed by the European Commission defined fairness with substantive and procedural dimensions [7]. Fairness is closely associated with the rights to non-discrimination, solidarity, and justice, which are closely linked with explicability and responsibility [7]. Based on the ethical principles, the HLEG identified seven requirements for achieving trustworthy AI [7]. One of these requirements is "diversity, non-discrimination and fairness", which consists of "avoidance of unfair bias", "accessibility and universal design" and "stakeholder participation" [7]. Additional guidelines addressing fairness exist such as the "Recommendation of the Council of Artificial Intelligence (OECD)" and the "Beijing AI Principles for R&D"[8, 9].

Apart from that, the EU has enacted regulations to protect fairness, such as the EU anti-discrimination law. The assurance of fairness is complicated by the black box problem of AI, which addresses the issue of the limited interpretability and the boundaries in explanatory functionality, as only the outcomes are visible to outsiders. This reinforces the importance of explainable artificial intelligence [10].

There are diverse AI guidelines from different affiliations that have conceptual intersections. However, domain-specific guidelines are missing. Furthermore, the guidelines are generic and abstract, making it difficult to understand what the proposed guidelines mean in practice and how to evaluate AI-assisted recruitment tools against the proposed principles. A comprehensive overview of the requirements with a focus on fairness in AI in recruitment is lacking.

Elaborating on the requirements that promote fairness in AI-assisted recruitment is a relevant research topic and can bring multiple benefits to various stakeholders. Academically, this thesis provides an overview of AI-assisted recruitment including the associated challenges, thereby raising awareness. It reviews existing AI guidelines and derives relevant requirements for AI-assisted recruitment tools, supporting responsible AI and fair recruitment. In practice, it serves as a decision-making aid for organisations seeking to select fairer AI-assisted recruitment software. By identifying potential problems, organisations can take action to avoid negative consequences. AI recruitment software must comply with the law. Violations of existing legal principles can lead to lawsuits. In addition, there is a social interest in taking ethical considerations into account. These arguments support the development of an assessment tool based on the requirements. It enables a more transparent selection process among software providers and promotes fairness in recruitment. This thesis also includes case studies to demonstrate the assessment of AI recruitment tools. Implicitly, the findings are relevant for AI recruitment tool vendors to reflect on and improve their products in accordance with the requirements.

1.3 Aim of the Thesis

This thesis aims to promote fairness in AI-assisted recruitment by developing an assessment tool based on fairness-enhancing requirements derived from existing literature. This tool is designed to guide in identifying aspects that should be addressed to generate clarity and mitigate potential issues. Using this tool in the software selection process supports informed decision-making and fosters a more transparent and equitable evaluation of AI systems. Ultimately, this artefact contributes to promoting trustworthy AI and fairness in recruitment practices.

While this thesis does investigate primary international guidelines for responsible AI, the final developed framework focuses on Europe. While legal aspects, including references to the European Artificial Intelligence Act (EU AI Act) and General Data Protection Regulation (GDPR), are considered, specific national laws are not examined. Additionally, each aspect is not explored in full detail, as the focus is not on legal analysis. The artefact serves as a baseline and encourages further discussion. Additional adaptations are needed to customise it for individual cases, such as considering applicable laws, regulations, and the organisation's values.

Research Questions

The main research question (RQ) is: "What are the requirements for AI applications in recruitment to promote a fairer process?" This can be divided into several sub-questions:
RQ1: What are the criteria for a fair recruitment process?
RQ2: How can bias arise in AI applications for recruitment ?
RQ3: Which concepts addressing responsible design and governance of AI exist already?
RQ4: How does the created artefact of fairness-enhancing requirements perform in the evaluation task of identifying critical aspects?

Methodology Summary

The development of the artefact used a design science research approach. After explicating the problem and explaining its background, the criteria for a fair recruitment process were discussed. In addition, the requirements of AI in recruitment for promoting fairness were elaborated based on the results of a systematic literature review. Then, an artefact—specifically a list of guiding questions based on these requirements—was built.

The evaluation was conducted in two ways: First, the created artefact was compared with research of similar objectives to validate its correctness and completeness. In the second step, qualitative case studies of representative AI applications in recruitment were performed systematically to assess the artefact's utility based on observations, using public documents about the applications and the literature. A comparison between the issues found using the artefact and issues mentioned in the literature on the corresponding topic (if they existed) was performed. Improvements to the artefact were made in each step. After finalising the artefact and conducting its discussion, recommendations were drawn. For more details on the methodology, see Chapter 3.

1.4 Outline

The structure of the following chapters is described below.

Chapter 2 presents the theoretical foundations relevant to the research. It begins by introducing human resource management and defining recruitment as used in this thesis, then progresses to explain the basics of Artificial Intelligence (AI) and Machine Learning (ML). Lastly, the chapter investigates the use of AI across multiple stages in the recruitment process, outlining its potential benefits and challenges.

Chapter 3 discusses the research methodology used in the study. It starts by presenting the design science research methodology and justifying its suitability. Following this, the research design is introduced, outlining the research questions and corresponding methods that guide the investigation.

Chapter 4 answers the Research Questions 1,2, and 3. It addresses the concept of fairness within the contexts of recruitment selection and AI. Subsections explore issues such as discrimination and bias. Then, it discusses the responsible design and governance of AI, comparing the most frequently cited guidelines. Dimensions supporting perceived fairness and objective fairness are identified.

Chapter 5 proposes answers to the main Research Question. It describes the requirements for AI recruitment tools to promote both objective and perceived fairness based on the dimensions identified in Chapter 4. To help select fairer tools, key guiding questions for each dimension are developed, capturing the main elements of the requirements.

Chapter 6 evaluates the proposed dimensions, requirements and key questions through alignment analysis and case studies, thereby answering Research Question 4. First, it assesses the correctness and completeness by comparing the artefact with research of similar objectives. Second, it examines the artefact's practicality and effectiveness in identifying issues through case studies in three areas: CV screening, chatbot, and video interview. In each application area, one specific tool is analysed in detail using the artefact. The assessment results of the tools are discussed, followed by a discussion of the artefact's evaluations.

Chapter 7 synthesises the research findings, providing a summary of the answers to the research questions. It explores the implications for theory and practice. The chapter also acknowledges the thesis's limitations and suggests areas for future work. Finally, it presents recommendations for different stakeholders.

CHAPTER 2

Background

This chapter provides the background for a better understanding of the topic. Section 2.1 defines the concept of Human Resource Management (HRM), followed by the definition of recruitment used in this thesis, along with the associated tasks. Section 2.2 elaborates on the basics of Artificial Intelligence (AI) and Machine Learning (ML). Section 2.3 explores the application of AI across various stages of the recruitment process, explaining its potential benefits and challenges.

2.1 Human Resource Management and Recruitment

Human Resource Management (HRM) has been defined in various ways, evolving significantly. Earlier definitions concentrated on management activities affecting the employment relationship [11, 12, 13]. In contrast, contemporary perspectives view HRM as a strategic approach to managing employment relationships, emphasising that maximising employees' abilities and commitment is essential for achieving sustainable competitive advantage or delivering high-quality public services [13]. This strategic approach is implemented through integrated employment policies, programmes, and practices shaped by organisational and societal contexts [13]. Ethical considerations are also important. HRM should support the organisation's objectives and build relationships with employees based on fair treatment, trust, openness, and personal development [14]. The role of HR has evolved from personnel administration to personnel management, and further to a service provider and strategic partner within organisations [15]. Key HRM functions include Planning, recruitment and selection, Staffing and assignment, Development, Incentives and remuneration, Leadership and motivation, and Controlling [15]. For the purpose of this thesis, the recruitment process will be examined in more detail.

Recruitment is a dynamic and complex process that plays a crucial role in organisational success [16]. This topic has caught the attention of both practitioners and researchers over the past century [17]. Some literature distinguishes between recruitment and selection phase [18]: Barber stated that recruitment involves any activity conducted by the organisation aiming at identifying and attracting potential employees primarily. According to her definition, recruitment comprises three phases, i.e. generating applicants, maintaining applicant status, and influencing applicants' job choice [19]. The selection process is the next step after recruitment. It involves evaluating the available candidates based on the job requirements and the candidates' profiles, intending to choose the most suitable future employees for the vacant position [20]. Some literature defines recruitment as encompassing the entire process from generating candidates to post-offer closure [21]. Despite varying definitions, the overall process shares common stages such as establishing objectives, developing strategies, executing activities, assessing candidates, and making final hiring decisions [22, 23, 24].

This thesis uses the hiring funnel model as a reference for the recruitment process to avoid unnecessary complexities. A modified version of the hiring funnel is shown in Figure 2.1. The model divides the process into four main stages: 1. sourcing, 2. screening, 3. interviewing, and 4. selection, with continuous evaluation and decreasing number of candidates [25]. Sourcing aims to attract potential candidates through various channels such as advertisements, job postings, and personal contacts. Screening involves assessing candidates based on their experience, skills, and characteristics against the job requirements. Interviewing allows for a more direct and personalised assessment of candidates through for example face-to-face interactions or virtual sessions. Selection is the final stage where the organisation makes the hiring and compensation decisions [25]. Throughout the process, communication with the applicants and maintaining the process status play an important role [19, 26].

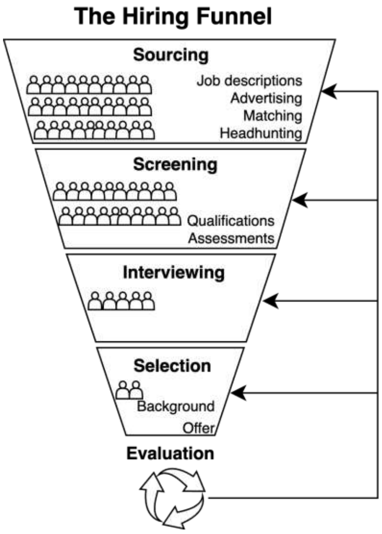

Figure 2.1: The hiring funnel adapted from [25]

2.2 AI and Technology

Artificial intelligence (AI) plays an essential role in digital transformation and is revolutionising the recruitment process. Before diving into this topic, it is necessary to understand the basic concept of AI. Although the idea of artificial intelligence already exists in antiquity, the term artificial intelligence was first introduced by John McCarthy et al. in a proposal for the Dartmouth Conferences in 1956. These researchers proposed that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" [27]. AI involves training computer systems using data and algorithms to perform tasks that typically require human intelligence, such as learning, reasoning, and self-correction [28].

AI is classified into weak (narrow) and strong (general) AI. Weak AI focuses on specific tasks and requires human oversight, encompassing most current AI systems like ChatGPT and autonomous vehicles. Strong AI would match human intelligence with independent learning and problem-solving abilities, a level that has not yet been achieved [29, 30]. Current weak AI technology encompasses several subfields, including machine learning, natural language processing, expert systems, speech recognition, vision, robotics, and planning [31]. These subfields are interconnected, often with overlapping areas, and each can be further divided into subdivisions. A hot topic within current AI research is Generative AI, a subset of machine learning that is capable of creating new content. Most modern generative AI models are based on deep learning architectures. Generative AI uses techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive models (e.g., GPT, which stands for Generative Pre-trained Transformer, based on Large Language Models (LLMs)), and Diffusion models (particularly effective for image generation) to generate various forms of data, including text, images, music, speech, video and code [32, 33].

For the purpose of this thesis, the basic concepts of machine learning (ML) are described to aid in understanding bias. Detailed information on bias is discussed in Chapter 4. ML is a subfield of AI concerned with the development of algorithms and statistical models that enable computers to perform a specific task without explicit instructions, by learning from data [34]. ML includes e.g., supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, deep learning, ensemble learning, and multi-task learning, each with its corresponding algorithms [34]. ML is commonly used for prediction (making accurate forecasts based on input data), classification (assigning input data to predefined categories), clustering and pattern discovery (uncovering structures or hidden relationships in data without predefined labels) [34]. Similarly to the CRISP-DM reference model for data mining, the machine learning life cycle typically follows these stages: 1. problem definition, 2. data collection, 3. data preparation, 4. data analysis, 5. model creation (including feature engineering if necessary, model selection and training), 6. evaluation and tuning, 7. deployment, 8. monitoring and maintenance, and eventually 9. disposal [35, 36].

AI is applied across diverse industries, transforming traditional processes. Table 2.1 illustrates AI adoption in organizations worldwide in 2022 by industry and function [37]. The functions for which organisations are most likely to use AI vary by industry. Overall, AI was used primarily in strategy & corporate finance at 21% in all industries examined [37]. In comparison, AI adoption in the Human Resources function accounted for 11% in total [37]. Among all examined industries, the Healthcare/pharma industry relied most on AI for its Human Resources function with 15%, while the Financial services industry had the lowest adoption rate of 1% [37].

Industry	**HR**	**Manuf.**	**M.&S.**	**P./S.D.**	**Risk**	**S.Ops.**	**S&CF.**	**Supply Chain**
All industries	11	8	5	10	19	19	21	9
Business, legal, & professional services	11	10	9	8	16	20	19	12
Consumer goods/retail	14	4	3	4	15	31	29	11
Financial services	1	8	7	31	17	24	23	2
Healthcare/pharma	15	7	2	4	22	12	8	8
High tech/telecom	6	6	4	7	38	21	25	8

Note: HR= Human Resources, Manuf.= Manufacturing, M.&S.= Marketing & Sales, P./S.D.= Product/Service development, S.Ops.= Service Operations, S&CF.= Strategy & Corporate Finance

Table 2.1: AI Adoption in Organizations Worldwide 2022, by Industry and Function (in %) according to [37]

2.3 AI in Recruitment

AI is used in various ways throughout the recruitment process. Tables 2.2, 2.3, 2.4, and 2.5 provide an overview of the main uses of AI in different stages specific for recruitment, along with examples of associated potential benefits (intended values), challenges, adopters, and vendors [25, 2]. Further tools, such as AI for translation, grammar correction, and text generation, are not listed due to their generic usage. The challenges, especially those related to bias, are addressed in more detail in Subsection 4.2.2. Note that this is a snapshot in time. In the fast-changing environment, potential services may be changed or no longer provided or adopted. AI tools aim to make the recruitment process more efficient and effective. These tools intend to automate certain tasks, support recruiters in decision-making and allow them to focus on more essential tasks rather than replace them.

There are many benefits of using AI tools for organisations, as mentioned by researchers. AI can help in reducing employee attrition and enhancing employer branding. For example, AI tools for vacancy prediction (e.g. Workday talent insights) can analyse employees' behavioural data to predict the probability of their resignation [2]. Based on the prediction, preventive measures can be taken. Such tools also help in reducing costs due to spontaneous resignation. Additional cost savings can be achieved, for example, through optimised job advertisements, or by minimising human errors through tools such as AI-powered background checks [2]. AI can handle processing and analysing massive amounts of data. Time spent on administrative tasks like scheduling tests, interviews or meetings can be reduced with the help of AI tools. Applying AI can save time in tasks such as job-candidate matching, multi-database candidate sourcing and CV screening, as it selects and ranks the best candidates automatically in real-time [2]. Especially AI-powered sourcing in multi-database can improve the sourcing rate, quality and quantity of candidate pool [2].

Another selling point claimed about AI tools is that they can enhance diversity and minimise the risk of indirect discrimination [2]. Research indicates that job descriptions that use stereotypical male language are likely to attract fewer female applicants [38]. To solve this problem, AI can be applied to make suggestions to improve job descriptions and customise the language to attract diverse candidates. Even if the AI suggestions are inaccurate, such tools encourage organisations to invest time in using more inclusive languages. Tools such as CV screening, psychometric testing and video interviewing are argued to promote diversity as they do not generate personal judgment and can avoid certain biases [2].

Certain AI tools aim to promote candidate engagement or enhance candidate experience [2]. AI optimisation for job descriptions and advertisements plays an important role in gaining candidates' awareness. Chatbots apply natural language processing to simulate human conversational skills [2]. They can be used to engage candidates and deliver quick answers to requests at any time. AI-powered psychometric testing and video interviewing can be used to evaluate candidates. Research indicates that using AI-assisted psychometric testing can enhance the candidate-to-hire ratio [2]. The integration of AI-supported gamified testing may relieve candidates's pressure and make the interview process more interesting. AI video interviews offer candidates flexibility in terms of time and location. AI can be used to predict the specific offers (e.g. salary, bonus and other benefits) that candidates are likely to accept which can increase candidates' chance of acceptance [25]. Further uses of AI include employer branding monitoring, which searches public data to evaluate overall sentiment and detect weaknesses in the hiring process [2]. It supports organisations to improve employer branding and talent pool quality, minimize time-to-hire, employee fluctuation and overall costs, as well as maintain a positive image for clients [2].

Although the application of AI in the field of recruitment appears promising, it is still immature. On the practitioner side, an overly optimistic attitude is found in many literature. On the academic side, the sparse literature is predominantly based on fictional credibility [2, 6]. The adoption of AI in recruitment also poses numerous challenges in data, technological, political, legal, policy, ethical, social, economic, organisational and managerial aspects [39]. The data used to build AI tools can pose various issues related to the quantity and quality of input data, transparency, reproducibility, lack of data collecting standards, data integration and continuity [39]. Technological challenges include AI security (with issues such as adversarial attacks which can manipulate the AI model), transparency and interpretability, design of AI systems, architecture issues and AI safety (including AI bias) [40]. Political, legal, and policy challenges encompass governance issues related to responsibility, accountability, privacy and safety, and copyright issues [39]. There is a lack of rules and official industry standards for the use of AI in recruitment and the evaluation of its performance [39]. From an ethical perspective, issues arise in responsibility and explanation of AI decisions, alignment of machine judgment with human value judgment, moral dilemmas, and AI discrimination [39]. For example, AI tools for recruitment are advertised as highly competent and objective decision-making instruments. However, an increasing number of research works reveal their imprecise outcomes and inherent inequities that show discrimination against women and people of colour [41, 42, 43]. Social challenges involve cultural barriers, human rights, unrealistic expectations of AI technology, limited knowledge about the challenges and benefits of AI adoption and a trust deficit in AI [39]. Economic challenges are related to costs and resulting profits [39]. From an organisational and managerial perspective, challenges include a lack of AI experts and AI development strategies, fear of replacement of human workforce, monetary factors, and resistance to cooperation [39].

As mentioned above, the use of AI tools in recruitment brings various benefits and challenges. To decide whether the use of AI tools is worthwhile or which tool should be used, organisations should pay attention to fairness criteria, among others. Not only to comply with the law but also to meet ethical principles, which can also promote the organisation's image.

Stage	**AI Usage**	**Challenges**	**Task Description**	**Intended Values**	**Adoption**	**Vendor**
Sourcing	Vacancy Prediction	Bias in training data, bad historical data	It analyses employees' behavioural data to predict the probability of their resignation.	Enhance employee attrition and employer brand, reduce costs due to spontaneous resignation, decrease time to hire	IBM, Facebook, Goldman Sachs	Workday, BambooHR, Monster Talent Management
Sourcing	Job Description Optimisation	Intransparency, reinforces stereotypical bias	It makes suggestions to improve job descriptions and customise the language to attract diverse candidates.	Enhance diversity, minimise the risk of indirect discrimination, promote candidate engagement	Cisco, American Express, Johnson & Johnson, Nvidia, Expedia, Evernote	Textio, 15Five
Sourcing	Job Advertisements Optimisation	Intransparency, barriers for certain demographics, limits underrepresented groups, skews distribution by gender or race	It optimises the distribution of targeted advertising for relevant candidates based on AI, machine learning, and data insights.	Enhance candidate experience, increase likelihood of candidate engagement, reduce advertising cost	Netflix, YouTube, Starbucks	ClickIQ, PandoLogic, Appcast, Wonderkind, Google
Sourcing	Job-Candidate Matching	Ranking bias, popularity bias, presentation bias, replicates cognitive bias, stereotypes users	It compares job opportunities with potential candidates and generates a ranked list of recommendations.	Let recruiters concentrate on more essential activities	Netflix, eBay	ZipRecruiter, LinkedIn
Sourcing	Multi-database Candidate Sourcing/ Headhunting	Hidden information, equity issues, reproduces cognitive bias, generates unconscious bias, risk of stereotyping, intransparency, predicts actions instead of direct signals, risks neglecting skilled candidates with no prior experience	It searches various databases (e.g., LinkedIn, Glassdoor, Indeed, social media profiles) to find qualified candidates.	Improve candidate sourcing rate, enhance quality and quantity of candidate pool, let recruiters concentrate on essential activities	Intel, eBay, Hilton, Verizon, IBM, Accenture, Warner Bros	Hiretual Pro, Ideal, HiredScore, Recruitment Smart, Eightfold, Engage Talent, Leoforce, Entelo, ZipRecruiter

Table 2.2: AI Applications in the Sourcing Stage [25, 2]

Stage	AI Usage	Challenges	Task Description	Intended Values	Adoption	Vendors
Screening	CV Screening	Reflects prior social biases, NLP absorbs racial and gender bias, disadvantages minority candidates	It selects and ranks the best candidates among numerous CVs in real-time.	Reduce CV reviewing time, minimise bias, promote diversity, lower costs, let recruiters focus on essential activities	IBM, LinkedIn, Hilton, Goldman Sachs, Amazon	IBM Kenexa, Ideal, CVViZ, Zoho Recruit, Talent Recruit, Talent Cube
Screening	Psychometric Testing	Discriminatory evaluation, reflects undesirable social patterns, traits not causally related to performance, biased training data, amplifies differences between candidates	It uses AI to generate appealing tests that enhance candidate experience while evaluating candidates.	Let recruiters focus on essential activities, enhance candidate experience, promote workplace diversity, improve applicant-to-hire ratio	Unilever, PwC, LinkedIn, Tesla, McKinsey, BCG	Arctic Shores, Empirical, Pymetrics, Vervoe, Fortay, Knack, Imbellus, Impress.ai
Interviewing	Video Interviewing	Speech recognition disadvantages accents, facial analysis issues with darker skin, physical features not related to success, infringes dignity and justice, discourages genuine preparation, rewards irrelevant criteria, penalises disabilities, lack of transparency	It analyses candidates' performance (verbal responses, tone, facial expressions) in video interviews to assess fit.	Minimise bias and discrimination, let recruiters focus on other activities, enhance candidate experience	Vodafone, Intel, Urban Outfitters, IBM, Hilton, Unilever	HireVue, MyInterview, Montage, Wepow, InterviewStream, Talview, Knockri

Table 2.3: AI Applications in Screening and Interviewing Stages [25, 2]

Stage	AI Usage	Challenges	Task Description	Intended Values	Adoption	Vendors
Selection	Background Check	Tends to disadvantage people of color, immigrants, and women; social media behaviors may not relate to professional performance; limited ability to identify the real intention due to linguistic ambiguity; hard to define toxic content; gathers personal sensitive data (i.e., pregnancy, sexual identity) that should not be considered during recruitment; restriction by corporate policies and legislation	It examines candidates' background information such as criminal records, credit scores, and references in several databases.	Minimize the costs caused by human errors; let recruiters concentrate on other essential activities	Uber, Axa Insurance, BT, McAfee	Intelligo, GoodHire, HireRight, Sterling Talent, Onfidox, Fama, Predictim
Selection	Offer Acceptance Prediction	May amplify the pay gap by gender and race; information asymmetry between candidates and employers when negotiating wages; violates laws that prohibit employers from evaluating a candidate's wage history	It predicts the specific offers (e.g., salary, bonus, and other benefits) that candidates are likely to accept.	Increase candidate's chance of offer acceptance; review own pay practices	Oracle	Oracle Recruiting Cloud

Table 2.4: AI Applications in the Selection Stage [25, 2]

Stage	AI Usage	Challenges	Task Description	Intended Values	Adoption	Vendors
Sourcing, Screening, Interviewing, Selection	Employer Branding Monitoring	May gather personal sensitive data; relying too much on data from certain platforms can skew sentiment analysis due to over- or underrepresentation of specific demographics; issues with language translation and cultural, emotional, or other aspects that affect the accurate understanding of sentiment	It searches public data to evaluate overall sentiment and detect weaknesses in the hiring process.	Improve employer brand; enhance talent pool quality; maintain positive image for clients; minimize time-to-hire, employee fluctuation, and overall costs	McKinsey, Oracle, HP	Lexalytics, Semantria, Microsoft, Thematic, DiscoverText
Sourcing, Screening, Interviewing, Selection	Candidate Engagement Chatbot	NLP can absorb society's racial and gender bias; can disadvantage minority candidates	It applies natural language processing to simulate human conversational skills. It can be used to engage candidates and deliver quick answers to requests at any time.	Minimize time-to-hire; let recruiters concentrate on other essential activities; enhance candidate experience and employer's image	Sephora, eBay, H&M, Pizza Hut, Burberry	IBM, Impress.ai, Nuance, Kore, Inbenta, Perso-netics, Stepstone, Beamery, AllyO, Xor, TextRecruit, Paradox, Wade and Wendy, SmashFly, Recruitment Smart, Capacity, Koru
Sourcing, Screening, Interviewing, Selection	Automated Scheduling	Bias in training data; might favor candidates in certain time zones while others receive less convenient times	It automatically handles administrative tasks such as scheduling tests, interviews, or meetings.	Let recruiters concentrate on more essential activities	AT&T, Disney, Coca-Cola, Walmart, General Electric, Survey Monkey	X.ai, Troops.ai, Tact.ai, Insight-Squared, My Ally

Table 2.5: AI Applications in Multiple Stages [25, 2]

CHAPTER 3

Methodology

This chapter describes the methodology used in this thesis. Section 3.1 briefly explains the design science research methodology and why it is used. Section 3.2 illustrates how this was applied in this thesis with strategies to address each research question.

3.1 Design Science Research Methodology

Design Science Research is a methodological framework that facilitates the creation of new, innovative artefacts to solve problems or make improvements [44]. It is predominantly used in, but not limited to, the field of information systems [45]. As identified by Hevner [45], the core activities constitute an iterative process of building and evaluating a design artefact. These activities, which aim to solve a real-world problem, are connected to the environment or rather, the application domain through the relevance cycle, and linked to the knowledge base of scientific foundations via the rigour cycle. The outputs can be categorised as a construct, a model, a method, or an instantiation [46]. Johannesson and Perjons expanded the Design Science Research framework defined by Hervner, introducing five key activities: explicating the problem, defining requirements, designing and developing the artefact, demonstrating it, and evaluating it, supported by the knowledge base, research strategies and methods [47]. According to Peffers et al., these activities do not necessarily have to be in sequential order [48]. Researchers can begin at different steps depending on the situation [48].

The Design Science Research approach was chosen because this thesis aims to solve a real-life problem by developing a new artefact. A detailed description of how this approach was used, combined with data strategies, can be found in the following section.

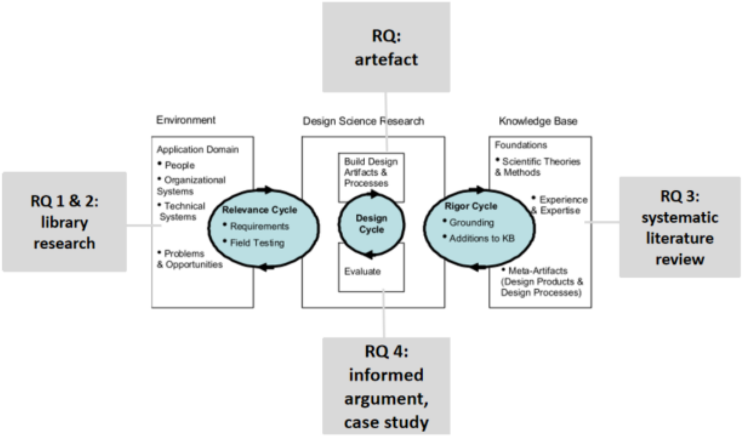

Figure 3.1: Methodology using the adapted design science research approach [45]

3.2 Research Design

The main research question (RQ) is: "What are the requirements for AI applications in recruitment to promote a fairer process?" Note that although referring to recruitment, the focus is on the phases where AI tools are used and directly impact the evaluation and selection of suitable candidates. The main RQ can be divided into several sub-questions:
RQ1: What are the criteria for a fair recruitment process?
RQ2: How can bias arise in AI applications for recruitment ?
RQ3: Which concepts addressing responsible design and governance of AI exist already?
RQ4: How does the created artefact of fairness-enhancing requirements perform in the evaluation task of identifying critical aspects?

To fulfil the research purpose, the design science research methodology was applied as shown in Figure 3.1. To gain sufficient knowledge of the underlying research area and address the research questions, topics such as fairness in recruitment, bias, AI design and governance were initially explored. To answer RQ1 and RQ2, **Library research** was performed, involving a review of the existing literature [49]. Data were gathered using Google Scholar, focusing on reliable research literature.

For RQ3, the topic of responsible design and governance of AI was explored through a **systematic literature review**, as it was essential for deriving comprehensive requirements concerning the fairness of AI-powered recruitment. A systematic literature review aims to determine, evaluate, and interpret the selected literature on a specific research question, area, or phenomenon, which helps to generate a foundation, summarise the state of the art, or identify gaps in current research [50]. Specific search terms, shown below, were determined for the search process. For the selection of the literature, inclusion and exclusion criteria should be defined [50]. Examples of these criteria are listed under Language, Time Range, and Database. The literature found was narrowed down by reading abstracts and conclusions and filtering for relevance. Relevant sources were noted and cited.

Search Terms: “responsible design AI ” or “governance AI” or “AI principles” or “ AI guidelines”.

Language: Only literature in English was considered.

Time Range: 2018-2022

Database: Google Scholar

16,800 results were suggested. To further narrow down the selection, the following selection criteria were added:

Guidelines focus: Primarily international guidelines, excluding those not specific to AI (e.g., those about big data, algorithms, or robotics). Guidelines from specific countries, such as the earliest or from countries leading in AI research, were also noted.

Holdout sets for comparison: Papers conducting similar research that examined multiple guidelines, such as [51] were used only for comparison. The comparison can be found in Section 6.1.

Constraints: Only the direct citations were considered. Linked literature within the papers was not further investigated to avoid duplication and over-complication.

A literature analysis was conducted by critically analysing the literature, building connections, and performing comparisons [49]. Over 63 guidelines were identified across government, science, and industry. The number exceeds 63 because there are additional guidelines, including country-specific guidelines and those from various companies. The citation frequency of the guidelines was counted and ranked. The top six cited guidelines (resulting in seven papers; see Subsection 4.2.3) were examined and compared. The categories identified were summarised. Additionally, guidelines from China and the USA were compared, as they belong to leading AI research countries, and also to examine potential cultural aspects.

After obtaining a broader understanding, requirements promoting the fairness of AI-based recruitment were derived within each relevant identified dimension. Based on this knowledge, an artefact with leading questions for each dimension was created.

For the evaluation of the built artefact for RQ 4, a combination of informed argument (to validate the artefact’s correctness and completeness) and case study (to assess the artefact’s practicability and ability to find issues) was used. **Informed argument** uses information gained from the knowledge base, such as relevant research, to create a convincing argument for the artefact’s utility [52]. Initially, the intention was to compare the built artefact with other similar frameworks. The purpose was to validate the artefact’s correctness and completeness. The library research did not find specific issued frameworks tailored to the issue of AI recruitment tools. Therefore, each part was broken down for separate examination, e.g., comparison with research in AI systems and the recruitment context (details see Section 6.1).

After checking for the alignment of the artefact, specific applications of AI in recruitment were examined in detail through case studies to test its practicability and ability to identify issues. **Case study research** aims to analyse a single unit at a specific time point or within a predefined time range to gain in-depth insights into that unit, which might provide an understanding of a larger category of similar units. In addition to focusing on a single unit in depth, the unit should be examined in its natural setting. [47, 53, 54]. By doing so, it can also illustrate how well the software tools meet the defined requirements.

The selected areas were CV screening, chatbots and video interviews. In each field, one tool was examined, as the goal was to demonstrate the practicability and applicability of the artefact across different stages of AI tools. Google Scholar and website searches were conducted to find AI tools in these three fields. The selection criteria were: The tool is often mentioned and has clients in the EU, as the thesis also has a primary focus on the EU. It should have more public information, including but not limited to white papers, demos, and case studies, to evaluate the tool. It should be mentioned in other research, to compare the issues identified through the artefact with issues mentioned in other research, if they exist. The tool should still be operating and have the relevant features at the time of evaluation. For example, although HireVue is well known for video interviews, it no longer uses its AI component (facial analysis) in its video hiring software and was therefore not selected.

For CV screening, CVVIZ was chosen; for chatbots, impress.ai was selected; and for video interviews, myInterview was chosen. For the assessment of the tools in each case study using the artefact, publicly available documents (including the official websites of each tool and demo videos) were used. In the next step, the built artefact was assessed based on how well it helped to detect critical aspects that should be addressed or identify areas of improvement in each case. A comparison between the issues found using the artefact and those mentioned in existing literature on the corresponding topic (if they exist) was conducted. After consolidating findings from three case studies, the artefact was finalised after iterations and discussed. Last but not least, recommendations were derived based on the systematic analysis.

CHAPTER 4

Fairness in Recruitment Selection and AI

To address the first sub-research question—identifying the criteria for a fair AI-based recruitment process—Section 4.1 examines the factors involved in fair recruitment, specifically focusing on personnel selection, which constitutes the subjective, perceived fairness concept. Section 4.2 then explores fairness in AI, contributing to the concept of objective fairness. Subsection 4.2.1 analyses various types of discrimination, while Subsection 4.2.2 investigates bias in AI and algorithmic fairness. This subsection also answers the second sub-research question by explaining how bias can arise in AI-based recruitment applications and outlining examples of technical measures to mitigate bias. Finally, Subsection 4.2.3 addresses the third sub-research question by focusing on existing concepts for the responsible design and governance of AI, comparing the most frequently cited guidelines, and identifying additional dimensions that support objective fairness.

4.1 Fairness in Recruitment Selection

As mentioned in the first chapter, definitions of fairness vary depending on the context. Different stakeholders might have different criteria. In a fair recruitment process, specifically during selection, the perception of fairness from candidates' perspectives plays a crucial role and is also the focus of this thesis. The perception of unfairness can lead to negative consequences such as reactions during hiring (leaving the talent pool or rejecting the job offer, damaging the organisation's image) or even reactions after hiring (negative influence on performance, organisational climate) [55, 56].

A well-known field of study addressing employees' perception of fairness in the workplace is called organisational justice [57]. Historic research on organisational justice has predominately concentrated on three dimensions - distributive, procedural and interactional justice [58]. Adams asserted in his equity theory that distributive justice relies on individuals' perceptions of fairness in allocating outcomes [59]. Similarities can be found in the theory of relative deprivation which also refers to individuals' evaluation of their contributions and outcomes with those of others to identify distributive justice [59]. If inequity exists in the distribution of rewards, individuals feel deprived [59]. Procedural justice deals with fairness in the process. Thibaut and Walker stated that process control and decision control affect an individual's perception of fairness [60]. Process control refers to having a "voice" in the process (i.e. being able to present information), while decision control implies having a "choice" in the process (i.e. having the ability to influence the decision)[61]. To assess the fairness of allocation procedures, Leventhal identified six justice rules: (1) consistency, (2) bias elimination, (3) accuracy, (4) correctablility, (5) representativeness and (6) ethicality [62]. Bies and Moag proposed the aspect of interactional justice, which focuses on perceptions of fair interpersonal treatment [63]. Some researchers view interactional justice as a subcomponent of procedural justice instead of an independent dimension [64]. Nevertheless, interactional justice comprises interpersonal sensitivity and explanations. Interpersonal sensitivity denotes fair treatment with respect and politeness while explanations provide reasons for the decision [61, 65].

Based on the theory of organisational justice Gilliland derived a model that explains influencing factors of applicants' perceived fairness in the organisational selection system as shown in Figure 4.1 [56]. Gilliland differs between procedural and distributive justice rules [56]. Both rules can influence the overall fairness of the selection process and outcome [56]. Procedural rules are classified into three categories: formal characteristics of the selection system, explanations provided during the process and interpersonal treatment [56]. Conditions such as the type of test, human resource policy and the behaviour of human resource personnel impact applicants' perceptions of the selection system's procedural justice [56]. Formal characteristics comprise job relatedness, opportunity to perform, possibility for reconsideration and consistency of administration [56]. Explanation includes feedback, selection information and honesty[56]. Interpersonal treatment involves interpersonal effectiveness of administrator, two-way communication and propriety of questions [56]. This aspect is sometimes referred to as interactional justice. Additional rules such as invasiveness of questions concerning privacy or ease of falsifying answers can be considered [56].

Conditions such as hiring decision, performance expectations, salience of discrimination and locus of special needs influence distributive justice rules which are based on equity, equality and needs [56]. Past application experience and the stage in the selection process may influence applicants' perceived fairness. The perceived fairness affects applicant's reactions not only during but also after the hiring process [56]. In case of unfairness, applicants may lose motivation, reject the job offer, advise others against applying, and engage in litigation [56]. Even if applicants accept the offer, the perceived fairness in the recruiting process may impact the applicant's job performance, organisational citizenship behaviour, job satisfaction and organisational climate [56]. Moreover, the perceived fairness can play a role in applicants' self-perceptions such as self-esteem, self-efficacy

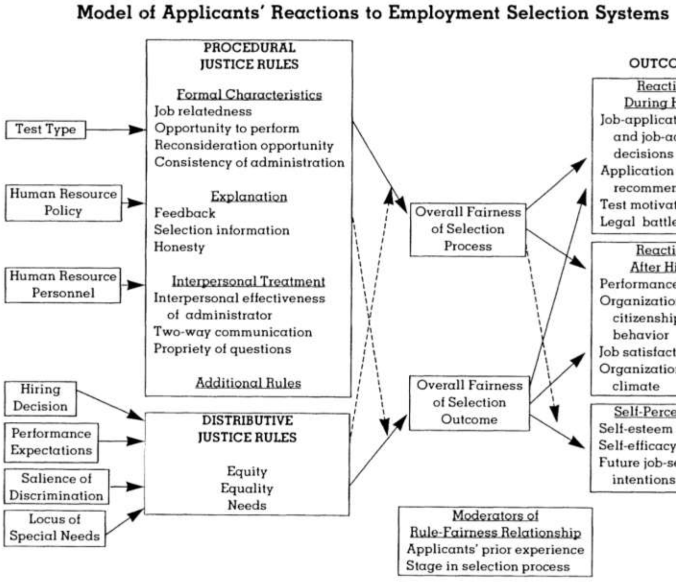
Model of Applicants' Reactions to Employment Selection Systems
PROCEDURAL JUSTICE RULES
Formal Characteristics
Job relatedness
Opportunity to perform
Reconsideration opportunity
Consistency of administration
Explanation
Feedback
Selection information
Honesty
Interpersonal Treatment
Interpersonal effectiveness of administrator
Two-way communication
Propriety of questions
Additional Rules
Test Type
Human Resource Policy
Human Resource Personnel
Hiring Decision
Performance Expectations
Salience of Discrimination
Locus of Special Needs
DISTRIBUTIVE JUSTICE RULES
Equity
Equality
Needs
Overall Fairness of Selection Process
Overall Fairness of Selection Outcome
OUTCOMES
Reactions During Hiring
Job-application and job-acceptance decisions
Application recommendations
Test motivation
Legal battles
Reactions After Hiring
Performance
Organizational citizenship behavior
Job satisfaction
Organizational climate
Self-Perceptions
Self-esteem
Self-efficacy
Future job-search intentions
Moderators of Rule-Fairness Relationship
Applicants' prior experience
Stage in selection process

Figure 4.1: Gilliland's model of applicants' reactions to employment selection systems [56]

and intentions in future job search [56]. This reaffirms the importance of fairness. A fair recruitment selection process should comply with the above-mentioned rules, with a particular emphasis on procedural justice. These form the subjective fairness part of the criteria.

Research on perceived fairness in AI-assisted Recruitment Selection showed controversial results. Some studies revealed that candidates viewed AI interviews as less fair as AI lacks human intuition [66, 67]. AI makes decisions based on keywords and might ignore qualities that are difficult to measure [66, 67]. It cannot make exceptions [66, 67]. Moreover, some candidates perceived algorithmic-based evaluation as demeaning and dehumanising [66, 67]. In opposition to these results, some studies showed that candidates did not perceive differences in fairness between AI interviews and interviews conducted by humans, although most of them showed a lower preference for AI interviews [68, 67]. Some applicants who had previously experienced discrimination argued that the selection process was fairer when decisions were made by algorithms rather than humans [69, 67]. Especially for people with a strong sense of personal uniqueness, AI-based selection has a negative impact on the attractiveness of a company [69, 67]. Different stakeholders may have different or even conflicting conceptions about fairness. Current research also suggested that influencing factors on perceived fairness in AI recruitment include diversity, ethics, bias, discrimination and explainability[70]. The greater the perceived fairness, the greater the acceptance of AI in recruitment [70].

4.2 Fairness in AI

Fairness in AI is not only restricted to technical aspects such as data quality, algorithm design and metrics but also ethical considerations and interconnected dimensions such as transparency and accountability. For example, the High-Level Expert Group on Artificial Intelligence (HLEG) on AI proposed that fairness concerning AI should be defined from a substantive and a procedural perspective [7]. The substantive aspect includes the preservation of equality and justice in the allocation of benefits and costs, and the prohibition of unfair bias, discrimination and stigmatisation of individuals and groups [7]. Furthermore, the employment of AI systems should never deceive individuals or affect their freedom of choice unjustifiably [7]. Moreover, fairness requires AI practitioners to follow the principle of proportionality between means and goals, as well as carefully evaluate how to balance opposing interests and objectives [7]. The procedural aspect involves the ability to challenge decisions and demand redress against judgments made by AI systems and by individuals managing them [7]. This requires the identifiability of the accountable entity and the explainability of the decision-making process[7]. Chapter 4.2.3 examines further interconnected dimensions as well as dimensions proposed by other guidelines.

4.2.1 Discrimination

The topic of discrimination is often explored under fairness. Discrimination theory encompasses multidisciplinary concepts such as legal theory, economics and the social sciences [71]. In terms of algorithmic fairness, the type of discrimination can be classified into explainable discrimination and unexplainable discrimination[71].

- **Explainable discrimination** means that the different treatment and outcomes of different groups can be justified and explained by other attributes and therefore it is not deemed to be illegal discrimination [71]. For example, in the UCI adult dataset, women have a lower average annual income than men because women work fewer hours per week on average than men [71]. If an algorithm proposes to make the average income of women and men equal, this would lead to reverse discrimination, as men would be paid less than women for the same number of working hours [71]. In this case, the discrimination is explained by the attribute working hours and hence, it is acceptable [71].

- In contrast, **unexplainable discrimination** implies that the discrimination is unjustified and thus illegal [71]. Unexplainable discrimination encompasses direct and indirect discrimination [71]. **Direct discrimination** occurs when protected attributes of individuals lead to unfavourable results for them [71]. A group of individuals sharing one or more protected or sensitive attributes is called a protected group. Discrimination against individuals as well as protected groups is prohibited by **Anti-discrimination laws**. Anti-discrimination laws differ per jurisdiction in terms of the sorts of discrimination that are forbidden, and the groups that are protected [72, 73, 74]. For example, the European Commission states that "[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited" [75]. This right is embodied in Article 21 of the Charter of Fundamental Rights [75]. In the United States, Title VII of the Civil Rights Act of 1964 as well as other federal and state acts, outlines anti-discrimination laws which are complemented by court rulings. For example, "Title VII of the Civil Rights Act of 1964 prohibits discrimination in hiring, promotion, discharge, pay, fringe benefits, job training, classification, referral, and other aspects of employment, based on race, colour, religion, sex or national origin" [74, 76]. In US labour law, the concept of direct discrimination is referred to "disparate treatment" which is often associated with intentional discrimination [77].

In **indirect discrimination**, individuals appear to be treated based on apparently neutral, non-protected attributes [71]. Nevertheless, protected groups or individuals are still treated unfairly because of the implicit influence of their protected attributes [71]. For instance, the use of residential postcode, an apparently non-sensitive attribute, in a decision-making process may lead to discrimination such as redlining, because residential areas may correlate with the protected attribute race [71]. In US labour law, this concept is described as "disparate impact" which is often known as unintentional discrimination [77]. Most discrimination by AI tools is indirect and arises unintentionally through machine learning.

European Union's **General Data Protection Regulation** (GDPR) is important for its implications on the usage of machine learning algorithms. The GDPR intends to protect EU citizens' rights regarding data privacy and security. The data protection principles include accountability, lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality [78]. GDPR regulates the way of collecting, storing and processing personal data [78]. Data protection may oblige organisations to apply AI fairness measures. Article 22(3) GDPR mentioned that data controllers shall implement appropriate measures to protect the rights, freedoms and legitimate interests of data subjects in case e solely automated decisions are permitted [78]. Bias reduction should belong to these measures as researchers [79] suggested. Recital 71 GDPR requires measures to correct data inaccuracies, reduce errors and prevent discriminatory impact on individuals based on "racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation" [78]. Algorithmic discrimination also poses challenges for regulatory entities. For example, authorities may lack technical expertise to assess complex AI systems. Existing regulations may not cover all cases. Several data protection mechanisms are retrospective corrective strategies after the damage has already occurred [79, 80].

On 1 August 2024, the **European Artificial Intelligence Act** (EU AI Act) entered into force aiming to foster responsible artificial intelligence development and deployment in the EU [81]. EU AI Act applies to both EU-based and foreign AI systems if they are sold or used within the EU. It creates a unified framework with a risk-based approach:

- **Minimal risk**: No obligations for most AI, like spam filters, though companies can adopt voluntary codes [81].
- **Specific transparency risk**: Systems like chatbots must disclose to users that they are AI, and certain AI-generated content must be labelled [81].
- **High risk**: Strict rules for AI in sensitive areas like healthcare or recruitment, requiring risk mitigation mechanisms, high-quality datasets, transparent user information, and human oversight, etc. [81].
- **Unacceptable risk**: AI applications, such as those enabling "social scoring" or posing a clear threat to fundamental rights, are banned [81].

The EU AI ACT also recommend taking into account the HLEG's ethics guidelines for trustworthy AI [81]. Although both emphasise fairness, a clear definition is lacking. This poses a challenge, as the concept of fairness varies by context and can be interpreted differently from an objective, technical perspective and a subjective viewpoint, shaped by the differing perceptions of stakeholders. It is essential to continue developing regulations and laws and stay up to date with the evolving environment.

4.2.2 Bias

Concerning Fairness, bias is an important topic that has to be addressed, leading to the second sub-research question (RQ2): How can bias arise in AI applications for recruitment?

During the recruitment process, cognitive bias of humans may occur, which may lead to discrimination. A cognitive bias describes a systematic deviation of human judgment from norm or rationality. Commonly known cognitive biases in recruitment includes implicit biases which refers to the tendency to generalise frequently exhibited characteristics in a group and apply them to all individuals within the group; order effect which describes the inclination to assign more weight to Information supplied at first and/or last than information given in the middle; contrast-effects leading to a biased assessment of an applicant based on unconscious benchmark derived from evaluating the performance of another applicant; halo-effect resulting in personality assessment based on a salient characteristic such as perceived attractiveness; confirmation bias which denotes the preference to seek evidence supporting preconceptions while disregarding or lessen the importance of contradictory facts; similar-to-me effect denotes a more affirmative judgment when a candidate shows similarity to the interviewer; stereotypes which indicates beliefs of individuals' characteristics or behaviours that may not correspond to reality; standardised measurements signifying a different evaluation of the same person depending on the performance of the group [15].

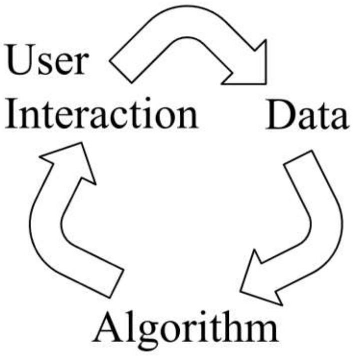

Figure 4.2: Bias feedback loop through user interaction, data and algorithm [71]

It is often expected that using AI in recruitment will reduce bias caused by humans. However, bias can occur at any stage of the AI decision-making pipeline. Human bias may already exist in the problem definition and requirement gathering stage before starting to implement the AI system. Then, the data collected to build the AI model may contain biases [71]. Algorithms trained on them may perpetuate or even amplify the existing biases [71]. Even if the used data is not biased, algorithms can still introduce systematic errors which may unfairly discriminate against specific individuals or groups while favouring others [71, 55]. The biased outcome of the AI system affects users' decision-making [71]. Even if the AI-generated result is unbiased, it cannot be assured that the user's final decision is bias-free. Furthermore, the biased result might be used to train future algorithms [71]. The feedback loop as illustrated in Figure 4.2 will continue to generate biases in case of unawareness and if no action is taken against it [82, 71].

Bias in the **data** that can lead to biased algorithmic results include e.g.: Measurement bias that occurs when features or labels are inaccurately chosen, used and evaluated [71]. A feature or label can be seen as a proxy used to approximate or predict a not directly observable construct. Problems can arise if the proxy is oversimplified or if the method of proxy creation differs across groups or if the accuracy of the proxy varies across groups [83]. Cause-effect bias induced by the mistaken belief that correlation implies causation; Omitted variable bias that arises when one or more essential variables are not included in the model [71]; Representation bias that happens when there is a lack of variety in the population sample data, such as missing subgroups and other anomalies [71]; Aggregation bias that appears when incorrect inferences about individuals are derived from analysing the entire population [71]. Simpson's paradox also belongs to this category [71]. When aggregated data is disaggregated into its underlying subgroup, a previously observed relation vanishes or reverses, resulting in a paradox [71]; Sampling bias resulting from the non-random sampling of subgroups [71]; Longitudinal Data Fallacy that arises from analysing temporal data of cohorts using cross-sectional analysis instead of longitudinal analysis, resulting in different outcomes [71]; Linking bias emerges when network attribute derived from user connections, activities or interactions deviate and do not reflect the actual user behaviour [71].

Bias resulting from **algorithmic** outputs that may affect user's behaviour encompasses e.g.: Algorithmic bias caused by incorrect algorithmic design such as the inaccurate choice of subgroups, estimators, applied models, optimisation functions or regularisations [71]; User interaction bias that is generated through user interface and user biased behaviour reinforced through the algorithm [71]. It can be affected by other types and subtypes of bias such as presentation bias and ranking bias [71]. Presentation bias is generated by the way how information is presented [71]. Users can only interact with content that is shown to them [71]. Ranking bias causes top-ranked outcomes to appear to be the most relevant ones resulting in gaining even more interactions through that [71]; popularity bias which states that more popular content is more likely to get more attention [71]. The evaluation of popularity may be manipulated through fake information, bots or other biased factors [71]. The recommender system or search engine would present the biased outcomes to the public and make them even more popular [71]; emergent bias that arises from real user engagement triggered through new change in population, cultural values or societal knowledge that is not included in the existing system design [71]; evaluation bias that is caused by incorrect and disproportionate benchmarks used for model evaluation [71]. For example, the Adience and IJB-A benchmarks used to evaluate facial recognition systems are skewed towards skin colour and gender [71].

Bias in **user interaction** which may be reflected in the data generated contains e.g.: historical bias, i.e pre-existing biases and socio-technical issues that can affect the data gathering process despite proper sampling and feature selection [71]; population bias that generates non-representative data because the user community differs from the intended target market in terms of statistics, demographics, representatives, and user attributes [71]; self-selection bias which is a subgroup of sampling bias [71]. It is obtained when the research subjects select themselves into a group resulting in a biased sampling [71]; social bias that occurs when individuals' judgment is influenced by others [71]; automation bias that is the tendency of individuals to prefer suggestions from automated decision-making systems and to neglect conflicting information provided without automation, regardless of its correctness [84]; behavioural bias due to the different user behaviour across platforms, contexts or datasets [71]; temporal bias due to the population and behavioural variations across time [71]; content production bias that is induced through structural, lexical, semantic and syntactic disparities in content created by users [71].

Human bias is integrated into the dataset used to train a machine learning model through different ways that can incorporate:

- **Training data**: Machine learning model learns the bias from training on biased data [5]. The biased data can arise from a biased sample of the population (e.g. if the sample is skewed, in which a group attaining one result has proportionally more records than another) or/and occurs from a contaminated dataset's labelled results (e.g. if a human manually tagged the dataset and human bias was passed to the labels) [5]. Both can lead to discrimination. Note also that the validation datasets and test datasets in machine learning models can be biased.

- **Label definitions**: The target label comprises a vague description of the result and therefore leads to erroneous predictions and a larger disparate impact [5]. For example, creating a simple binary classification model to classify a job candidate as a suitable hire— without considering the various aspects that contribute to a candidate's suitability —can lead to many important factors being overlooked by the model's prediction [5]. Employee motivation, person-job fit, and person-environment fit are exemplary aspects that commonly affect how well a candidate fits a company and will perform once employed [5].

- **Feature selection**: The chosen feature to build the model may lead to incorrect predictions [5]. For example, certain properties may be irrelevant to the model's application in the real world causing bias against protected groups [5]. Furthermore, certain traits may be acquired from untrustworthy/inaccurate data leading to decreased prediction accuracy for particular populations [5].

- **Proxies**: Even if protected attributes are deleted from the dataset, they may still be detected in other attributes leading to biased outcomes [5]. For example, Amazon's hiring tool was discriminatory even without using the gender attribute as it derived it from the educational institution stated on candidates' CV (e.g., female-only or male-only colleges) [5].

- **Masking**: New features used to replace protected attributes or their proxies may result in disparate impact when new biases emerge from human-selected features that mask the protected features [5, 77]. In Addition, other types of bias may occur, e.g. technical bias due to technical constraints or technical considerations in the design [55]. For instance, It can arise from "limited computer technology, including hardware, software, and peripherals" or from formalising human constructs that are difficult for computers to quantify [85, 55].

Bias can occur in **AI applications** and cause several **problems**.

- **Sourcing**: For example, vacancy prediction, which analyses employees' behavioural data to predict the probability of their resignation, may suffer from bias in training data and bad historical data. Moreover, the processing of employee data is often constrained by specific legal regulations. In the European Union, such practices must comply with the General Data Protection Regulation (GDPR). Job description optimisation often lacks transparency and can reinforce stereotypical bias. Job advertisement optimisation is performed based on user behaviour such as the number of clicks or job applications [25]. This may create a barrier for demographics that are less inclined to take those actions in the past. It would limit the number of underrepresented groups to whom opportunities are offered [25]. Furthermore, such tools could skew recipient distribution by gender or race even if they aim to be inclusive. Recommender systems for matching candidates and jobs may reinforce cognitive, unconscious and stereotyping biases and lead to discrimination [25].

Matching tools as well as headhunting tools may predict the actions of recruiters or jobseekers instead of direct signals such as “job success” [25]. In addition, AI-supported headhunting tools aim at measuring candidates’ fit to the company and culture posing a risk of neglecting skilled candidates with no prior working experience in similar companies [25].

- **Screening**: If screening systems are designed to mirror an employer’s past hiring decisions, the generated outcomes may reflect past interpersonal, institutional and systemic social biases [25]. AI-assisted assessment can be discriminatory as it may screen out skilled candidates who do not show the defined characteristics. Candidates with different cultural backgrounds may act differently. The tested traits may not have a causal relationship with the working performance. Furthermore, it may disadvantage candidates with disabilities [25]. The differences in assessment scores may amplify the real differences between candidates. The assessment may generate biased statistical accuracy affecting recruiters’ view [25]. It would be problematic if there is no frequent reevaluation and update of the model as well as bad quality of training data such as biased historical employee performance data.

- **Interviewing**: In video interviews, speech recognition may disadvantage candidates with accents [25]. Facial analysis may have issues in recognising candidates with darker skin [25]. The algorithm may reward for irrelevant or unfair criteria such as exaggerated expressions and penalise for disabilities such as visible disabilities or speech disorder [25]. Analysing physical features and facial expressions has no causal relation with workplace success [25]. It may discourage candidates from preparing in good faith to demonstrate that they are qualified for a job [25]. In addition, examining immutable features may infringe on principles of dignity and justice [25]. Automated rejections can be biased. There is a lack of transparency [25]. Automated rejections can negatively influence candidates’ experience if they are rejected in their application without proper explanation or any human interaction. Apart from that, if companies only give automatic rejections, they miss the chance to build relationships with candidates. Bad candidate experience may affect a company’s image and lead to loss of talent and revenue. According to Article 22 of GDPR, data subjects have the right to request human intervention, express their opinion, and challenge automated decisions [78].

- **Selection**: Automated background checks tend to disadvantage people of colour, immigrants and women [25]. Some of these tools collect information from social media [25]. However, social media behaviours may not relate to professional performance [25]. Furthermore, it has limited ability to identify the real intention of the content due to linguistic ambiguity which makes it difficult to determine toxic content [25]. Apart from that, social media background checks are limited by many regulations and corporate policies [25]. The data collected from candidates raises privacy concerns. It may gather personal sensitive data (e.g. pregnancy or sexual identity) that should not be considered during recruitment [25]. Inappropriate processing of personal data may cause legal issues, such as violating GPDR, resulting in monetary losses. Offer prediction tools may generate information asymmetry between candidates and employers when negotiating wages and amplify the existing pay gap by gender and race [25]. Apart from that, offering prediction tools may violate laws that prohibit employers from evaluating a candidate's wage history [25].

- **Stage-independent**: Employer branding monitoring may gather sensitive personal data and rely more heavily on data from certain platforms where specific demographics are over- or underrepresented, leading to a skewed representation of sentiment. There may also be issues with language translation and cultural, emotional, or other aspects that affect the accurate understanding of sentiment. Chatbots relying on natural language processing can absorb society's racial, and gender bias and disadvantage minority candidates due to e.g., expected linguistic patterns [25]. Automated scheduling can also suffer from historically biased data. It might exhibit time zone bias, favouring candidates in certain time zones, while others receive less convenient times.

Algorithmic bias can cause unequal access to resources and opportunities, unfair distribution, loss of trust, incorrect decisions, systematic acceptance of unintended divergence and a vicious cycle of harmful consequences [86]. Numerous **measures** have been proposed to assess algorithmic fairness in past literature. These measures can be categorised into group fairness, individual fairness and subgroup fairness [71].

- **Group fairness** intends to handle different groups equally [71]. Prominent measures of group fairness include, for example, demographic parity and equalised odds [71]. Demographic parity, also referred to as statistical parity, indicates that the probability of a positive prediction is the same across different groups [87, 5]. It complies with laws that require fair hiring procedures, such as the four-fifths rule in the United States [88][1]. Equalised odds, also known as the disparate treatment, demand that protected and unprotected groups have the same rates for true positives and false positives, i.e. the likelihood of an individual in the positive class being correctly assigned a positive result and the likelihood of an individual in a negative class being incorrectly assigned a positive result should both be the same regardless of group [71].
- **Individual fairness** aims to make similar forecasts for similar individuals. It incorporates: fairness through awareness that requires an algorithm predicting comparable outcomes for similar individuals [71, 5]; fairness through unawareness

[1] EEOC states in the Uniform Guidelines for Employee Selection Procedures that "A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact" [88].

that considers the algorithm to be fair if no protected attributes are explicitly used [71]; and counterfactual fairness that states that a decision is fair towards an individual if the outcome is the same in reality as it would be in a counterfactual world, in which the individual falls into a different demographic [89, 71].

- **Subgroup fairness** combines group and individual notions of fairness by choosing a group fairness constraint and testing whether this constraint applies to a large number of subgroups [71]. Examples of further types of fairness measures can be found in [71, 90, 91]. Research [92] has shown that it is not possible to fulfil some of the fairness conditions simultaneously except in extremely limited cases, as some of the fairness conditions are incompatible. Which fairness measure to use depends on the context and the application. Furthermore, one should consider time and temporal analysis of the effects that these definitions may have [71]. Research [93] revealed that the fairness definition did not always support improvement for sensitive groups and may even lead to harm in some cases when studied over time. Moreover, it is recommended to consider the source and type of bias when dealing with fairness-related issues.

Algorithmic bias **mitigation** methods can begin with the correct framing of the problem and continue throughout the entire AI life cycle [94]. Nazer et al. created a checklist to assist in bias mitigation during the development and implementation of AI algorithms [94]. Bias detection can be conducted through auditing or discrimination discovery [95]. The methods for fairness management in ML vary across different application stages [95]. In the following, examples are given:

- **Pre-processing**: This entails modifying the dataset before using it to train the model [5]. The training data should be balanced and representative. Running a statistical analysis is helpful. One may create synthetic data to balance underrepresented groups. Oversampling and undersampling can be considered to balance the dataset. Sensitive attributes such as gender which could introduce bias should be removed. Its proxies need to be examined as well. Note that including sensitive attributes in the data can sometimes help in creating a fair model [96]. Biased or erroneous labels should be corrected [97]. Before classification, algorithms like reweighting and optimized preprocessing can be applied to alter the features and labels in the data to meet fairness criteria [5]. This can be used to eliminate unrelated protected attributes or to change features that could contribute to bias [5]. Careful feature selection and engineering are also crucial.

- **In-processing**: The model can be optimised to remove discrimination during the model training process to fulfil the fairness definition [5]. This is done by changing the objective function or introducing a constraint [5]. Depending on the selected fairness measure, the accuracy of the classifier may be modified[5]. For instance, accuracy parity may lead to unqualified candidates being hired to achieve equal results [5]. Sometimes, changing the model may not be an option, e.g. if the recruitment process is outsourced [5].

- **Post-processing**: After training, post-processing is done by accessing a holdout set that was not used during the model's training [71]. If the algorithm can only use the learned model as a black box and cannot adapt the training data or learning algorithm, the labels generated by the black-box model should be reassigned based on a function [71, 5]. In addition, the result of the model can be modified by establishing a threshold for generated classifications or by offering transparency through counterfactuals [5]. By demonstrating what specific changes could result in a different outcome, counterfactuals can provide transparency, build trust in the model, and enable candidates to understand areas for improvement [5]. This not only helps individuals enhance their future performance but also promotes the perceived fairness in the decision-making process [5].

Orphanou et al.described additional methods in [95]. Both model and outcome explainability are important [95]. In this context, the papers [95, 5] list toolkits such as AI Fairness 360, SHAP, and Lime.

4.2.3 Responsible Design and Governance of AI

Achieving fairness in AI extends beyond technical aspects like data quality, algorithm design, and metrics. It also includes ethical considerations and other closely related dimensions, such as transparency and accountability, as previously mentioned. The following section explores the responsible design and governance of AI to examine further the factors supporting fairness.

Following the approach and criteria described in the methodology, over 63 guidelines were identified, as presented in Table 4.1. The number exceeds 63 because additional country- and company-specific guidelines exist. The top six most frequently cited guidelines were selected, with the sixth position shared by three guidelines. "AI in the UK: Ready, Willing and Able?" was not examined because the focus was not on country-specific guidelines [98]. Guidelines, ranked from most to least frequent, include **Ethics Guidelines for Trustworthy AI** by the High-Level Expert Group on AI (HLEG) appointed by the European Commission[7], **Ethically Aligned Design** by IEEE [99], **OECD AI Principles** [8], **Asilomar AI Principles** developed at the Beneficial AI 2017 conference [100], **Google's Responsible AI Practices** [101], and both the **Montreal Declaration for a Responsible Development of AI** coordinated by the University of Montreal [102] and **AI4People** by Floridi et al. [103]. The content analysis examined the following categories derived from the guidelines (see Table 4.2):

Row Number	Guidelines/Principles	Mentioned by Paper	Frequency	Rank
1	Ethics Guidelines for Trustworthy AI (HLEG)	[104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129], [130]	27	**1**
2	Ethically Aligned Design (IEEE)	[104], [105], [110], [131], [132], [112], [133], [113], [132], [134], [135], [114], [136], [115], [137], [117], [138], [139], [140], [141], [142], [130], [143], [144]	24	**2**
3	OECD principles	[108], [110], [112], [133], [114], [118], [138], [145], [124], [125], [126], [128], [130]	13	**3**
4	Asilomar AI Principles	[108], [133], [113], [114], [136], [146], [117], [118], [121], [39], [128]	11	**4**
5	Google's Responsible AI Practices	[132], [114], [136], [128], [130], [143], [133]	8	**5**
6	AI in the UK: ready, willing and able?	[108], [114], [136], [117], [124], [147], [39]	7	**6**
7	AI4People/five principles key to any ethical framework for AI	[137], [138], [108], [110], [112], [114], [113]	7	**6**
8	Montreal Declaration for a Responsible Development of AI	[106], [108], [133], [114], [136], [117], [148]	7	**6**
9	Microsoft's Office of Responsible AI	[132], [121], [130], [133], [114], [136]	6	9
10	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	[110], [146], [149], [137], [117]	6	9
11	Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence	[106], [108], [133], [150], [147], [131]	6	9
12	Beijing AI principles	[106], [114], [121], [124]	4	12

Table 4.1 – continued from previous page

Row Number	Guidelines/Principles	Mentioned by Paper	Frequency	Rank
13	IBM	[132], [114], [136], [133]	4	12
14	Toronto Declaration Declaration	[133], [148], [130], [106]	4	12
15	ACM code of ethics	[110], [146], [130]	3	15
16	OpenAI	[106], [136], [121]	3	15
17	Axon's AI Ethics Board for Public Safety	[106]	2	17
18	Canada	[129], [133]	2	17
19	G20 Ministerial statement on trade and digital economy	[118], [125]	2	17
20	Intel	[114], [133]	2	17
21	More Countries	[133], [131]	2	17
22	Preparing for the future of artificial intelligence (Whitehouse)	[142], [125]	2	17
23	SAP's Guiding Principles for Artificial Intelligence	[136], [130]	2	17
24	Tenets of Partnership on AI	[136], [117]	2	17
25	The Partnership on AI to Benefit People and Society	[106], [137]	2	17
26	UNESCO : Report of World Commission on the Ethics of Scientific Knowledge and Technology	[113], [118]	2	17
27	White House principles (Vought)	[108], [125]	2	17
28	Universal Guidelines for Artificial Intelligence	[136],[130]	2	17
29	Accenture report outlining a framework to assist US federal agencies to evaluate, deploy and monitor AI systems.	[117]	1	29

Table 4.1 – continued from previous page

Row Number	Guidelines/Principles	Mentioned by Paper	Frequency	Rank
30	AI Next programme	[117]	1	29
31	AI Policy Principles	[136]	1	29
32	AI R&D Principles	[136]	1	29
33	Australia's Ethics Framework	[113]	1	29
34	Baidu	[133]	1	29
35	The Centre for Humane Technology	[106]	1	29
36	Denmark	[133]	1	29
37	DeepMind Ethics & Society Principles	[136]	1	29
38	Developing AI for Business with Five Core Principles	[136]	1	29
39	Draft AI Utilization Principles	[136]	1	29
40	Ethical principles and democratic prerequisites, European Group on Ethics in Science and New Technologies	[136]	1	29
41	Fairness, Accountability and Transparency in Machine Learning	[106]	1	29
42	France	[133]	1	29
43	Harmonious Artificial Intelligence Principles	[136]	1	29
44	International Association of Privacy Professionals	[113]	1	29
45	Malta	[133]	1	29
46	More companies	[133]	1	29
47 48	Sage Principles for Algorithmic Transparency and Accountability by ACM	[133] [136]	1 1	29 29
49	Principles for the Governance of AI	[136]	1	29

Table 4.1 – continued from previous page

Row Number	Guidelines/Principles	Mentioned by Paper	Frequency	Rank
50	Sony Group AI Ethics Guidelines	[136]	1	29
51	Stanford Human-Centered AI Initiative	[136]	1	29
52	Summary report 2018 global governance of AI roundtable	[118]	1	29
53	Tecent	[133]	1	29
54	The Council of Europe published a draft recommendation on the human rights impacts of algorithmic systems.	[128]	1	29
55	The EU Declaration of Cooperation on Artificial Intelligence	[137]	1	29
56	The European Union (EU) strategy for AI	[137]	1	29
57	The Japanese Society for Artificial Intelligence Ethical Guidelines	[136]	1	29
58	The UK's House of Lords Artificial Intelligence Committee	[142]	1	29
59	The UNI Global Union	[128]	1	29
60	Three Rules for Artificial Intelligence Systems by the CEO of Allen Institute for Artificial Intelligence	[136]	1	29
61	Top 10 Principles For Ethical Artificial Intelligence	[136]	1	29
62	White Paper on Artificial Intelligence Standardization	[106]	1	29
63	Workday	[133]	1	29

Table 4.1: Guidelines/principles ranked by frequency of mentions

Category	**R1.EU HLEG**[7]	**R2.IEEE**[99]	**R3.OECD**[8]	**R4.Asilomar**[100]	**R5.Google**[101]	**R6.Montreal**[102]	**R6.AI4People**[103]	**Beijing AI**[9]	**China New Gen**[151]	**U.S. 2016**[152]	**U.S. 2020**[153]
Release Version	2019	2019	2019	2017	2018	2018	2018	2019	2019	2016	2020
Human rights, agency, oversight	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩
Technical robustness, safety, security	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Transparency	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Privacy, data governance	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Diversity, non-discrimination, fairness	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Societal, environmental wellbeing	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Accountability, responsibility	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Awareness, education, discussion	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
AI arms races, weapons	🟩			🟩	🟩		🟩			🟩	
Regulations, governance frameworks	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩
Standards, certifications	🟩	🟩	🟩	🟩	🟩		🟩		🟩	🟩	🟩
Science-policy link	🟩	🟩		🟩			🟩	🟩	🟩	🟩	🟩
Research funding	🟩		🟩	🟩			🟩			🟩	
Research culture (diversity, collaboration)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Specific industry mentioned	🟩					🟩				🟩	
Affiliation	gov	sci	gov	sci	ind	sci	sci	sci,ind	gov	gov	gov

Note: Shaded cells indicate that the dimension is addressed in the respective guideline. Column headers are abbreviated. R stands for Rank as in Table 4.1

Table 4.2: Comparison of selected guidelines/principles, inspired by [154]

These categories are not exhaustive. Each can be further subdivided into more detailed and specific subcategories. The aspects explicitly mentioned are highlighted in green.

The category of **Human rights, agency, and oversight** encompasses topics related to human fundamental rights, human agency and human oversight [7, 99, 8, 100, 101, 102, 103]. It mandates that AI systems respect, promote, and protect human rights while upholding the rule of law [7, 99, 8]. Additionally, AI systems should incorporate human oversight and enable users to make informed autonomous decisions [7, 8]. All seven guidelines address this category. Specific guidelines such as [7, 102, 8, 103] emphasise the principles of human autonomy.

Regarding the category of **Technical robustness, safety, and security**, all seven guidelines emphasise the necessity of developing reliable and resilient AI systems, and advocate for the prevention of harm [7, 102, 8, 103, 99, 100, 101]. Protecting AI systems from cyber threats and ensuring data integrity are shared concerns across these guidelines [7, 102, 8, 103, 99, 100, 101]. Notably, the HLEG explicitly addresses the need to prevent both physical and psychological harm, which may be overlooked in discussions about AI safety [7].

In explainable AI, terms like **Transparency**, interpretability, explainability, and understandability are often used interchangeably or defined inconsistently [155]. Verhagen et al. suggest that transparency is the system's openness in disclosing its external and functional elements to users [155]. Interpretability refers to how easily users can understand and analyse the system based on the disclosed information [155]. Explainability clarifies the system's elements by explaining their relationships and causes to help users understand its behaviour [155]. Understandability refers to how well users comprehend the system's operation through the combination of transparency and explainability [155].

All seven guidelines highlight the importance of transparency in AI systems [7, 102, 8, 103, 99, 100, 101]. The HLEG integrates traceability, explainability, and communication within the concept of transparency [7]. The Montreal Declaration for the Responsible Development of AI does not mention explainability [102]. Even when some guidelines mention explainability, they often lack specificity. For example, the Asilomar AI principles state the need for a "satisfactory explanation", which leaves room for interpretation [100]. The AI4people framework introduces "Explicability", which combines intelligibility (providing clear explanations of how AI systems operate) and accountability (establishing who is responsible for the actions and decisions of AI systems) [103]. This also illustrates that the categories are closely tied together. Understandability and interpretability have subjective characteristics, as they depend on the user's familiarity with the system, the current state of mind, and background knowledge[155]. Understandability and interpretability are rarely addressed in guidelines. For example, they are not mentioned in [100, 101, 99]. Apart from ensuring the general accessibility of information, explanations need to be tailored to different audiences and contexts.

All seven guidelines mention **Privacy and data governance** aspects such as respect for privacy, data protection, access, and control over data [7, 102, 8, 103, 99, 100, 101].

The category **Diversity, non-discrimination and fairness** is also mentioned in all the guidelines. Some guidelines such as [7, 100] explicitly mention cultural diversity. Fairness is often mentioned with no concrete definition, although the definition and metrics could vary depending on the context. Within these guidelines, concepts of fairness or justice vary, referring to aspects such as fair resource distribution, eliminating discrimination, equality (no biased output), ensuring shared benefits, and the ability to contest AI decisions [103].

A further shared concern across the seven guidelines is that AI should be developed and used in ways that are beneficial to **Society and the environment** [7, 102, 8, 103, 99, 100, 101]. Related concepts include sustainability and environmental friendliness, social impact, society and democracy, although the aspect of democracy is rarely mentioned [7, 102, 8, 103, 99, 100, 101].

All guidelines mention **Accountability and responsibility** perspectives [7, 102, 8, 103, 99, 100, 101]. Designers and operators should ensure AI systems are ethical, function correctly, and are protected against misuse [7, 102, 8, 103, 99, 100]. Organisations should address the moral implications of AI and maintain proper operations [7, 102, 100, 8, 103, 99]. Again, human oversight is essential to control AI decision-making [7, 102, 8, 103, 99, 100, 101]. Concepts falling into these categories include auditability, minimisation and reporting of negative impacts, risk management, trade-offs and rationale documentation, and the possibility of redress [7, 102, 8, 103, 99, 100, 101]. The guidelines from HLEG provide the most comprehensive aspects[7].

All the guidelines mention aspects of **Awareness, education, and discussion**. This involves ensuring that stakeholders recognise when they are interacting with AI systems and understand their capabilities and limitations. This category also includes educating the public about AI technologies, their risks, and responsible usage, and engaging stakeholders in conversations about the ethical implications of AI use and misuse to promote responsible actions [7, 102, 8, 103, 99, 100, 101].

Clear differences could be found within the category of **AI arms races and weapons**, which is explicitly opposed in [7, 103, 100, 101]. While this issue might be interpreted under the societal and environmental wellbeing category, it is only explicitly mentioned in these sources.

The category **Regulations and governance frameworks** for AI encompasses key elements such as the responsibility of policymakers at local, national, and international levels to oversee the development and deployment of AI technologies. AI laws and regulations should be continuously developed and updated to address emerging technologies and challenges. Similarly, organisations should ensure that their AI governance frameworks are aligned with both evolving regulations and their core values [7, 102, 8, 103, 99, 100, 101]. While some aspects are not explicitly stated, they could be inferred. For example, the Asilomar principles emphasise the importance of “constructive and healthy exchange between AI researchers and policymakers,” which falls into the category of Science-policy link, and does not directly mandate specific regulatory actions [100]. The Asilomar principles raise critical questions about AI's beneficial use and address issues related to law, ethics, and governance, they do not explicitly call for direct regulation by policymakers. Instead, they emphasise the need for research funding to explore these complex issues [100]. On the other hand, Google's AI principles focus on internal organisational governance and alignment and do not explicitly call for higher-level policy intervention, which is understandable given their focus on ensuring the organisation complies with existing laws and regulations.

The category **Standards and certifications** deals with aspects related to establishing standards and certifications for ethical AI. Most of the guidelines explicitly emphasise standards, whereas certification is less frequently mentioned [7, 102, 8, 103, 99, 100, 101]. The Montreal Declaration does not mention these aspects.

The category **Science-policy link** addresses aspects related to promoting exchange and collaboration between scientists and policymakers, as mentioned in [7, 103, 99, 100]. The category **Research funding** concerns aspects related to the need for governments to invest in AI research and development related to ethical AI, and to provide financial incentives. This is highlighted in [7, 103, 8, 100].

The **Research culture** category relates to diversity among researchers and the necessity for cooperation. The HLEG calls for all stakeholders to work towards a global framework [7]. The IEEE stresses cooperation between academia and industry, while the OECD underscores international cooperation [7, 8]. The Asilomar Principles advocate for fostering cooperation, trust, and transparency among AI researchers and developers [**?**]. Google commits to working with a range of stakeholders, and the Montreal Declaration states that AI research should be open, accessible, and inclusive, reflecting societal diversity [101, 102]. AI4People promotes cross-disciplinary and cross-sectoral cooperation [103]. Advancing AI responsibly requires collaborative efforts across different sectors, disciplines, and international borders, fostering diversity and involving stakeholders.

A **Specific industry** is rarely addressed, and when it is, the discussion is brief. For example, the Montreal declaration states that healthcare using AI systems must consider the importance of a patient's relationships with family and healthcare staff [102]. The HLEG mentions lethal autonomous weapon systems in a military context [7]. Each industry has its additional points that should be considered, which also shows the importance of this thesis in customising principles into a specific domain, i.e., AI in HR recruitment and selection.

The **Affiliation** category shows whether the guidelines originate from government, industry, science, NGO, etc. Note that even if they are published under a government affiliation, researchers and practitioners might also have participated in creating these. Among the seven most mentioned principles, two are from the government, four from science and one from the industry. Government is more mentioned, which could be because government affiliations are often taken more seriously due to potential consequences. It is important to have binding regulations and laws to promote ethical AI.

Considering the potential **cultural variations**, well-known guidelines from China and the United States are analysed, as they belong to leading countries in AI research and technology. No additional relevant category was found for the main purpose of this thesis. From China, the “Beijing Artificial Intelligence Principles” a collaboration between research institutions and industries, and the government’s “New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence” are examined [9, 151]. In the United States, two significant documents are reviewed: “Preparing for the Future of Artificial Intelligence” from the Obama administration and “Guidance for Regulation of Artificial Intelligence Applications” issued during Trump’s presidency [152, 153].

An observation is that China’s “Beijing Artificial Intelligence Principles” advocate respecting human privacy without explicitly mentioning data privacy and protection [9]. While these principles do mention data security [9], the explicit mention of data protection emerges in a subsequent government publication titled “China’s New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence” [151]. This later document also introduces a perspective on standards, an aspect absent from the earlier principles [9, 151]. Guidelines should be continually adjusted to reflect the changing environment. The United States’s guidelines also mention privacy, focusing more on the security perspective [152, 153]. All four guidelines stress the importance of open data and sharing for collective benefit [9, 151, 152, 153]. In comparison, the guidelines from HLEG place a stronger focus on data protection, likely influenced by the General Data Protection Regulation (GDPR) [7].

Notably, the United States’s “Preparing for the Future of Artificial Intelligence” lacks explicit guidelines on Human rights, agency, and oversight[152]. The “Guidance for Regulation of Artificial Intelligence Applications” does not address diversity [153]. While China’s “New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence” does not specifically address the AI arms race or weapons, it does highlight the aim of societal benefit [151]. The “Beijing Artificial Intelligence Principles” state the importance of avoiding a “malicious AI race” [9]. Although the AI arms race or weapons are not explicitly mentioned, they could be interpreted as a subset of a malicious AI race. Both Chinese documents reflect a cultural emphasis on “harmony” [151, 9]. The United States’s “Guidance for Regulation of Artificial Intelligence Applications” does not mention the topic of AI arms races or weapons [153]. It emphasises avoiding creating rules that could hinder the United States’s innovation and competitiveness, considering the potential benefits and costs of AI [153].

Among the four guidelines [9, 151, 152, 153], [152] explicitly mentions research funding and specific industry aspects. The “Beijing Artificial Intelligence Principles” highlight the importance of considering various fields and scenarios in the application of AI to guide the development of more specific and detailed guidelines [9]. This reinforces the relevance of this thesis.

The focus areas among the guidelines vary. For example, while the Asilomar AI Principles focus on long-term alignment with human values and global governance, Google’s AI Principles are more oriented towards practical applications and internal company policies [100, 101]. This difference is also understandable when considering the affiliation of the organisations issuing the guidelines. The OECD principles have specific sections calling for policy-makers and governments. The US guideline emphasises American values, including innovation and competitiveness [153]. Although some aspects are categorised differently—such as traceability, which the OECD classifies under accountability and the HLEG under transparency—the similarity can be found in the overall message.

The following categories are particularly relevant for deriving fairness-enhancing requirements for AI Recruitment tools: 1.Human rights, agency, and oversight; 2.Technical robustness, safety, and security; 3.Privacy and data governance; 4.Transparency, 5.Diversity, non-discrimination, and fairness; 6.Societal and environmental wellbeing; 7.Accountability and responsibility [7, 102, 8, 103, 99, 100, 101]. These categories are closely interconnected and support each other.

The primary issue with current guidelines is their abstract and generic nature, which leaves significant room for interpretation. It is recommended to establish a global unified baseline framework for AI systems that can then be adapted following applicable laws and regulations. Creating a flexible AI governance system that addresses the diverse nature of AI while considering cultural differences and various national legal systems is challenging. More concrete recommendations, such as standards and best practices on implementation, would be helpful. Industry-specific guidelines need to be further developed, and ideally break down further. Organisation-specific values and policies can be considered within the organisation. All these need to be updated to keep pace with current developments.

To derive fairness-enhancing requirements for AI recruitment tools, specifically for objective fairness, the HLEG's guidelines were chosen because they provide comprehensive and relevant aspects, and the focus of this thesis is on the EU. The HLEG's dimensions are called: 1.Human agency and oversight; 2.Technical robustness and safety; 3.Privacy and data governance; 4.Transparency; 5.Diversity, non-discrimination and fairness; 6.Societal and environmental wellbeing; 7.Accountability [7]. For subjective fairness, the perceived fairness framework from Gilliland was selected [56].

CHAPTER 5

AI Recruitment Tool Requirements for Promoting Fairness

Chapter 5 applies the dimensions of objective and subjective fairness to AI recruitment tools, and derives key requirements to promote fairness in the recruitment process. Guiding questions for each dimension are developed to help select fairer tools and ultimately support fair recruitment practices.

5.1 Requirements for Promoting Objective Fairness

Starting with the perspective of trustworthy AI from the High-Level Expert Group (HLEG), which supports objective fairness:

Diversity, non-discrimination and fairness:

The dimension of Diversity, non-discrimination and fairness is divided into three subdimensions: **Avoidance of bias**, **Accessibility and universal design**, and **Stakeholder participation** [7]. As mentioned in Chapter 4, fairness can have various definitions. It's important to determine whether the AI system uses a purpose-serving definition of fairness and ensures its correct implementation. The AI recruitment tool should be designed and continually evaluated to avoid unfair bias. This includes implementing strategies to prevent and mitigate bias in both the data and the algorithm's design, ensuring it does not perpetuate historical biases or discrimination based on race, gender, age, or other protected characteristics. Additionally, cultural perspectives should be considered throughout development. For example, training data should not be sourced solely from one cultural context, as this may lead to misinterpretation of candidates from different backgrounds. One might consider additional cultural aspects, such as different cultural norms, values, and communication styles. Beyond the tool's design, it is important to assess whether the AI recruitment tool leads to fair hiring outcomes in practice. This can include evaluating the impact on diversity within the organisation and whether the tool contributes positively to equitable employment practices.

Second, the AI recruitment tool should be user-centric, accessible and usable by all users, regardless of their age, gender, abilities, or characteristics [7]. This includes adhering to universal design principles and following accessibility standards to ensure that people with disabilities can equally participate in the recruitment process [7]. For example, the tool can offer features such as screen reader compatibility, adjustable text sizes, and clear, understandable language to cater to a wide range of users. Third, there should be active and ongoing stakeholder participation in the development and deployment of the AI recruitment tool [7]. This involves consulting with a diverse group of stakeholders, such as users (job applicants, HR professionals), legal and ethical experts throughout the AI system's life cycle. Engaging individuals from diverse backgrounds in the design and implementation of AI recruitment tools can help identify potential biases and cultural misunderstandings. Regular feedback should also be gathered after deployment. There should be mechanisms for stakeholders to report concerns and suggest improvements. Considering the perspectives mentioned, the main questions in each subdimension of Diversity, non-discrimination and fairness are:

1. Avoidance of bias
Does the AI tool have an appropriate, justified fairness definition and effective strategies implemented to ensure fairness, including mitigating inappropriate bias in data, algorithm design and outputs?

2. Accessibility and universal design
Is the AI recruitment tool user-centric, adhering to universal design, and accessible to all users, including those with disabilities?

3. Stakeholder participation
Does the design, development, and continuous improvement of the AI recruitment tool incorporate input from a diverse range of stakeholders, such as job applicants, HR professionals, legal experts, and ethicists?

Human agency and oversight:

The dimension of Human agency and oversight includes **Fundamental rights**, **Human agency** and **Human oversight** [7]. AI systems have the potential to both enable and restrict fundamental rights [7]. In the context of recruitment, AI tools should be assessed for their negative impact on these rights before development [7]. In the European context, this necessitates a comprehensive evaluation of the system's potential to affect rights such as human dignity, individual freedom, democracy, justice, the rule of law, equality, non-discrimination, solidarity and citizens' rights [7]. AI recruitment tools should guarantee equal respect for all individuals, going beyond mere non-discrimination.

This involves ensuring that the system does not produce unjustified biased outcomes and that its operations are inclusive, representing diverse population groups [7]. Special attention might be given to potential groups to prevent exclusion or unfair treatment.

AI recruitment tools should adhere to legal and regulatory frameworks. For example, in Europe considering the General Data Protection Regulation(GDPR), an evaluation should be conducted on how the tool processes personal data, the necessity of this processing, and its proportionality in relation to the tool's purpose [78]. The HLEG recommends conducting a fundamental rights impact assessment [7, 36]. This assessment should evaluate whether risks can be mitigated or justified as necessary in a democratic society, ensuring the preservation of balanced rights and freedoms before the system's development [7]. In terms of the risk management system, the EU AI ACT Article 9 specifies requirements that high-risk AI systems shall follow, including risk analysis for intended use, risk estimation and assessment, evaluation of additional risks from post-market monitoring, and implementation of risk management measures [156]. Furthermore, the AI recruitment tool should have mechanisms to regularly review and address any potential infringements on fundamental rights [7]. Concerning the risk management system, various standards exist, such as ISO/IEC 23894 AI - Guidance on risk management.

4. Fundamental rights
Does the AI recruitment tool undergo evaluation to identify and mitigate potential risks[a] to fundamental rights[b] during its design, development, and operational phases?

[a] Certain risks can be justified as necessary in a democratic society to ensure the preservation of balanced rights and freedom.

[b] In the EU, it includes human dignity, individual freedom, democracy, justice, the rule of law, equality, non-discrimination, solidarity, and citizens' rights.

AI systems should support, not replace, human decision-making in recruitment. Both HR personnel and candidates, as users, should be aware they are interacting with AI and have access to necessary information for effective engagement [7]. This includes instructions about the system's setup, objectives, and submission process. The AI recruitment tool can offer tutorials or tests to help users become familiar with its features. In the EU, GDPR compliance is crucial, necessitating transparent explanations about data collection and its purpose [78]. For more information, please refer to the dimension of Privacy and data governance.

Procedures should be established to prevent unintended AI influences, including those involving subconscious processes such as unfair manipulation and conditioning, which may harm human autonomy [7]. Unfair manipulation can occur if the AI influences the decisions of recruiters by prioritising certain profiles over others based on biased data or algorithms. This manipulation might not be noticeable but can lead to unfair hiring practices. Over time, the use of AI in recruitment might condition recruiters or candidates to respond in specific ways. For example, recruiters might ignore certain aspects of a candidate's profile because the AI does not emphasise these aspects, or candidates might tailor their applications to what they believe the AI system favours, rather than presenting their genuine skills and experiences. HR personnel, as human decision-makers, should be informed of the tool's limitations and receive adequate training and support to use the AI effectively. Users should be able to challenge AI decisions and provide feedback, enabling the system to learn and improve. The functionality of the AI tool must centre on human autonomy [7].

5. Human agency
Does the AI recruitment tool focus on human autonomy by empowering users to make informed autonomous decisions, including implementing safeguards against manipulation, over-reliance, and confusion between AI and human interactions?

The AI recruitment tool should be designed to incorporate effective human oversight to safeguard human autonomy and minimise adverse effects [7]. The tool provider should conduct periodic audits and evaluations, focusing on the tool's performance, accuracy and fairness. Involving human resources professionals in these evaluations might help align the tool with the organisation's objectives and principles. One should not rely solely on algorithms. Recruitment personnel should be able to review and, if necessary, override decisions made by the AI tool throughout the recruitment process to prevent inappropriate outcomes such as biases and errors [7]. To facilitate this, transparency and clear explanations are essential, as detailed in the dimension of Transparency.

Incorporating AI into the recruitment process requires careful consideration of its extensive influence on candidates, the organisation, and adherence to legal and ethical standards. It is essential to assign a skilled individual or team with the authority to restrict or discontinue the use of the AI recruitment tool based on its performance, ethical considerations, and legal compliance [7]. Article 14 of the EU AI ACT describes similar concepts of human oversight of high-risk AI systems [156]. For specific high-risk systems, additional measures require decisions based on AI outputs to be verified by at least two humans with the necessary competence, training and authority [156].

6. Human oversight
Does the AI recruitment tool incorporate mechanisms for effective human oversight, including options for modification to prevent inappropriate outcomes?

Technical robustness and safety:

When selecting an AI tool for recruitment, it's crucial to focus on technical robustness [7]. This involves ensuring the AI system is designed to proactively mitigate risks and operate reliably as intended, minimising unintended and unexpected negative impacts [7]. Additionally, it's essential to safeguard these systems against vulnerabilities to prevent exploitation by malicious entities. The tool provider should assess the vulnerability of the AI recruitment tool to cyber-attacks, including data poisoning, model evasion and model invasion [36].

Data poisoning involves manipulating training data, to contaminate the learning process, thereby degrading model performance [36, 157]. For example, in AI recruitment tools, inserting false or manipulated candidate profiles labelled as top candidates can corrupt training data. This could lead to biased hiring decisions favouring certain attributes that reflect the attacker's goals rather than the true qualifications needed for the job.

Model evasion involves manipulating input data in the test phase to exploit weaknesses in the model, causing the model to make incorrect predictions or classifications [157, 36]. In AI recruitment, model evasion can occur if applications are artificially enhanced by overusing specific keywords that the AI tool favours, thereby gaining an advantage despite not meeting the job requirements. For example, "collaborated with ML teams" does not necessarily mean that the applicant has actual expertise in ML. The model may misinterpret the resume and place the applicant higher in the ranking due to the keyword "ML".

Model inversion attacks attempt to infer sensitive information about the training data by analysing the outputs produced by the model [158]. In an AI recruitment context, this can involve using an AI tool trained on actual candidates' profiles to uncover sensitive information.

When dealing with generative AI, potential issues include data security and privacy (e.g., data leakage), data integrity and quality (e.g., data poisoning), model security and integrity (e.g., adversarial attacks, backdoor attacks, jailbreaking), input manipulation (e.g., prompt injection), and output reliability (e.g., faithfulness, RAG (Retrieval-Augmented Generation) hallucination, toxicity) [159, 160].

The AI tool should possess **robust security** measures to **protect against attacks** throughout its life cycle, including regular security updates and penetration testing [36]. The AI tool should be certified for security (e.g., through the certification scheme created by the Cybersecurity Act in Europe [161]) or comply with recognised security standards [36]. According to Article 15 of the EU AI Act, high-risk AI systems shall achieve appropriate levels of accuracy, robustness, and cybersecurity throughout their life cycle[156]. To effectively assess the accuracy, robustness, and other performance metrics of high-risk AI systems, the commission encourages the development of benchmarks and measurement methods in collaboration with relevant stakeholders [156]. Various standards organisations, such as ISO, IEEE, and ETSI, have published cybersecurity standards that could be considered. Additionally, the AI recruitment tool provider should implement safeguards against social engineering attacks and measures to protect physical security.

7. Resilience to attack and security
Is the AI recruitment tool certified for security or compliant with security standards, including having measures to counter cyber-attacks [a] through ongoing security

updates and penetration testing?

[a] such as data poisoning, model evasion and model inversion

The AI recruitment tool should have mechanisms to switch from AI to rule-based processing or human oversight in cases of anomalies or failures, to prevent harm [7]. Ongoing risk assessments are necessary, and stakeholders, such as HR professionals, should be informed [36]. It is crucial to identify potential threats to the AI system, including malicious use or misuse, and to assess the risks they pose to ethical and fair recruitment practices [36]. Crucial risks such as unacceptable harm should be mitigated. The AI recruitment tool should be stable and reliable [36]. There should be mechanisms for fault tolerance through backup systems or other parallel systems [36].

8. Fallback plan and general safety
Does the AI recruitment tool incorporate mechanisms for switching to rule-based processing or human oversight in cases of anomalies or system failures, along with fault tolerance mechanisms such as a backup system?

Accuracy is critical because it determines the AI's capability to make correct judgments, such as classifying information correctly and making accurate predictions, recommendations, or decisions based on the data and model it processes [7]. However, there are also other performance metrics suitable depending on the application [36]. For example, accuracy can be used if the class distribution is similar, while the F1-score is a better metric when dealing with imbalanced classes [36]. The AI recruitment tool should inform HR about how reliable or accurate it is expected to be to set realistic expectations and avoid misinformation or overestimation of its capabilities. The AI tool should be developed using complete, high-quality, up-to-date data that accurately represents the environment in which the system will be deployed [36]. This necessitates frequent retraining of the model. Furthermore, the data should come from a reliable source, preferably verified by a third party. Synthetic data can be considered reliable if it is generated using scientifically sound methods, validated against real data, used within the appropriate context, and comes with thorough documentation and transparency. The provider of the AI recruitment tool should consider that the AI tool's operation can invalidate the data or assumptions used in its training, potentially leading to adverse effects such as exacerbating biases or generating misleading information [36].

Implementing a clear, structured development and evaluation process with thorough documentation is essential to manage, mitigate, and correct the unintended risks associated with incorrect predictions. The tool should be capable of assessing and communicating the likelihood of errors when they will inevitably occur [7]. Article 10 of the EU AI ACT also requires that high-risk AI systems' training, validation, and testing data sets, if applicable, be subject to data governance and management [156]. The data shall be relevant, representative, and suitable for the intended purpose, considering context and geographical factors [156]. This includes the examination of potential biases and the evaluation of the availability, quantity, and suitability of the data sets [156]. The AI recruitment tool should be **reliable**, and the outcomes it produces should be **reproducible** [7]. The provider should implement comprehensive verification and validation methods, including detailed documentation and logging, to evaluate the AI system's reliability and reproducibility [36].

9. Accuracy + Reliability and reproducibility
Is the AI recruitment tool's evaluation or prediction accuracy high, reliable, reproducible, and well-documented, using up-to-date, high-quality, complete, and representative data of the current environment [a]?

[a] This also implies that the model should be retrained frequently.

Privacy and data governance:

Tightly associated with the concept of preventing harm is the right to **privacy**, which is impacted by AI systems [7]. Protecting privacy in AI involves implementing strict data management, which includes maintaining data quality, ensuring its relevance, controlling access, and processing data securely to prevent privacy violations. The AI recruitment tool should ensure privacy and **data protection** throughout its entire life cycle [7]. This includes safeguarding the personal information provided by candidates and the data generated about them during their interaction with the tool [7]. It is essential to protect against the unlawful or unfair use of data that could lead to discrimination [7].

The AI tool should respect the human right to physical, mental and moral integrity. Regarding physical integrity, ensuring accessibility and user-friendly interfaces for candidates are crucial but not sufficient. Other aspects should also be considered, such as implementing measures to prevent the theft of physical identity information, including biometric data protection and secure data transmission. Furthermore, the AI tool should safeguard candidates' psychological well-being and cognitive autonomy. While certain assessments may aim to evaluate candidates' ability to handle stress and perform under pressure, candidates should not be subjected to undue stress, manipulation, or unfair treatment during the recruitment process. The AI tool should also respect ethical values, which intersect with e.g., fairness, non-discrimination, bias mitigation, cultural sensitivity (respecting diverse cultural backgrounds and practices), and transparency.

10. Privacy and data protection
Does the AI recruitment tool implement measures to safeguard privacy rights, ensure the protection of physical, mental, and moral integrity, and maintain data security throughout its life cycle[a]?

[a] This includes avoiding invasive practices, preventing the use of data for unlawful or unfair discrimination, and protecting both original and subsequently generated user data.

The **quality of data** used to train the AI tool significantly impacts its performance [7]. Biases tied to gender, ethnicity, or other irrelevant factors could unjustly influence hiring decisions. These biases should be eliminated before training. The tool provider should examine **data integrity** such as verifying the data source. The insertion of harmful data into an AI recruitment tool can potentially disrupt its functionality, especially in self-learning systems [7]. It is important to test and document the processes and data sets used throughout the entire process, including planning, training, testing, and deployment [7]. These records should be accessible for review to ensure accountability and transparency, as discussed later. There are also various standards that could be taken into consideration, such as ISO/IEC 5259 AI Data quality for analytics and machine learning (ML) and the IEEE P7003 Standard for Algorithmic Bias Considerations.

Data access should be regulated by data protocols that specify the authorized individuals for accessing and modifying the data, along with the conditions under which they may do so [7]. Only qualified personnel with the necessary expertise and a legitimate reason should be granted permission to handle individuals' data [7]. If the AI recruitment tool processes personal sensitive data, it must ensure compliance with GDPR or equivalent regulations. Concerning the EU, a Data Protection Impact Assessment (DPIA) is required when there is "a high risk" to other people's personal information [162]. The AI tool should incorporate privacy-by-design and default measures, such as encryption, pseudonymisation, aggregation, and anonymisation, to protect users' data [36]. It should also enable data minimisation, consent withdrawal, objection, and data erasure following GDPR [36]. The tool provider should consider privacy and data protection implications throughout the AI system's life cycle, including the collection, generation, and processing of both personal and non-personal data [36]. Non-personal data can still pose risks such as re-identification, bias, and unauthorized access for misuse. To ensure its reliability, it is advised that the AI recruitment tool conforms to relevant standards (e.g., ISO [163], IEEE[164]) or adheres to widely accepted data management and governance protocols [36].

11. Quality and integrity of data + Access to data
Are the AI recruitment tool's data management and processing in compliance with applicable regulations such as GDPR, and the EU AI ACT and clearly documented?

Transparency:

Transparency is related to the principle of explicability [7]. To achieve transparency, the AI recruitment tool should be **traceable** throughout its entire lifecycle [36]. According to Article 11 of the EU AI ACT, the technical documentation for a high-risk AI system shall be prepared before the system is introduced to the market or put into service, and it shall be maintained up to date [156]. ANNEX IV of the EU AI ACT provides the minimum requirements for technical documentation [156]. Logging practices are beneficial. Also, Article 12 of the EU AI ACT states that high-risk AI systems shall be designed to automatically log events continuously throughout their lifespan [156]. Documentation should be maintained not only for the AI's output decisions but also for its data collection, processing, labelling, and the algorithms used [7]. Such documentation should provide sufficient detail to trace the AI tool's decision-making process from data input to evaluation. This facilitates the detection and correction of errors or biases.

Moreover, the tool provider should commit to continuous quality control and improvement. The EU AI Act requires providers of AI systems to conduct a conformity assessment, ensuring their quality management systems comply with Article 17 [156]. Providers shall review technical documentation to verify that their AI systems meet essential requirements and confirm that the design, development, and post-market monitoring processes align with this documentation, as detailed in Articles 16 and 72 [156]. Additionally, the EU AI Act describes responsibilities across the AI value chain, requiring obligations for other stakeholders including deployers, importers, and distributors of AI systems [156]. An example of a standard concerning transparency is ISO/IEC 42001 - AI Management Systems, which also covers other relevant dimensions, such as human oversight.

12. Traceability
Does the AI recruitment tool have documentation and logging systems in place that enable the traceability of its decision-making process, from data input through to output evaluation?

the AI recruitment tool should be **explainable**, meaning its decision should be understandable to humans [7]. This encompasses not only the technical processes but also the reasoning behind the AI's decisions or predictions [7]. When relevant, the tool can display feature importance and provide global and local explanations, such as SHAP (SHapley Additive exPlanations) values which are also mentioned in Chapter 4. Moreover, explanations should be provided in timely and tailored to the expertise level of the concerned stakeholders such as HR professionals. A trade-off may be necessary between explainability and accuracy [7]. In situations where the AI tool operates as a "black box" with limited explainability, alternative methods to ensure the system's explainability are essential [36]. These methods include ensuring that the tool's operations can be traced, audited and that there is clear communication about what the AI can and cannot do, to help ensure that the AI tool respects fundamental rights [36].

13. Explainability
Does the AI recruitment tool provide explanations of its decisions or predictions that are understandable to decision-makers, regardless of their technical expertise?

At any phase of the recruitment process that uses an AI tool such as a chatbot, it is crucial to **communicate** to all users - not just recruiters but candidates as well - that they are engaging with an AI system [7]. This is essential for upholding fundamental rights, as also mentioned in the section of Human agency [7]. Users should receive clear information about the AI tool's purpose, capabilities, criteria and limitations [7], tailored to the context and role. For example, candidates need to understand the objectives of the AI recruitment tool and its influence on the decision-making process; for further details on candidate communication, refer to sections related to perceived fairness.

Decision-makers should not only receive information about the AI recruitment tool's benefits but also its performance, limitations and potential risks, such as its accuracy level and/or error rates [7]. Before selecting and using a tool, decision-makers should ensure its limitations etc. are acceptable for the intended context. Additionally, the tool provider should provide appropriate training material to users on how to adequately use the AI recruitment tool and receive information for better interpretability. The EU AI ACT specifies in Article 13 transparency and provision of information to deployers [156]. Instructions for the use of high-risk AI systems shall include the provider's identity and contact details, characteristics, capabilities and performance limitations of the high-risk AI system, changes, human oversight measures, computational and hardware resources requirements, and a description of logging mechanisms [156].

14. Communication - minimise confusion
Does the AI recruitment tool clearly inform all users, including both recruiters and candidates, that they are interacting with an AI system?

15. Communication - minimise over-reliance
Does the AI recruitment tool provider establish mechanisms to inform decision-makers (such as recruiters) about the tool's characteristics, capabilities, and limitations, while also providing appropriate training?

Societal and environmental wellbeing:

When considering the impact of AI recruitment tools, it is crucial to look beyond immediate user and business needs to consider the wider **societal** and **environmental** implications. These tools should be designed and developed **sustainably** and with ecological responsibility in mind throughout the entire supply chain [7]. For example, adopting more energy-efficient algorithms and choosing less harmful options [7]. This contributes to a form of fairness that extends to environmental justice. AI tools can contribute to addressing global issues, such as the Sustainable Development Goals, by e.g., improving access to employment opportunities for underrepresented or disadvantaged groups, thereby promoting economic inclusion and social equity. Creating trustworthy systems that do not perpetuate biases or inequalities is crucial. These systems should be continuously improved through feedback from a diverse range of stakeholders.

Providers should establish mechanisms to evaluate the environmental and **social impacts** of their AI recruitment tools throughout their life cycle such as considering resource usage, energy consumption, and their potential to exacerbate social inequalities. The aim is to make these tools as environmentally friendly as possible and minimise potential harm to **society and democracy** [7]. As mentioned in the section on Human agency, AI recruitment tools should support human decision-making rather than replacing it. It should promote meaningful work and enhance the recruitment process [36].

Organisations using AI recruitment tools should self-assess the impact on workers by evaluating e.g., how AI affects job roles, tasks, work arrangements, and the development of skills [36]. Involving and consulting with affected workers is essential to foster transparency and prepare them for changes AI may bring to the work environment [36]. Offering re-skilling and up-skilling opportunities for HR professionals can help them adapt to new technologies and methodologies in recruitment [36]. Understanding how the AI recruitment tool operates, its capabilities, and its limitations is crucial [36]. Furthermore, organisations should establish ongoing mechanisms for monitoring and evaluating the AI recruitment tool's impact on recruitment practices, workplace diversity, and candidate experience. The evaluation should encompass not only the AI tool's effectiveness in task performance and meeting recruitment objectives but also its adherence to ethical standards and social responsibility.

Although not specified in the HLEG's guidelines for trustworthy AI, sustainability can encompass not only environmental and social aspects but also economic sustainability. In the context of selecting an AI recruitment tool, factors such as long-term viability and cost-effectiveness are important in the decision-making process. However, economic sustainability does not necessarily guarantee the promotion of objective fairness. For example, a focus on cost-effectiveness might lead to compromising on fairness measures, since implementing comprehensive fairness initiatives can be expensive and the benefits of fairness are not immediately apparent in financial terms. Although the concepts in the derived question overlap, each aspect is important. Therefore, sustainability, environmental responsibility, social impact, and societal aspects are intentionally listed explicitly to emphasise that all should be considered as an example. Using a broad term would be too vague and might not adequately highlight the significance of each concept. At the same time, it should not be overly complex considering its usability.

16. Sustainable and environmentally friendly AI + Social impact + Society and democracy
Is the AI recruitment tool designed and developed with sustainability in mind such as environmental responsibility and social impact, minimising harm to society?

Accountability:

Accountability is closely related to the principles of fairness, necessitating the establishment of procedures to ensure responsibility and accountability for AI systems and their outcomes throughout their life cycle [7]. The trustworthiness of AI recruitment tools can be enhanced through evaluations conducted by both internal and external auditors [7]. This involves enabling the assessment of algorithms, data, and design processes, which does not necessarily mean that business models and proprietary information should be disclosed [7]. AI recruitment tools should be designed to support independent audits, particularly in areas that may impact fundamental rights or involve safety-critical applications [7]. As emphasised within the dimension of transparency, traceability and documentation are essential for facilitating **auditability**. Documentation should encompass not only technical aspects, such as how the AI recruitment system operates and the sourcing of data but also the design process, including critical decision-making protocols and stakeholder involvement.

17. Auditability
Does the AI recruitment tool enable independent third-party auditing e.g., of its algorithms, data, and design processes [a]?

[a] without necessarily disclosing proprietary business information. For example, can external auditors access necessary logs, documentation, or review mechanisms that allow them to evaluate the tool's compliance with ethical standards and regulations? Notably, a history of successful audits would be advantageous.

Ensuring both the ability to document and explain the rationale behind an AI system's results, as well as addressing the consequences of those outcomes, is crucial [7]. AI recruitment tools should incorporate mechanisms to identify, assess, document, and **mitigate** potential **negative impacts**, such as bias in candidate selection, unfair treatment, or discrimination [7]. Conducting algorithmic impact assessments, including red teaming or other forms of Algorithmic Impact Assessment throughout the AI tool's life cycle, is beneficial [7]. These assessments should match the level of risk posed by the AI systems [7].

As previously mentioned, under the EU AI Act, AI tools used for recruitment should be classified as high-risk due to their potential impact on individuals' career opportunities, livelihoods, and workers' rights [156]. Nevertheless, AI recruitment tools serve various functions and purposes. For example, AI systems that handle automated interview scheduling are less critical compared to those that make hiring recommendations. Ideally, all AI recruitment tools should be managed comprehensively. Considering the practicability, tools involved in critical decision-making processes require more extensive measures to mitigate associated risks effectively.

18. Minimisation and reporting of negative impacts
Does the AI recruitment tool undergo regular impact assessments, such as red teaming or forms of Algorithmic Impact Assessment, to identify, document, and mitigate potential negative impacts?

Conflicts between different requirements may lead to inevitable **trade-offs**, such as the trade-off between explainability and accuracy, as previously mentioned [7]. Addressing these trade-offs rationally and methodically, following current best practices, is vital [7]. This involves recognising the interests and values affected by the AI system and, in cases of conflict, clearly acknowledging and assessing these trade-offs against their impact on ethical standards and fundamental rights [7]. The development, deployment, and use of AI recruitment tools should not continue in that manner if no ethically acceptable trade-offs can be found [7]. Decisions on which compromises to accept should be well-reasoned and thoroughly documented, with the responsible individual(s) accountable for how these determinations are made [7]. This person or these persons should also regularly review the appropriateness of these decisions to ensure the system can be adjusted as needed [7].

19. Trade-offs
Is there a systematic and transparent process for assessing and addressing trade-offs among various requirements, ensuring that decisions on compromises are ethically acceptable, well-reasoned, and documented?

AI tools should provide clear and accessible means for users to seek **redress** for unjust adverse impacts, paying special attention to vulnerable individuals or groups [7]. This includes offering accessible channels for user feedback and ensuring that feedback reflecting an actual issue leads to meaningful revisions of the AI recruitment tool [7]. The AI recruitment tool provider can establish an AI ethics review board or a similar mechanism to oversee the tool's ethical adherence and ambiguous areas [7]. Implementing a process for continuous monitoring and assessment, such as against the Assessment List for Trustworthy AI (ALTAI), is advisable [36]. Providers should also pursue internal risk training and possibly seek external guidance or third-party audits to ensure ethical compliance [36]. It is also recommended that the adopter establish an internal controlling entity to oversee AI ethics and regulatory issues.

20. Redress
Does the AI recruitment tool offer a clear and accessible process for users to report issues (such as potential vulnerabilities, biases, or risks), and does it have an effective mechanism to address and mitigate these concerns?

5.2 Requirements for Promoting Perceived Fairness

After elaborating on the tool requirements that promote objective fairness in AI recruitment, this thesis proceeds to derive additional tool requirements aimed at enhancing subjective or rather perceived fairness within the AI recruitment process.

Formal characteristics:

A key factor influencing procedural fairness perceptions in selection processes is **job relatedness**, i.e. how closely a test appears to measure content relevant to the job situation or perceived validity [56]. The perceived validity can be captured in both content validity, which examines how test content aligns with job requirements, and criterion-related validity, which assesses how well test performance predicts job performance [165, 56]. It is crucial to distinguish job relatedness from face validity, which refers to the test's superficial appearance of measuring what it claims to measure [166, 56]. When AI recruitment tools are used to assess candidates, they should be designed with criteria directly relevant to the job. This requires the use of algorithms to evaluate skills, experiences, or qualifications that are essential for the position, ensuring that the selection criteria align with job requirements. The design of the system should be scientifically based to confirm its validity. To demonstrate this, the tool provider can offer evidence through means such as references, studies, and results from pilot tests.

21. Job relatedness
Is the AI recruitment tool scientifically based, and when used for assessments, are its evaluations directly relevant to the job and predictive of job performance?

The **opportunity to perform** and express oneself before a decision is made positively affects perceptions of fairness [60, 56]. AI recruitment tools should provide candidates with an equal chance to showcase their abilities and qualifications. This means ensuring the tool is accessible to diverse applicants, including those with disabilities, in accordance with HLEG guidelines on accessibility and universal design. Incorporating various assessment methods enables candidates to display a broad spectrum of relevant skills and competencies.

22. Opportunity to perform
Does the AI recruitment tool ensure equal opportunities for all candidates to showcase their abilities and qualifications through accessible, diverse assessment methods?

Another factor enhancing perceived fairness in recruitment is providing candidates the **opportunity** to challenge or seek **reconsideration** of the decision-making evaluation process [167, 62, 168, 56]. AI recruitment tools should not only grant HR professionals access to results but also allow candidates to review their scores [169, 56]. This approach promotes transparency, enabling candidates to better understand their assessments and thereby enhancing their perception of fairness. Providing access to assessment results builds a foundation for challenging and modifying decisions. To further improve perceived fairness, it is crucial to offer candidates the chance to contest decisions, particularly to address fairness and bias concerns. This can also be done through feedback collection after the assessment using other tools. When appropriate, a second evaluation with human oversight can be introduced to ensure fairness and accuracy. Human review can provide insights that AI might overlook, leading to a more thorough assessment.

23. Reconsideration Opportunity
Does the AI recruitment tool enable candidates to access their evaluation results, offering role-specific information?

AI recruitment decision procedures should ensure **consistency** and equal treatment for all candidates across diverse demographics. This requires standardised assessments, equal access to information about the process, and uniform evaluation of candidates' responses. AI recruitment tools should use algorithms that apply consistent criteria and weighting, thereby avoiding arbitrary or biased decision-making. Not only should the process be consistent, but the outcome should also be consistent when possible. The outcome perspective is covered in the subdimension of Accuracy, Reliability, and Reproducibility. Additionally, the tool provider should regularly audit the tool's performance to maintain consistency over time and across different candidate groups, correcting any deviations from established fairness standards.

24. Consistency
Does the AI recruitment tool apply consistent standards and procedures for evaluation across all demographics to ensure unbiased decision-making and adherence to established fairness standards?

Explanation:

To enhance perceptions of interactional justice, the AI tool should be designed to offer timely and informative **feedback** [170, 56]. This involves providing insights into candidates' performance in assessments or interactions with the AI, as well as ensuring transparency in feedback mechanisms. Offering automated, personalised, constructive feedback would be advantageous. Additionally, the tool should be capable of explaining decisions or assessment outcomes understandably and helpfully to users, possibly providing guidance on areas for improvement. While individuals may perceive understandability differently, a common baseline can be considered.

Some AI screening tools currently in use automatically reject candidates deemed not a good fit based on predefined algorithms. To enhance transparency, it would be better to provide feedback indicating the specific categories—such as skillset mismatch or other criteria—that led to the decision. Geographical location is used as a hard criterion in LinkedIn for AI screening. While restricting candidates based on region may streamline the hiring process, it risks overlooking highly qualified individuals who are cross-border workers or possess valid work permits for the job location. This issue is particularly significant within the EU, which allows for the free movement of people and the freedom to work in any Member State. Any EU citizen is entitled to equal treatment in recruitment [171]. AI recruitment tools should be designed with options that allow for additional candidate information to be considered. For example, candidates should have the opportunity to indicate their legal eligibility to work in the target region. It is also recommended for applicants to proactively provide information that might influence their suitability for a position.

25. Feedback
Does the AI recruitment tool provide timely, informative, and understandable feedback to users?

The perceptions of fairness in selection processes are influenced by **information** regarding the validity of these processes, the clarity of scoring methods, the application of scores in decisions, and the rationale behind decisions[56]. Lounsbury et al. observed that individuals showed more positive attitudes towards testing when they understood its relevance to future job performance [172]. Presenting evidence of validity is especially critical for tests that may not appear relevant at first glance, such as cognitive ability assessments, to enhance their acceptance [173]. Prior information about the selection procedure plays a critical role in shaping perceptions of fairness [56]. To enhance perceived fairness, the AI recruitment tool, when evaluating candidates, should be capable of offering explanations regarding its process, as well as its validity. Should the tool lack the capability to self-explain, HR professionals should supply this information to candidates to foster perceived fairness. Moreover, it is essential to communicate the role of AI in the assessment process transparently. Despite the integration of AI, humans should remain the final decision-makers, in accordance with the High-Level Expert Group (HLEG) guidelines on human agency [7].

26. Selection information
Does the AI recruitment tool provide transparent and understandable explanations of its process and validity to candidates?

Honesty and truthfulness are crucial in communications with candidates, as emphasised by Gilliland [174, 56]. In the context of AI, honesty refers to an AI's commitment to making statements it believes to be true, avoiding deception. Truthfulness involves an AI system's efforts to truthfully describe the world, essentially focusing on the accuracy and reliability of the information it provides relative to the actual state of the world [175]. This includes providing information that is grammatically correct, accurate, and relevant. Ideally, AI recruitment tools should embody both honesty and truthfulness in interactions with humans. Evans et al. recommend prioritising truthfulness, acknowledging that perfect truthfulness is unattainable [175]. Developers can consider measures such as preventing negligent falsehoods, which occur when an AI system makes avoidable errors by not accurately using available information [175].

In the context of generative AI systems such as chatbots, only improvements in prompt engineering and retrieval-augmented generation are insufficient. AI guardrails are also essential, which can monitor and filter the input and output of LLMs [176, 177]. For example, guardrails can prevent LLMs from handling harmful requests or adjust their outputs to align with the deployer's specific moral requirements [177]. Some existing guardrails include Nvidia NeMo and LlamaGuard, both of which use simple approaches [176]. However, the design of effective guardrails also faces several challenges [176]. To address these, Dong et al. recommended a multidisciplinary approach [176].

Furthermore, the information provided by AI should also be useful. Misleading candidates, intentionally or not, can harm the fairness perceptions and the organisation's image. The tool provider should accurately represent the tool's capabilities and limitations, avoiding overstatements about its effectiveness.

27. Honesty
Does the AI recruitment tool have measures to promote truthfulness and usefulness in communication with candidates, thereby ensuring the delivery of accurate and relevant information while preventing the misleading of candidates?

Interpersonal treatement:

Gilliland identifies **interpersonal effectiveness** as a factor influencing perceived fairness, referring to the extent to which candidates are engaged with warmth and respect [56].

AI recruitment tools should use inclusive and respectful language and tone in their communications with candidates.

28. Interpersonal effectiveness
Does the AI recruitment tool use inclusive, respectful language and tone in communication with candidates?

Two-way communication involves allowing candidates to provide input or to have their opinions considered during the selection process [170, 56]. AI recruitment tools should facilitate two-way communication, particularly in relevant scenarios such as chatbots. Beyond the user-friendliness of these tools, when candidates offer inputs, they should receive responses based on the interactions. Furthermore, two-way communication also implies allowing candidates to inquire about the job, the organisation, or the selection process[56]. In cases where AI tools are not designed for such interactions, HR professionals can conduct interactive Q&A sessions in advance using other tools. This approach enables candidates to gather information, voice concerns and make informed decisions. It is also advisable to have staff available to address issues during the evaluation process with AI tools. While implementing two-way communication throughout all phases of the recruitment process may not be feasible, it is important to identify which steps critically require this feature and whether it should be integrated into the AI tools.

29. Two-way communication
Does the AI recruitment tool support two-way communication in relevant cases?

Bies and Moag indicated that the **propriety of questions** asked during recruitment impacts recruitees' perceptions of fairness [63, 56]. It is important to avoid improper questioning and prejudicial statements [56]. The suppression of personal bias is highlighted as a crucial element of procedural justice by several scholars [62, 168, 170, 56]. Moreover, the propriety of questions is linked to **perceived invasion of privacy** which may affect fairness perceptions[178, 56]. This implies that AI recruitment tools should incorporate measures to mitigate inappropriate bias in questioning, ensuring that questions are relevant, job-related, non-discriminatory, and respect privacy rights. The tool provider should regularly audit and update the questions database to ensure compliance with fairness and legal standards. These aspects are integrated into the questions described previously to avoid redundancy and ensure coherence.

Additional rules:

Ease of faking answers pertains to the degree to which applicants perceive the possibility of manipulating their responses in a socially desirable manner throughout the selection process [56]. In evaluating candidates, the AI recruitment tool should have mechanisms to minimise the potential for the fabrication or manipulation of responses. This might include the use of advanced algorithms capable of detecting inconsistencies or patterns indicative of less honest responses. The AI recruitment tool should rely on a wide range of data points. This can be incorporating a variety of assessment methods when evaluating candidates and conducting cross-checking.

30. Ease of faking answers
Does the AI recruitment tool incorporate mechanisms to minimise the potential for candidates to fabricate or manipulate their responses?

Distributive justice encompasses rules of equity, equality and need [56]. The equity distribution rule posits that perceptions of fairness are influenced by the comparisons between expected and actual outcomes [56]. In the context of AI recruitment tools, this implies that their outcomes should align with the contributions of individuals, highlighting the importance of consistent evaluation and transparency to avoid misunderstandings. Violations of equality such as decisions based on irrelevant characteristics, are perceived as unfair. This reinforces the necessity for AI recruitment tools to focus on job-relatedness and the mitigation of inappropriate biases to prevent discrimination based on gender or other irrelevant factors. The needs distribution rule can refer to preferential treatment for disadvantaged groups, such as affirmative action or accommodations for individuals with disabilities. For AI recruitment tools, this emphasises the importance of accessibility and the opportunity to perform.

No new dimensions are added, as the relevant ones have already been covered in the preceding dimensions. Organisations using AI recruitment tools can assess whether their procedural fairness measures effectively achieve the intended outcomes and uphold principles of distributive justice. Additionally, defining key performance indicators (KPIs) for monitoring purposes can support the measurement of improvements and the actual impact of these tools.

Satisfaction of one rule may lead to the violation of another, posing a challenge in achieving perceived fairness in the selection process [56]. When selecting AI recruitment tools, decision-makers can prioritise requirements based on the use case. Additionally, aspects such as evolving laws, regulations, standards, organisational values and policies should also be considered. The mentioned rules proposed by Gilliland account for much of the variance in the perceptions of fairness of selection systems [56]. While the possibility of additional rules exists, it is essential to avoid overcomplication and ensure the scope of the thesis is maintained. Factors influencing perceived fairness include, for example, also candidates' past experiences in similar processes [56]. Limiting the number of guiding questions is essential to preserve practicality, avoiding them being deemed too laborious for use.

Finally, an additional question has been included to evaluate ongoing assessment and improvement across the aforementioned dimensions. For simplification purposes, this thesis treats AI recruitment tool providers as unified entities, even though some may outsource the development of their products et cetera. Nonetheless, all entities involved should fulfil the necessary criteria to ensure that the end product meets the requirements.

31. Ongoing evaluation and improvement
Does the AI recruitment tool undergo continual performance evaluation, risk assessment, auditing, and improvement across the aforementioned dimensions to ensure adherence to legal and ethical standards?

CHAPTER 6

Evaluation

Chapter 6 presents an evaluation of the developed artefact. Section 6.1 assesses its correctness and completeness by comparing it to research with similar objectives. Section 6.2 examines the artefact's practicality and effectiveness in identifying issues through case studies in three areas: CV screening, chatbot, and video interview. In each application area, one specific tool is analysed in detail using the artefact, followed by a discussion of the artefact's evaluation. Then, a comparison between the tools' assessment results is conducted, leading to the final conclusions of the artefact's evaluation.

6.1 Framework Alignment

There is no specific framework that details the requirements for promoting fairness in AI recruitment tools. Research on fairness in AI-based recruitment primarily focuses on topics related to avoiding discrimination and bias, and ensuring algorithmic fairness (e.g. [179, 5]). Similar findings are noted in the paper [67]. These aspects are already included in the dimensions of the built artefact. The developed artefact is divided into two main components: the first refers to the principles of trustworthy AI that support objective fairness, while the second addresses subjective, perceived, fairness in the recruitment selection process. The completeness of each component is evaluated.

As revealed through the database search results presented in Chapter 4.2.3, numerous guidelines exist for the responsible design and governance of AI. The guidelines for trustworthy AI from HLEG cover a wider range of dimensions compared to the other six guidelines within the top six most frequently mentioned, as illustrated in Table 4.2. Although the perspectives of the six guidelines exhibit some differences, this variation is to be expected given their origins in governance, science, and industry. Despite these differing viewpoints, the overall theme aligns, showcasing a similar direction such as 1.Human rights, agency, and oversight; 2.Technical robustness, safety, and security; 3.Privacy and data governance; 4.Transparency, 5.Diversity, non-discrimination, and fairness; 6.Societal and environmental wellbeing; 7.Accountability and responsibility; 8.Awareness, education, and discussion. These dimensions are interconnected. The relevant ones are already included in the artefact. Considering the potential cultural variations, well-known guidelines from China and the United States are analysed as they belong to leading countries in AI research and technology[9, 151, 152, 153]. The comparison revealed no additional relevant dimensions for the purpose of this thesis.

Comparison was also made to studies examining principles and guidelines, but no further relevant dimensions were identified. For example, Jobin et al. analysed 84 global principles and guidelines concerning ethical AI and identified 11 categories: Transparency, Justice and fairness, Non-maleficence, Responsibility, Privacy, Beneficence, Freedom and Autonomy, Trust, Sustainability, Dignity, and Solidarity [51]. No additional relevant dimensions were found for the objective of this thesis. Hagendorff mapped 21 major AI ethics guidelines into categories [154]. He mentioned cultural differences explicitly, which fall into the Artefact's Diversity, non-discrimination and fairness dimension. Again, no new relevant dimensions aligning with the thesis's goal were identified. While this artefact is aimed at being relevant in the EU, it may also serve as inspiration outside the EU. However, adherence to regional law and other region-dependent factors should be considered.

In terms of perceived fairness in recruitment selection, Gilliland's framework is a well-established model in organisational fairness and explains most of the variance in perceived fairness [56]. Several papers in the area of AI-assisted recruitment cited his model, such as [55], where the authors also advocated for considering both objective fairness perceptions and subjective fairness perceptions among applicants and employees regarding algorithmic decision-making. Yu et al. also used Gilliland's model to explain perceived fairness and found that meeting the requirements of consistency, voice, explainability and human involvement could enhance applicants' perception of fairness in AI-based hiring decisions [180]. The fairness rules for talent intelligence management systems proposed by Zhang et al. included consistency, representativeness, bias suppression, accuracy, correctability, ethicality, interactivity, and explanation [181]. These rules were also based on distributive justice, procedural justice, and interpersonal justice [181]. No additional relevant dimensions were identified.

Perceptions of fairness can vary among stakeholders, which might include individuals from departments such as Strategy & Corporate Governance, Human Resources and Operations, Procurement, and Information Technology. For example, a stakeholder focused on the organisation's strategy might perceive an AI recruitment tool as a fair choice if it aligns with strategic goals. Others might judge its fairness based on its ability to improve hiring quality, its cost-effectiveness, or its ability to integrate with current systems. These perspectives are typically those that decision-makers already have in mind. This artefact aims to focus on the objective fairness of the tool and its perceived fairness by candidates. The perception of fairness among candidates can impact their behaviour during the hiring process—this includes decisions to apply or accept job offers, willingness to recommend the application process to others, motivation during testing phases, and potential for legal disputes [56]. It also influences post-hiring outcomes such as performance, engagement in organisational citizenship behaviours, job satisfaction, and the overall organisational climate [56]. No new dimensions are found through searches conducted with perceived fairness in AI recruitment selection.

During the Iteration, the guiding questions were changed from an open-ended to a yes/no format to streamline the evaluation process for shortlisting purposes. However, these questions serve only as preliminary indicators and a discussion of the underlying factors is encouraged. In evaluating use cases, the binary yes/no answer option was changed to a scale of one to five stars for better measurement, as shown in Table 6.1: One filled star, Barely Fulfilled, means only a small portion of the requirements are met, with significant improvement needed. Two filled stars, Partially Fulfilled, indicates some criteria are met, but considerable gaps remain. Three filled stars represent Moderately Fulfilled, while four filled stars, Largely Fulfilled, signifies that most criteria are met with only minor shortcomings. Finally, five filled stars denote Fully Fulfilled. This change acknowledges that not everything can be clearly identified or meet every ideal standard. In practice, organisations may need to prioritise the dimensions to find the most appropriate recruitment tool, considering their values, policies, and further applicable laws and regulations.

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool	Comments
1	Trustworthy AI, Procedural justice rules	Diversity, non-discrimination and fairness, interpersonal treatment	Avoidance of bias, Propriety of questions	Does the AI tool have an appropriate, justified fairness definition and effective strategies implemented to ensure fairness, including mitigating inappropriate bias in data, algorithm design, and outputs?	☆☆☆☆☆	
2	Trustworthy AI	Diversity, non-discrimination and fairness	Accessibility and universal design	Is the AI recruitment tool user-centric, adhering to universal design, and accessible to all users, including those with disabilities?	☆☆☆☆☆	

Continued on next page

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
3	Trustworthy AI	Diversity, non-discrimination and fairness	Stakeholder participation	Does the design, development, and continuous improvement of the AI recruitment tool incorporate input from a diverse range of stakeholders, such as job applicants, HR professionals, legal experts, and ethicists?	☆☆☆☆☆	
4	Trustworthy AI	Human agency and oversight	Fundamental rights	Does the AI recruitment tool undergo evaluation to identify and mitigate potential risks[1] to fundamental rights[2] during its design, development, and operational phases?	☆☆☆☆☆	
5	Trustworthy AI	Human agency and oversight	Human agency	Does the AI recruitment tool focus on human autonomy by empowering users to make informed autonomous decisions, including implementing safeguards against manipulation, over-reliance, and confusion between AI and human interactions?	☆☆☆☆☆	
6	Trustworthy AI	Human agency and oversight	Human oversight	Does the AI recruitment tool incorporate mechanisms for effective human oversight, including options for modification to prevent inappropriate outcomes?	☆☆☆☆☆	

Continued on next page

[1] Certain risks can be justified as necessary in a democratic society to ensure the preservation of balanced rights and freedom.

[2] In the EU, it includes human dignity, individual freedom, democracy, justice, the rule of law, equality, non-discrimination, solidarity, and citizens' rights.

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
7	Trustworthy AI	Technical robustness and safety	Resilience to attack and security	Is the AI recruitment tool certified for security or compliant with security standards, including having measures to counter cyberattacks[3] through ongoing security updates and penetration testing?	☆☆☆☆☆	
8	Trustworthy AI	Technical robustness and safety	Fallback plan and general safety	Does the AI recruitment tool incorporate mechanisms for switching to rule-based processing or human oversight in cases of anomalies or system failures, along with fault tolerance mechanisms such as a backup system?	☆☆☆☆☆	
9	Trustworthy AI	Technical robustness and safety	Accuracy, reliability, and reproducibility	Is the AI recruitment tool's prediction accuracy high, reliable, reproducible, and well-documented, using up-to-date, high-quality, complete, and representative data of the current environment[4]?	☆☆☆☆☆	
10	Trustworthy AI, Procedural justice rules	Privacy and data governance, additional rules	Privacy and data protection, Perceived invasion of privacy	Does the AI recruitment tool implement measures to safeguard privacy rights, ensure the protection of physical, mental, and moral integrity, and maintain data security throughout its life cycle[5]?	☆☆☆☆☆	

Continued on next page

[3] Such as data poisoning, model evasion, and model inversion.

[4] This also implies that the model should be retrained frequently.

[5] This includes avoiding invasive practices, preventing the use of data for unlawful or unfair discrimination, and protecting both original and subsequently generated user data.

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
11	Trustworthy AI	Privacy and data governance	Quality and integrity of data, access to data	Are the AI recruitment tool's data management and processing in compliance with applicable regulations such as GDPR, and the EU AI Act and clearly documented?	☆☆☆☆☆	
12	Trustworthy AI	Transparency	Traceability	Does the AI recruitment tool have documentation and logging systems in place that enable the traceability of its decision-making process, from data input through to output evaluation?	☆☆☆☆☆	
13	Trustworthy AI	Transparency	Explainability	Does the AI recruitment tool provide explanations of its decisions or predictions that are understandable to decision-makers, regardless of their technical expertise?	☆☆☆☆☆	
14	Trustworthy AI	Transparency	Communication - minimise confusion	Does the AI recruitment tool clearly inform all users, including both recruiters and candidates, that they are interacting with an AI system?	☆☆☆☆☆	
15	Trustworthy AI	Transparency	Communication - minimise over-reliance	Does the AI recruitment tool provider establish mechanisms to inform decision-makers (such as recruiters) about the tool's characteristics, capabilities, and limitations, while also providing appropriate training?	☆☆☆☆☆	

Continued on next page

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
16	Trustworthy AI	Societal and environmental well-being	Sustainable and environmentally friendly AI, social impact, society and democracy	Is the AI recruitment tool designed and developed with sustainability in mind such as environmental responsibility and social impact, minimising harm to society?	☆☆☆☆☆	
17	Trustworthy AI	Accountability	Auditability	Does the AI recruitment tool enable independent third-party auditing (e.g., of its algorithms, data, and design processes)[6]?	☆☆☆☆☆	
18	Trustworthy AI	Accountability	Minimisation and reporting of negative impacts	Does the AI recruitment tool undergo regular impact assessments, such as red teaming or forms of Algorithmic Impact Assessment, to identify, document, and mitigate potential negative impacts?	☆☆☆☆☆	
19	Trustworthy AI	Accountability	Trade-offs	Is there a systematic and transparent process for assessing and addressing trade-offs among various requirements, ensuring that decisions on compromises are ethically acceptable, well-reasoned, and documented?	☆☆☆☆☆	

Continued on next page

[6] Without necessarily disclosing proprietary business information. For example, can external auditors access necessary logs, documentation, or review mechanisms that allow them to evaluate the tool's compliance with ethical standards and regulations? Notably, a history of successful audits would be advantageous.

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
20	Trustworthy AI	Accountability	Redress	Does the AI recruitment tool offer a clear and accessible process for users to report issues (such as potential vulnerabilities, biases, or risks), and does it have an effective mechanism to address and mitigate these concerns?	☆☆☆☆☆	
21	Procedural justice rules	Formal characteristics	Job relatedness	Is the AI recruitment tool scientifically based, and when used for assessments, are its evaluations directly relevant to the job and predictive of job performance?	☆☆☆☆☆	
22	Procedural justice rules	Formal characteristics	Opportunity to perform	Does the AI recruitment tool ensure equal opportunities for all candidates to showcase their abilities and qualifications through accessible, diverse assessment methods?	☆☆☆☆☆	
23	Procedural justice rules	Formal characteristics	Reconsideration opportunity	Does the AI recruitment tool enable candidates to access their evaluation results, offering role-specific information?	☆☆☆☆☆	
24	Procedural justice rules	Formal characteristics	Consistency	Does the AI recruitment tool apply consistent standards and procedures for evaluation across all demographics to ensure unbiased decision-making and adherence to established fairness standards?	☆☆☆☆☆	
25	Procedural justice rules	Explanation	Feedback	Does the AI recruitment tool provide timely, informative, and understandable feedback to users?	☆☆☆☆☆	

Continued on next page

No.	Underlying concepts	Dimensions	Subdimensions	Questions	Tool Name	Comments
26	Procedural justice rules	Explanation	Selection information	Does the AI recruitment tool provide transparent and understandable explanations of its process and validity to candidates?	☆☆☆☆☆	
27	Procedural justice rules	Explanation	Honesty	Does the AI recruitment tool have measures to promote truthfulness and usefulness in communication with candidates, thereby ensuring the delivery of accurate and relevant information while preventing the misleading of candidates?	☆☆☆☆☆	
28	Procedural justice rules	Interpersonal treatment	Interpersonal effectiveness	Does the AI recruitment tool use inclusive, respectful language and tone in communication with candidates?	☆☆☆☆☆	
29	Procedural justice rules	Interpersonal treatment	Two-way communication	Does the AI recruitment tool support two-way communication in relevant cases?	☆☆☆☆☆	
30	Procedural justice rules	Additional rules	Ease of faking answers	Does the AI recruitment tool incorporate mechanisms to minimise the potential for candidates to fabricate or manipulate their responses?	☆☆☆☆☆	
31	All	All	Ongoing evaluation and improvement	Does the AI recruitment tool undergo continual performance evaluation, risk assessment, auditing, and improvement across the aforementioned dimensions to ensure adherence to legal and ethical standards?	☆☆☆☆☆	

Table 6.1: Questions for AI recruitment tool assessment based on [7, 36, 56]

6.2 Case Studies

To evaluate the artefact's practicality and effectiveness in identifying issues, three case studies are conducted across different AI applications: CV screening, chatbot, and video interview. Each case study begins with an explanation of how AI is applied within the specific area, followed by a discussion of examples of potential benefits and challenges. The selected AI tool (selection criteria outlined in Chapter 3, Methodology) and its provider are then briefly introduced. The tool is analysed using the artefact's criteria. Subsequently, the artefact's effectiveness is examined through a literature review to assess existing studies on the tool and determine whether the artefact could identify previously documented issues, if any. It is important to note that the scoring system used in this thesis's case studies relies on the accuracy of publicly available official information. However, these claims may not always be consistently accurate or truthful, and verification with the provider, as well as actual testing of the tool, should be done.

6.2.1 AI-powered CV screening

CV screening, also known as resume screening, occurs at the early stage of recruitment and aims to shortlist applicants based on their CVs for a potential role[182]. Traditionally, HR professionals manually searched through a pool of applicants to decide who should advance to the next stage. However, technology has evolved: algorithms were developed to scan resumes for specific keywords and phrases [67]. Nowadays, AI technologies extend beyond simple keyword matching. Current tools, including chatbots and resume parsers, can assess candidates' suitability by searching for semantic connections and related terms [67]. Additionally, some AI tools can predict a candidate's potential job performance by analysing indicators related to tenure or productivity, or the absence of indicators associated with tardiness or disciplinary issues [67]. Moreover, certain algorithms can recommend the most suitable job vacancies for candidates [67].

CV screening tools are viewed as highly efficient in simplifying the process, particularly for organisations that receive a large volume of applications for each position [67]. These tools can be time-saving and cost-effective [182]. AI-assisted CV screening can help mitigate certain forms of human bias. In manual CV evaluations, HR professionals may possess unconscious biases, such as implicit bias [182]. Implicit bias refers to the unconscious attitudes and stereotypes that influence perceptions and behaviours towards specific groups [182]. Such bias can lead to hiring discrimination when individuals unconsciously favour or discriminate against candidates based on characteristics unrelated to job positions, such as race, gender, age, or even the name or appearance on a CV.

While CV screening tools may hide certain characteristics, other types of biases, such as algorithmic bias, can still exist. These biases may result in highly qualified applicants being overlooked [67]. For instance, Amazon's former resume screening tool, which used NLP and ML, was designed to identify top job candidates by learning from the resumes of successful applicants and looking for similar characteristics in new submissions [183, 184]. By the end of 2014, the tool was widely used within the company as it significantly saved time [184]. However, by 2015, it was discovered that the tool did not evaluate resumes for technical positions, like software developers, in a gender-neutral way [183, 184]. This issue stemmed from the training data, predominantly composed of resumes from male employees, reflecting the male dominance in the tech industry at that time [183, 184]. Consequently, the tool developed biased associations, such as downgrading resumes mentioning "women's" groups or all women's colleges [183, 184]. Amazon's engineers modified the algorithms to mitigate the bias[183, 184]. Nevertheless, it remains a risk that such AI systems could still develop new forms of bias and inadvertently discriminate against certain candidates based on the data they process and the patterns they learn [183, 184]. To address algorithmic biases, Albaroudi et al. highlighted two primary AI techniques: vector space correction and data augmentation [182].

However, CV screening tools, like other AI recruitment tools, pose ethical risks, including privacy concerns [67]. Decision-makers in organisations should be aware of these risks and the limitations of these tools. In addition to AI regulations, ethical thinking and conscious use of AI tools should be promoted [185, 67]. organisations should ensure that their recruiters receive adequate training in these areas to foster an ethical and unbiased hiring process [185, 67].

CVViZ

For the evaluation, CVViZ was chosen. CVViZ is an AI-powered, cloud-based recruitment software founded in 2017 [186]. This platform automates candidate sourcing and matches suitable candidates with appropriate positions [186, 187]. It provides insights into the recruitment process and aims to enhance the quality of hiring [187]. CVViZ envisions bringing "intelligent automation in the hiring space where it helps companies and recruiters in optimizing their hiring efforts" [188]. Its mission is "making recruiters lives simpler with awesome software" [189]. CVViZ offers solutions ranging from advertising jobs to making job offers [187]. It is used by hundred of companies, from startups to large enterprises, with a company size of 11-50 employees, and is located in Mumbai, Maharashtra [187, 186].

For the purposes of this thesis, the analysis focuses on the resume screening component of CVViZ that uses natural language processing (NLP) and machine learning algorithms to enhance the candidate-job matching process[187]. It analyses uploaded resumes contextually and learns continuously from past and current recruitment activities [187]. CVViZ not only reviews historical recruitment data but also adjusts its criteria based on the outcomes of recent candidate evaluations—approving or rejecting applications[187]. This AI-driven method aims to identify the most suitable candidates through relative resume ranking, whom recruiters can then contact [190, 187].

Tool assessment

Table 6.2 provides an overview of the assessment results of CVViz.

No.	Dimensions	Subdimensions	CVViZ
1	Diversity, non-discrimination and fairness, interpersonal treatment	Avoidance of bias, Propriety of questions	★☆☆☆☆
2	Diversity, non-discrimination and fairness	Accessibility and universal design	★☆☆☆☆
3	Diversity, non-discrimination and fairness	Stakeholder participation	★☆☆☆☆
4	Human agency and oversight	Fundamental rights	☆☆☆☆☆
5	Human agency and oversight	Human agency	☆☆☆☆☆
6	Human agency and oversight	Human oversight	★★☆☆☆
7	Technical robustness and safety	Resilience to attack and security	★☆☆☆☆
8	Technical robustness and safety	Fallback plan and general safety	★☆☆☆☆
9	Technical robustness and safety	Accuracy, Reliability and reproducibility	☆☆☆☆☆
10	Privacy and data governance, additional rules	Privacy and data protection, Perceived invasion of privacy	★★★★☆
11	Privacy and data governance	Quality and integrity of data, Access to data	★★★★★
12	Transparency	Traceability	☆☆☆☆☆
13	Transparency	Explainability	☆☆☆☆☆
14	Transparency	Communication - minimise confusion	☆☆☆☆☆
15	Transparency	Communication - minimise over-reliance	★★☆☆☆
16	Societal and environmental well-being	Sustainable and environmentally friendly AI, Social impact, Society and democracy	☆☆☆☆☆
17	Accountability	Auditability	☆☆☆☆☆
18	Accountability	Minimisation and reporting of negative impacts	☆☆☆☆☆
19	Accountability	Tradeoffs	☆☆☆☆☆
20	Accountability	Redress	★★★☆☆
21	Formal characteristics	Job relatedness	★★★★☆
22	Formal characteristics	Opportunity to perform	★★★★☆
23	Formal characteristics	Reconsideration opportunity	N/A
24	Formal characteristics	Consistency	☆☆☆☆☆
25	Explanation	Feedback	N/A
26	Explanation	Selection information	N/A
27	Explanation	Honesty	N/A

Continued on next page

No.	Dimensions	Subdimensions	CVViZ
28	Interpersonal treatment	Interpersonal effectiveness	N/A
29	Interpersonal treatment	Two-Way communication	N/A
30	Additional rules	Ease of faking answers	☆☆☆☆☆
31	All	Ongoing evaluation and improvement	★☆☆☆☆

Note: N/A = not applicable

Table 6.2: CVViZ assessment results overview by dimensions based on [7, 56]

Starting with the dimension of **Diversity, non-discrimination and fairness**, particularly with the subdimension of Bias avoidance, the definition of fairness was lacking. Regarding strategies to ensure fairness, CVViZ only stated that “Influential factors such as demographic details can be completely ignored while screening candidates” [191]. It claimed that using AI for resume screening could help eliminate unconscious bias [191]. However, bias can manifest in other forms beyond demographics, including in data, algorithm design and outputs. No information on these potential biases was provided. Regarding the subdimension of Accessibility and universal design, CVViz only noted that their “recruiting tools are modern, intuitive and easy to adapt” [192]. However, it left questions about accessibility for people with disabilities and other aspects of universal design unanswered. As for the subdimension of Stakeholder participation, the information provided was limited to the overall team composition. It mentioned having “recruited engineers, programmers, marketing and sales people for his organizations” but did not specify the involvement of legal experts or ethicists[191].

In the dimension of **Human agency and oversight**, CVViZ did not specify whether it conducted evaluations to identify and mitigate potential risks to Fundamental rights during its design, development, and operational phases. Additionally, there was no available information on whether the tool focused on Human autonomy by enabling users to make informed, autonomous decisions. This should include implementing safeguards against manipulation, over-reliance, and confusion between AI and human interactions. Regarding the subdimension of Human oversight, CVViZ offered a skill match slider to filter suitable candidates and enabled HR professionals to view each instance [193]. While it was unclear whether direct modifications to AI outputs were possible to prevent inappropriate outcomes, the tool did offer the option to add notes to each candidate [193]. Furthermore, it remained unknown whether it allowed for the adjustment of AI parameters, such as weighting certain CV attributes.

Regarding the subdimension of Resilience to attack and security within the broader dimension of **Technical robustness and safety**, CVViZ claimed that they “take our customers data and its security very seriously. All your data is encrypted and stored in world class data centers managed by Amazon Web Services (AWS)”[189]. However, the company did not provide information on security certifications or compliance with recognised security standards. Regarding the subdimension of the Fallback plan and general safety, CVViZ mentioned its use of numerous AWS services to ensure frequent data backups and availability [189]. The Accuracy, reliability, and reproducibility of the CV Screening tool were not disclosed.

Regarding the dimension of **Privacy and data governance**, CVViZ clarified that personal data would not be sold [194]. The company stated that "The Websites and Service(s) have industry-standard security measures in place to protect against the loss, misuse, and alteration of the information under our control" [194]. Furthermore, CVViZ claimed full compliance with GDPR in their role as a data processor [189]. It did not explicitly address the protection of user-generated data during interaction with the tool. Information was missing concerning the measures in place to protect physical, mental and moral integrity.

Regarding the dimension of **Transparency**, CVViZ provided vague information about the tool's purpose and capabilities, falling into the subdimension of Communication to minimise over-reliance. However, the statement, "CVViZ provides training using live demos and webinars. Apart from that our support team is available by email, live chat and on the phone calls. You also get user guide within app", indicates supportive customer service, and suggests that information on the tool's limitations might be available upon request [195]. Nonetheless, no information was available on Traceability, Explainability, or Communication strategies to minimise confusion about the tool.

No information was found on **Societal and environmental wellbeing**. In the dimension of **Accountability**, CVViZ provided users with Redress opportunities, stating that their support team was accessible via email, live chat, phone calls, and an in-app user guide. However, they did not mention any information about the system for mitigating these issues. Additionally, no information was available on the subdimensions of Auditability, Minimisation and reporting of negative impacts, and Tradeoffs.

In the subdimension of Job relatedness, which falls under the broader dimension of **Formal characteristics**, the tool's evaluation appeared to be directly relevant to the job and predictive of job performance, as it analysed data contextually extracted from CVs [190]. This tool provided all candidates who submitted CVs with equal Opportunities for assessment, employing diverse criteria for evaluation. The process of matching and resume ranking was tailored to the specific needs of the organisation, including the types of candidates it typically hired, the nature of the work they performed and similar factors[190]. It is important to recognise that relying on past hiring decisions can perpetuate existing biases. Notably, detailed scientific validation and information on this aspect were lacking.

The primary question within the Reconsideration opportunity subdimension was not applicable because the CV screening tool was not designed to allow candidates direct access to their evaluation results or to provide role-specific information. Nevertheless, candidates could receive feedback via email from the organisation, depending on its policies. Offering high-potential candidates the opportunity to remain in a talent pool for future job openings that might more closely align with their profiles could serve as a complementary service. Nevertheless, the tool should be capable of reassessing talent and considering them for other relevant current roles. Information regarding the Consistency of AI evaluations was also not found.

Regarding the **Explanation** dimension, whether candidates receive procedural feedback messages (such as confirmation of a successful CV submission) depends on the website's design. Additionally, whether candidates received a response from organisations depended on their practices. The Feedback subdimension did not apply to the AI CV Screening tool as it was not intended for direct interaction with candidates. Similarly, Selection information relied on the website design and the content generated by HR. Since the CV screening tool did not interact directly with candidates, subdimensions such as Honesty, along with Interpersonal effectiveness and Two-way communication within the **Interpersonal Treatment** dimension, were also not applicable.

Regarding the subdimension that addresses the Ease of faking answers in the dimension of **Additional rules**, CVViZ did not publicly disclose whether its CV Screening tool can detect suspicious content or inconsistencies on CVs. In the dimension of **Continuous improvement**, CVViZ minimally addresses improvements in the privacy aspect, noting that they were "making continuous efforts to help our clients protect data"[194]. They also mentioned the use of analytics and cookies to improve user experience [194]. However, aspects like Continuous performance evaluation, risk assessment, auditing, and improvement for other dimensions were not discussed.

Due to limited public information, the applicable dimensions that received few stars included Diversity, non-discrimination and fairness, Human agency and oversight, Technical robustness and safety, Transparency, Societal and environmental well-being, Accountability, Additional rules, and Continuous improvement. The subdimension Consistency also received no stars. These aspects should be clarified. Additionally, subdimensions that received four stars could be discussed if they were prioritised by the organisation.

Framework evaluation

A literature review was performed to assess existing studies on CVViZ and determine if the artefact could identify issues that had been documented in the past. The review found no specific studies focused on CVViZ from these angles. One research indicated that CVViZ was more efficient and faster than the previous methods used by HRM staff (the organisational standard) for selecting candidates for specialised positions [196]. Additionally, the incident database, designed to document actual or potential damages caused by the deployment of artificial intelligence systems, reported no incidents related to CVViZ [197].

The evaluation of CVViZ revealed that the tool was practical and effective at identifying issues. As previously mentioned, certain dimensions may not be relevant for every type of AI recruitment tool, such as those that do not interact directly with candidates, making dimensions like perceived fairness in direct interactions inapplicable. Nevertheless, these dimensions are still worth considering when interactions occur through other means. For example, although interpersonal effectiveness does not directly apply to the AI component of CV screening, it remains relevant for the overall screening process and should be considered in interactions with candidates. If the tool cannot cover certain perspectives, these might be complemented through actions from the organisational side.

6.2.2 AI-powered Chatbot

Chatbot is a virtual and autonomous agent that uses AI, including deep learning and natural language processing, to conduct human-like conversations via text or voice [198, 199]. These tools find their use across various stages of recruitment, from sourcing and screening to interviewing and selection. With advancements in generative AI, chatbots such as ChatGPT, Gemini, or Claude can be used to write job descriptions or messages to engage candidates. In sourcing, organisations can deploy chatbots on their websites to approach passive candidates and access a wider talent pool, capable of engaging multiple candidates simultaneously. Moreover, chatbots can act as question-answering systems, addressing repetitive inquiries from candidates [200]. Offering personalised interactions and quick responses, can improve candidate engagement, and increase the likelihood that they will pursue opportunities with the organisation.

During the screening phase, Conversational AI technologies can simplify resume evaluation and the shortlisting of candidates. Instead of manually reviewing each application, conversational AI systems can automatically analyse resumes by extracting essential information and matching them to predefined requirements. This automated process can lessen recruiters' workload, promote consistency, and accelerate the process of identifying qualified candidates. Furthermore, chatbots can be applied for AI-powered assessments and virtual interviews. For example, a chatbot can conduct case study-based assessment to gain insights into candidates' problem-solving and decision-making skills with automated scoring systems [201, 199]. In the selection stage, chatbots can be used to schedule meetings [200]. In a one-way online interview, where candidates record their responses to predetermined questions, chatbots can guide the process. They serve as a helpful resource, allowing candidates to instantly gather information and offering support throughout the process. Chatbots provide several advantages, including 24/7 availability, which addresses challenges related to time zones. They offer rapid and convenient access to information. They can enhance operational efficiency, save time and reduce costs. When operating properly, chatbots can contribute to a positive candidate experience [202, 199].

Challenges chatbots encounter include ethical issues and technical difficulties. Ethical considerations are crucial, especially regarding the risk of bias. While some might claim that chatbots can mitigate human bias, decision-makers should be cautious about overly relying on these systems. This is because other forms of bias, such as algorithmic bias, may still exist, as illustrated in Chapter 4.2.2. Chatbots often face difficulties with complex edge cases, and ensuring they provide relevant and accurate content is vital. A failure to do so can result in user dissatisfaction, characterised by e.g., frequent inaccuracies or frustrating interaction loops. Therefore, incorporating options for human assistance is essential. Additionally, when selecting a chatbot, attention should be paid to technical compatibility. Another challenge is the limited emotional intelligence of chatbots. Although recent models have improved in understanding context, tone, and certain idiomatic expressions, current chatbots struggle to detect sarcasm and fully understand more complex idioms. Moreover, ensuring data privacy and security is critical, as emphasised in Chapter 5.

Impress.ai

For the purposes of this thesis, an examination of the impress.ai chatbot, specifically for case study-based assessments, was conducted. Impress.ai, operating under its registered name Ideatory Pte. Ltd., is a no-code, self-service platform designed to streamline and speed up different stages of the recruitment process launched in 2017. Its mission is "to revolutionize the recruitment process with AI and intelligent automation" with a vision of creating fair hiring practices for candidates and providing recruiters with the right tools to achieve this [203]. It offers a wide range of solutions for various stages of the recruitment process, including candidate sourcing, screening, assessment, evaluation, engagement, and the onboarding phase[204]. Serving more than 50 enterprise and government clients, Impress.ai has over a hundred team members and operates in five locations, with its headquarters located in Singapore [203].

Tool assessment

The evaluation, based on publicly available information, is presented in Table 6.3

No.	Dimensions	Subdimensions	Impress.AI
1	Diversity, non-discrimination and fairness, interpersonal treatment	Avoidance of bias, Propriety of questions	★★☆☆☆
2	Diversity, non-discrimination and fairness	Accessibility and universal design	★★☆☆☆
3	Diversity, non-discrimination and fairness	Stakeholder participation	★★★☆☆
4	Human agency and oversight	Fundamental rights	☆☆☆☆☆
5	Human agency and oversight	Human agency	★★★☆☆
6	Human agency and oversight	Human oversight	★★☆☆☆
7	Technical robustness and safety	Resilience to attack and security	★★★★★
8	Technical robustness and safety	Fallback plan and general safety	★★☆☆☆
9	Technical robustness and safety	Accuracy, Reliability and reproducibility	★★★☆☆
10	Privacy and data governance, additional rules	Privacy and data protection, Perceived invasion of privacy	★★★★☆

Continued on next page

No.	Dimensions	Subdimensions	Impress.AI
11	Privacy and data governance	Quality and integrity of data, Access to data	★★★★★
12	Transparency	Traceability	★★★★★
13	Transparency	Explainability	★★★★☆
14	Transparency	Communication - minimise confusion	☆☆☆☆☆
15	Transparency	Communication - minimise over-reliance	★★☆☆☆
16	Societal and environmental well-being	Sustainable and environmentally friendly AI, Social impact, Society and democracy	☆☆☆☆☆
17	Accountability	Auditability	☆☆☆☆☆
18	Accountability	Minimisation and reporting of negative impacts	☆☆☆☆☆
19	Accountability	Tradeoffs	☆☆☆☆☆
20	Accountability	Redress	☆☆☆☆☆
21	Formal characteristics	Job relatedness	★★★★☆
22	Formal characteristics	Opportunity to perform	★★★★☆
23	Formal characteristics	Reconsideration opportunity	☆☆☆☆☆
24	Formal characteristics	Consistency	★★★★☆
25	Explanation	Feedback	★★★★☆
26	Explanation	Selection information	★★★☆☆
27	Explanation	Honesty	☆☆☆☆☆
28	Interpersonal treatment	Interpersonal effectiveness	★★★★☆
29	Interpersonal treatment	Two-Way communication	★★★★★
30	Additional rules	Ease of faking answers	★★★★☆
31	All	Ongoing evaluation and improvement	★☆☆☆☆

Table 6.3: Impress.AI assessment results overview by dimensions based on [7, 56]

Beginning with the dimension of **Diversity, non-discrimination, and fairness**, specifically with the subdimension of Bias avoidance, the fairness definition was not mentioned. The described strategies for ensuring fairness included “hiding biasing info”, converting personally identifiable information to non-personally identifiable formats, and “anonymizing candidate data” [205, 206]. Notably, approaches for addressing bias in data or algorithms remained unexplored. The provider stated that its recruitment automation platform eliminated the potential for human bias by leveraging AI-powered assessments [206, 207]. However, this statement overlooks the potential for other forms of bias, as discussed in Chapter 4.2.2, indicating a gap in fulfilling this dimension. Further clarification from the provider is needed, especially when the objective is to select a fairer tool.

In the subdimension of Accessibility and universal design, explicit mention of support for disabilities was absent. However, the platform enabled candidates to "self-identify their Nationality, Disability, or other requirements during the Chatbot screening & interview"[205]. Impress.ai offered organisations that recognised diversity issues in their hiring processes the option to collect specific information from candidates during chatbot interviews to achieve diversity goals [205]. There were limitations regarding the devices and browsers compatible with the chatbots operation [208]. The user interface seemed to be intuitive and customisable [208]. Regarding Stakeholder participation, the information provided was not specific to the development of the chatbot but referred to the overall team composition. There were no mentions of legal experts or ethicists among the team members. The expertise of the team was described as "with a set of expertise ranging over Artificial Intelligence, Computer Science, Recruiting, and I/O Psychology" [203].

In the dimension of **Human agency and oversight**, impress.ai did not publicly disclose whether the AI recruitment tool underwent evaluations to identify and mitigate potential risks to fundamental rights throughout its design, development, and operational phases. Regarding the subdimension of Human agency, impress.ai claimed to "ensuring you make informed hiring decisions"[201]. However, it did not mention any safeguards to prevent user over-reliance and confusion. In terms of the subdimension of Human oversight, transparency regarding the availability of reports and audit trails was acknowledged. It was stated that the chatbot could "collect and analyze data from candidate interactions, providing you with actionable insights"[209]. Nevertheless, the option for modifications to prevent inappropriate outcomes was not addressed. Similarly, information concerning bias monitoring and correction was lacking. It was unclear whether the provider conducted periodic evaluations to address these concerns.

Regarding the subdimension of Resilience to attack and security within the broader dimension of **Technical robustness and safety**, impress.ai claimed to maintain "the highest standards of cybersecurity" and "Compliance with International Standards". They conducted "Regular Vulnerability Assessment and Penetration Testing" alongside "Vulnerability Management" [210]. In the Fallback plan and general safety subdimension, impress.ai offered backup saving and management[210].

Regarding mechanisms for switching to rule-based processing or human oversight in cases of anomalies or system failures, no specific information was found on their website. However, an option for editing existing reviews under "Rating candidate answers" was observed from the displayed figure on the website [205]. This suggested the possibility of mechanisms being available. The provider stated that their "platform learns to answer candidate queries and reach 95%+ accuracy in answering correctly" [205]. New inquiries from candidates were incorporated as new learning implying frequent model retraining. However, this accuracy claim does not specifically address case study-based chatbot assessments, leaving the accuracy of these assessments unclear. While impress.ai mentioned "reliable data" in general, no concrete information regarding Reproducibility was found [205].

The average score in the dimension of **Privacy and data governance** was high, assuming that the claims were true. Impress.ai stated, “The Owner takes appropriate security measures to prevent unauthorised access, disclosure, modification, or unauthorised destruction of the Data”. Additionally, the statement “The Owner has taken appropriate safeguards to require that Personal Data will remain protected in accordance with this Privacy Policy and applicable laws” implies the protection of subsequently generated user data. Furthermore, the chatbot enabled organisations to integrate their privacy policies, as demonstrated in the online case experience [208].

However, information on measures to avoid invasive questions, such as those that may comply with data protection laws but could pose issues under employment law, was not found. There was insufficient information to judge its protection of individuals’ physical, mental, and moral integrity. Impress.ai stated that its data privacy policy conformed to the standards set by the Singapore Personal Data Protection Act (PDPA), the European Union General Data Protection Regulation (EU GDPR), and all relevant regulations and laws relevant to their operations [210]. Their privacy statement was prepared based on the “provisions of multiple legislations, including Art. 13/14 of Regulation (EU) 2016/679 (General Data Protection Regulation)” [211].

In the context of the dimension of **Transparency**, the sub-dimension of Traceability appeared to be satisfactorily addressed, as evidenced by the provision of “complete visibility into the hiring process through comprehensive audit trails and documentation. This allows for the tracking of candidate progress, monitoring of team member actions, and review of decisions with ease” [205]. Regarding explainability, impress.ai “enables transparency” by offering insights into decision-making processes and the factors influencing the AI’s actions, “rather than using them as buzzwords“[205]. However, the comprehensibility of these insights was uncertain due to a lack of information.

The publicly available information did not specify measures for minimising confusion, such as the tool itself being implemented to proactively inform candidates and recruiters that they were communicating with an AI system. While it is theoretically feasible to integrate such a statement directly into the tool, as content customisation should be possible, in instances where this is not viable, alternative communication channels, like email, can serve to provide such clarifications beforehand. In addressing the challenge of Minimising over-reliance, impress.ai explained its purpose and capabilities, along with a general overview of limitations, in a white paper it published. This document discussed ethical considerations, bias in conversational AI systems, technical limitations and potential pitfalls, and data privacy and security concerns [209, 199]. However, these discussions were vague and broadly applicable to conversational AI rather than being specific to the services provided by impress.ai. The provider might be able to offer seminars on these topics upon request, as a means to further educate users and mitigate over-reliance.

For the dimensions of **Societal and environmental well-being**, no information could be found. This also applied to the **Accountability** dimension.

In the subdimension of Job relatedness within the **Formal characteristics** dimension, impress.ai's chatbot assessment appeared to assess content directly relevant to the job and predictive of job performance. Evidence included statements such as "Objective and data-driven Focus on the candidate's suitability for the job and hiring decisions are made based on their merit" , and "skills-driven evaluation with impress.ai's competency-based assessments. Our platform empowers you to design customised assessments that target the core skills and competencies required for each role, ensuring you identify candidates who are the perfect fit for your organisation" [205, 201]. Impress.ai supported the integration of third-party assessment tests through its assessment marketplace [201]. Multiple data sources could be combined, thereby enabling the offering of a variety of assessments to obtain a wide-ranging overview of candidates' performance[201]. However, the scientific basis of these methods was not explicitly stated but might be available upon request.

Regarding the Opportunity to perform, impress.ai claimed that removing human biases, "ensures a fair and inclusive evaluation process, giving everyone an equal opportunity to showcase their talents"[206]. Nevertheless, mechanisms to address accessibility issues, such as those related to disabilities, were not described. The subdimension of Reconsideration opportunity for candidates was unknown. Impress.ai stated that for decision makers, they provided analysis and "culumative scores of each candidate's strengths and areas for improvement"[201]. It was unknown whether these could be accessible to candidates. The subdimension of consistency appeared to be largely satisfied based on the statement, "By having a structured screening tool in place, you are ensuring that all applicants are being screened to the exact same criteria, thus ensuring a fair evaluation" [205]. However, it remained uncertain whether this procedure would produce consistent outcomes. Additionally, adherence to established fairness standards could not be assessed due to insufficient information.

Regarding the dimensions of **Explanation**, the chatbot seemed to offer timely feedback and was stated to be available 24/7 [205]. However, there was insufficient information to assess its understandability. Based on the online case experience, an explanation of its process was given, but its validity was not mentioned [208]. Since chatbot content can be customised, incorporating such information should be feasible if the assessment is valid. It was unclear whether the tool had measures to promote truthfulness and usefulness in communication with candidates, thereby ensuring the delivery of accurate and relevant information while preventing the misleading of candidates.

The dimension of **Interpersonal treatment** received high average ratings. This was supported by evidence showing that "This two-way interaction not only ensures completeness but also adds a layer of personal touch to the recruitment process, making candidates feel valued and understood" and "with natural, engaging, and human-like interactions" [212, 213]. However, no information on the aspect of inclusiveness was found.

In the subdimension of Ease of faking answers within the **Additional rules** dimension, the tool performed authenticity checks. According to impress.ai, their "platform vigilantly monitors candidates during assessments, detecting any suspicious behaviour and ensuring a fair process"[201]. However, further measures to minimise the potential for candidates to fabricate or manipulate their responses, such as consistency checks or cross-verifications in candidate responses, were not mentioned.

In the dimension of **Continuous improvement**, only the security and infrastructure monitoring aspects were mentioned. Evidence included “continuously monitor and identify unauthorised or suspicious activities within our cloud”[210]. Furthermore, impress.ai reported that their employees received ongoing education and awareness updates on cybersecurity, including phishing email simulations [210]. Their security policy was regularly reviewed “to ensure it remains effective and aligned with evolving security best practices and regulatory requirements” [210]. Impress.ai also had infrastructure monitoring to observe the use and behaviour of application components, aiding in improving performance, operation, maintenance and troubleshooting [211]. Other dimensions were not addressed concerning continual performance evaluation, risk assessment, auditing, and improvement to ensure adherence to legal and ethical standards.

This analysis was based on currently available public information and assumed the claims were true. It is recommended to verify details with the provider, particularly concerning the dimension of Diversity, non-discrimination and fairness, such as bias and disability issues. Special attention should be given to areas receiving zero stars, especially in the accountability dimension, followed by the dimension of human agency and oversight. For perceived fairness among candidates, it is important to provide opportunities for reconsideration, allowing candidates to understand their results and areas for improvement. Additionally, measures to ensure honesty—promoting truthful and useful communication—should be clarified to ensure the delivery of accurate and relevant information and prevent the misleading of candidates. While other issues remain relevant, the focus depends on which dimensions organisations aim to prioritise and to what extent they want to achieve them.

Framework Evaluation

A literature review was conducted to identify existing analyses of impress.ai, aiming to compare whether the developed artefact could detect previously published issues. This review revealed no specific examination of impress.ai, except for one paper that analysed impress.ai’s assessment using public information, which mentioned limited information about the fairness aspect [214]. At the time of their research, Raghavan et al. noted the absence of publicly available information on what they termed the validation process, which is related to the accuracy, reliability, auditability, explainability and job relatedness subdimensions [214]. They did not find statements on vendor websites addressing concerns over bias [214]. However, the description on the website at the time of this thesis’s evaluation mentioned bias, although still vaguely. This indicated that the provider was improving. It is acknowledged that the result of this analysis might change in the future. Additionally, the incident database, which aims to compile a comprehensive record of actual or potential harms caused in the real world by the implementation of artificial intelligence systems, recorded no cases involving impress.ai [197].

The evaluation of impress.ai demonstrated that the artefact was practical and capable of identifying additional potential issues. In practice, these questions can serve as a foundation, allowing for a deeper investigation into the dimensions to uncover more issues.

6.2.3 AI-powered Video Interview

Video Interviews are job interviews conducted online to select candidates during the screening process of recruitment. These interviews can be categorised into two types: one-way and two-way. One-way interviews are asynchronous, allowing candidates to record and submit their answers to predefined questions at their convenience using a camera [215]. These recordings enable decision-makers in the organisation to review them and make decisions later [215]. Two-way interviews require real-time communication between candidates and interviewers, often focusing on evaluating behavioural traits, personality, and cultural fit within the company [216, 215]. In the context of this thesis, an "AI-powered video interview" refers specifically to one-way interviews. However, AI can also be employed in two-way interviews, for example, through AI-generated notes and summaries. Asynchronous video interview models may use various data types, including verbal content (like sentence length), paraverbal elements (such as tone of voice), and visual cues (such as facial expressions) to infer characteristics related to personality and suitability [217]. The technology used includes, for example, natural language processing (NLP) speech recognition, and emotion recognition [216]. Candidates are assessed and scored or ranked by automated systems, providing human recruiters with a structured foundation for their subsequent decision-making processes [216].

AI-powered video interviews offer several advantages in the recruitment process. For candidates, they provide convenience by allowing them to choose their interview time and location independently. For recruiters, they enhance efficiency by enabling the screening of a larger number of candidates in a shorter timeframe [218]. Researchers in both psychology and computer science proposed that AI could outperform humans in identifying or predicting an applicant's personality to screen job candidates [219]. They stated that using AI methods on audio-visual data sets can provide more reliable and predictive results than human evaluators [219]. Moreover, providers of AI video interview solutions stated that these technologies help organisations cultivate a more ethnically diverse talent pool by including candidates from a wider range of educational backgrounds, including lesser-known colleges[218]. They also claimed that AI-powered video interview software eliminated various biases, including affinity, gender, school, and racial biases, which could hinder a fair recruitment process [218].

However, AI applications can perpetuate and automate existing biases because these technical systems rely on human input and data that are often derived from discriminatory social environments [216]. Algorithmic bias can manifest in several ways. For example, algorithmic hiring models may negatively evaluate brief responses from candidates with speech impairments, or inaccurately assess input from candidates with visual impairments, such as difficulties maintaining eye contact [217]. Research indicated that algorithms trained on video signals frequently resulted in significant biases against protected groups

[220, 221, 222, 217]. Consequently, visual signals, which may inadvertently disclose sensitive information like race and gender, were removed from some products due to their questionable relevance and validity in employment contexts [223, 217]. Even when hiring algorithms based on facial analysis could perform well in controlled settings, they may not reliably generalise across different conditions, making them less effective than those based on more established data types correlated with job performance [217]. Research showed that facial recognition technology showed variations in interpreting emotions across different races [224, 215].

Regarding the prediction of candidates' personalities, AI systems may not accurately assess personality traits based on elements of a candidate's video presentation, potentially leading to misjudgments [225]. For example, variables such as wearing glasses or a headscarf could negatively affect the AI's evaluation of traits like conscientiousness or neuroticism in the Big Five personality assessment [225]. Moreover, changes in the background, such as adding artwork or a bookshelf, had been shown to influence AI's perception of a candidate, enhancing perceived conscientiousness while reducing perceived neuroticism [226, 225]. These findings raise concerns about AI tools' claims of accurately classifying candidates and the potential for creating misleading associations during the hiring process [225]. Therefore, caution is advised regarding new hiring technologies that claim to assess motivation and personality traits through computer vision, even when efforts are made to reduce bias and ensure fairness [217].

Regulatory proposals are emerging to restrict the use of facial recognition for inferring emotions, mental states, or intentions in the workplace [227, 156, 217]. For example, the Artificial Intelligence Video Interview Act, enacted in Illinois in 2020, requires companies using AI for job candidate assessments to disclose the characteristics being evaluated [228, 225]. Additionally, the European Union's AI Act categorises AI-based hiring and management systems as "high risk," necessitating strict compliance measures [156, 225]. Failure to meet these requirements could result in restrictions or the withdrawal of such systems [156, 225]. This classification highlights the significant potential impact of AI on employment and the workforce [225]. In addition, considerations around data privacy remain critical, especially in the EU (GDPR).

MyInterview

For the evaluation, the vendor myInterview was selected. MyInterview specialises in video interview services that enable job candidates to showcase their personality, experience, and qualifications, allowing hiring managers to identify suitable candidates [229]. Founded in 2016 and headquartered in New York, it has current a company size of 11- 50 employees[230]. It serves over 1,500 clients worldwide [231]. The company's mission is "to put the personality back into the application process" [231], although no vision statement was found.

This evaluation focuses on the asynchronous video interview component of myInterview, where candidates respond to predefined questions via recorded videos [232]. In addition to voice-to-text transcription, the platform offers a searchable word cloud feature, enabling

HR professionals to filter candidates using keywords or industry-specific terms and directly access the segments in each video where these keywords appear [233]. Furthermore, it evaluates candidates' personalities based on the Big Five personality traits [233]. The platform also provides automated shortlisting based on specific criteria and ranks interviews using machine learning algorithms [233].

Tool assessment

Table 6.4 shows an overview of the assessment results of myInterview.

No.	Dimensions	Subdimensions	myInterview
1	Diversity, non-discrimination and fairness, interpersonal treatement	Avoidance of bias, Propriety of questions	★★★☆☆
2	Diversity, non-discrimination and fairness	Accessibility and universal design	★★★★☆
3	Diversity, non-discrimination and fairness	Stakeholder participation	★★★★☆
4	Human agency and oversight	Fundamental rights	☆☆☆☆☆
5	Human agency and oversight	Human agency	☆☆☆☆☆
6	Human agency and oversight	Human oversight	★★★☆☆
7	Technical robustness and safety	Resilience to attack and security	★★☆☆☆
8	Technical robustness and safety	Fallback plan and general safety	☆☆☆☆☆
9	Technical robustness and safety	Accuracy, Reliability and reproducibility	★★☆☆☆
10	Privacy and data governance, additional rules	Privacy and data protection, Perceived invasion of privacy	★★★★☆
11	Privacy and data governance	Quality and integrity of data, Access to data	★★★★★
12	Transparency	Traceability	★★☆☆☆
13	Transparency	Explainability	★☆☆☆☆
14	Transparency	Communication - minimise confusion	★☆☆☆☆
15	Transparency	Communication - minimise over-reliance	★★★★☆
16	Societal and environmental well-being	Sustainable and environmentally friendly AI, Social impact, Society and democracy	☆☆☆☆☆
17	Accountability	Auditability	★☆☆☆☆
18	Accountability	Minimisation and reporting of negative impacts	☆☆☆☆☆

Continued on next page

No.	Dimensions	Subdimensions	myInterview
19	Accountability	Tradeoffs	★★☆☆☆
20	Accountability	Redress	★★★★☆
21	Formal characteristics	Job relatedness	★★★☆☆
22	Formal characteristics	Opportunity to perform	★★★★★
23	Formal characteristics	Reconsideration opportunity	☆☆☆☆☆
24	Formal characteristics	Consistency	★★★★☆
25	Explanation	Feedback	☆☆☆☆☆
26	Explanation	Selection information	★★★★☆
27	Explanation	Honesty	☆☆☆☆☆
28	Interpersonal treatement	Interpersonal effectiveness	☆☆☆☆☆
29	Interpersonal treatement	Two-Way communication	★★★★★
30	Additional rules	Ease of faking answers	☆☆☆☆☆
31	All	Ongoing evaluation and improvement	★★☆☆☆

Table 6.4: MyInterview assessment results overview by dimensions based on [7, 56]

Starting with the dimension of **Diversity, non-discrimination, and fairness**, particularly the subdimension of Bias avoidance, the definition of fairness was missing. MyInterview claimed that it strived “to use the most bias free analysis techniques available to ensure fair shortlisting” [233]. However, the specifics of these techniques were not provided, making the statement vague. Relying solely on analysis techniques was insufficient, as bias could occur at multiple stages of the AI process. In its AI Policy, effective from April 3, 2024, MyInterview expanded on its approach, stating that “the Company shall actively identify and monitor for potential bias in AI tools, services, or outputs that are incorporated into its products. Company employees and AI developers shall pay special attention to detecting and preventing potential bias in AI tools and systems based on discriminatory factors such as gender, race, age, religion, or ethnicity. The Company shall conduct periodic reviews and testing of its AI models and outputs to mitigate and eliminate bias risks, including reviewing potential bias in datasets or algorithms” [234]. While this policy indicated a positive shift towards improved bias avoidance, it remained challenging to assess the extent of its current implementation.

Regarding the subdimension of Accessibility and universal design, myInterview stated that its “profiling tools are quick and easy to use” and that “Using everyday language, profiling is accessible and easy-to-understand” [233]. Additionally, it asserted that its tools were “Optimized for best in class accessibility and inclusion standards”[232]. The website offered an accessibility mode featuring various navigation adjustments (e.g. text reader), colour options (e.g. monochrome), and content customisation (e.g. font sizing) [233]. Further aspects of universal design were not mentioned. The service was accessible on both computers and mobile devices [235].

In the context of Stakeholder participation, beyond traditional roles such as software engineering and marketing, myInterview noted that its machine learning models learned from its “team of diverse psychologists across the world”[233]. However, it did not provide detailed information such as demographics and training information. The lack of concrete information could raise concerns about the validity and inclusivity of the AI model. MyInterview’s recent AI policy stated “The relevant stakeholders for the development of AI include legal, compliance, and data protection officers, to ensure that the AI tools and systems comply with all applicable laws, regulations, and ethical principles. The development team shall also seek approval from the Company’s AI point of contact before deploying any AI tools or systems to the production environment or releasing them to customers or partners” [234]. As of its effective date on April 3, 2024, it remained unclear whether these legal and ethical aspects had been considered during the development of the current version of the tool, or if adjustments were still being made.

In the dimension of **Human agency and oversight**, no information was found regarding the subdimension of Fundamental rights and Human agency. The subdimension of Human oversight was partially fulfilled, as HR professionals could review and comment on the videos collaboratively [236]. The tool also provided video transcripts and categorised candidates, such as placing them on a shortlist [233, 236]. These categories could be rearranged [236]. However, it was uncertain whether AI-generated evaluations, such as those assessing candidates’ personalities, could be altered. The company’s AI policy stated that “The training, development, and usage of AI tools, services, and models by the Company should be overseen by natural persons to ensure that they are not misused in any harmful or illegal manner and do not incorporate any unintentional bias or discrimination” [234]. This suggested a degree of awareness and concern from the company regarding these issues. Nonetheless, the use of the term “should” implies a recommendation rather than a mandate, raising questions about the policy’s actual implementation.

In examining the subdimension of Resilience to attack and security within the broader category of **Technical robustness and safety**, myInterview had not explicitly indicated that it held certifications for cybersecurity or complies with specific security standards. However, the company’s data processing addendum emphasised that the processor shall maintain industry-standard technical and organisational measures to protect personal data [234]. These measures included safeguards against unauthorised processing, accidental destruction, loss or alteration, unauthorised disclosure of personal data, and ensuring its confidentiality and integrity [234].

Moreover, myInterview’s AI policy stated that “AI tools, systems and outputs used in the development of Company’s products should be checked and validated for security vulnerabilities regularly in accordance with the Company data security policies and applicable laws” [234]. This demonstrated a commitment to addressing potential cybersecurity threats, which may involve ongoing security updates and penetration testing. Cybersecurity encompasses more than just data protection. It also includes areas such as network security, endpoint security, and application security. While the publicly available information primarily focused on data protection, other critical aspects required further clarification as they were not explicitly mentioned.

Regarding the subdimension of the Fallback plan and general safety, no specific information was available. MyInterview indicated that its professional version offered hosting and storage customisation, allowing data to be stored on the organisation's cloud infrastructure [232]. This feature could potentially enable a backup system, although explicit details were not provided. In terms of the subdimension of Accuracy, reliability and reproducibility, myInterview presented some evidence of prediction accuracy within a particular context. For example, in a video, the company claimed that its personality profile test highly correlated with the results of a real personality test conducted in a study in South Africa [237]. Furthermore, the AI policy stated that "Company shall take steps to review its AI tools and systems to ensure that they produce accurate outputs" [234]. However, there remained gaps in the available information regarding the frequency of model retraining, data quality measures, and the representativeness of the data used. For a more definitive evaluation, additional information on these aspects was necessary.

In the context of **Privacy and data governance**, myInterview claimed that its "Privacy and Security [were] 100% GDPR Compliant" [238]. Aspects such as protection of mental integrity (e.g., ensuring that the tool does not cause psychological harm) were not specified.

Regarding the subdimension of **Transparency**, specifically Traceability, the available information suggested that the AI recruitment tool likely employed documentation and logging system to ensure the traceability of its decision-making processes. In its AI policy, myInterview emphasised explainability at various stages of AI development and integration [234]. This included understanding the reasoning behind decisions, identifying the individuals involved in creating the AI tools, knowing the data used for specific decisions, and outlining the measures taken to minimise bias or inaccuracies [234]. A video demo of myInterview showcased the tool's outputs, such as candidate scoring and personality ranges [237]. However, whether the system had a documentation and logging system that enabled traceability from data input through output evaluation needed to be verified. MyInterview stated that "The development or the integration of any AI tools within Company's products, shall be done so that the Company can explain why a specific output or decision was made"[234]. Nonetheless, it was unclear whether the tool had features to explain its decisions or predictions, and if those explanations were understandable to decision-makers. This should be clarified.

Regarding the subdimension of Communication - minimise confusion, myInterview emphasised the importance of transparency in AI interactions. It stated that "When using a client facing AI-based chatbot, Company shall clearly disclose to the client in advance that it is interacting with an AI Chatbot" [234]. It was unclear whether the chatbot provided this information directly to the candidates. However, technically, it should be easy to integrate. Myinterview seemed to have this commitment, as mentioned in its AI Policy, stating that "Company will be transparent with its clients and candidates about the involvement and capabilities of AI tools within its recruitment platform" [234].

In relation to the subdimension Communication - minimise over-reliance, myInterview outlined on its website the purpose, capabilities, and limitations of AI tools. It noted, "The more diverse your data is from the start, the more balanced the algorithm. But it's important to remember that the algorithm isn't a determinate. It's a tool developed for the express purpose (there's that word again) of facilitating the hiring process. We are fueling this process with diverse data, making sure everyone benefits." [239].

However, some statements were imprecise or ambiguous. For example, it asserted that "people are quick to claim that AI is responsible for misinformation and creating biases. When, in reality, we should be looking to ourselves and the patterns we see in society. Perhaps AI would be gender or racially biased if it reflects existing societal biases supported by data. On the contrary, this can help us acknowledge societal biases we may not even be aware of. Ultimately, the software can be trained to whatever you want it to be. So, instead of getting nervous about how the AI can view others, think of it as a tool you can manipulate to analyse others in an unbiased manner" [240]. The claims that "the software can be trained to whatever you want it to be" and that it can be "manipulated to analyse others in an unbiased manner" were questionable [240]. These statements overlooked several critical limitations and challenges such as data limitations, computational constraints, algorithmic challenges, ethical and regulatory aspects. Additionally, uncertainties like the "black box" problem and the issue of achieving 100% accuracy argue against these claims. While the tool may be designed to mitigate certain biases, claiming it could analyse individuals in an "unbiased manner" is potentially overpromising, given the complexity of bias reduction in AI systems [240].

Despite these concerns, the commitment to minimising over-reliance on AI could be seen in its AI policy. MyInterview stated that "Company shall be transparent with its client and prospects about the capabilities of AI features that are incorporated into its products or services" [234]. Moreover, myInterview pledged to "communicate clearly to its clients, prospects, candidates, employees, and other third parties' information about why it uses AI in this context, where the Company use AI, and simple detail on how the AI tool or feature works" [234]. While specific training for decision-makers on the tool's purpose, capabilities, criteria, and limitations was not mentioned, it may be possible upon request.

Despite thorough research, no information was found regarding the **Societal and environmental well-being** dimension. In the dimension of **Accountability**, information was similarly lacking. MyInterview did not address auditing areas beyond the scope of DPA (Data Processing Addendum) compliance, only mentioning that "an annual audit will be conducted to assess deeply the accuracy of the system" [234]. Additionally, there was no information on minimising and reporting negative impacts. The AI policy vaguely stated that the "Company shall use AI tools and systems responsibly and ethically, avoiding any actions that could harm others, violate privacy, or facilitate malicious activities", without providing concrete measures or outcomes [234]. In terms of Redress, MyInterview offered "24/7 Built in Support" [234]. No further information regarding the mechanism to Address and Mitigate Issues was provided.

For the subdimension of Trade-offs, the AI Policy mentioned that "Company shall document key decisions made in the development, training, purchase and deployment of its AI models, products, systems, or services, to allow review and monitoring. Company shall document the important justifications and choices made through the development process and deployment of AI tools into its products" [234]. It could not be determined whether the justifications were ethically acceptable, or well-reasoned. Furthermore, there was uncertainty about the actual implementation, potential conflicts of interest, systematic assessment and prioritisation of trade-offs, and proactive integration of ethical considerations.

In relation to the subdimension of Job relatedness under the dimension of **Formal characteristics**, myInterview used the widely recognised "Big 5 Personality Mode" [241]. According to myInterview, its algorithms generated a personality insights summary for each candidate, estimating their position on the Big 5 personality traits [241]. For example, one candidate might be characterised as "outgoing" and "competitive", while another might be described as "organized" and "sensitive" [241]. MyInterview claimed that aligning personality traits with the key qualities required for a specific role streamlined the candidate identification process, making it faster and more efficient [241]. While personality traits may suggest or correlate with certain tendencies in work behaviour and preferences, there is an ongoing debate about the extent to which personality assessments should influence hiring decisions. The AI-based classification of traits can be problematic, as it may rely on training data embedded with deep-seated biases such as social biases. The AI's classification process may be biased, leading to incorrect associations and reinforcing existing stereotypes. Apart from that, the predefined interview questions were adjustable, allowing them to be tailored to be job-related [236].

Regarding the dimension of Opportunity to perform, myInterview enabled the assessment of role fit and Big Five personality traits [240]. According to its whitepaper, "Algorithms are used to intepret body language and automated transcripts" [240]. However, myInterview claimed on its website that "By concentrating only on what a candidate is saying (and not how they look when they say it), our Machine Learning ensures that automated shortlisting focuses on personality to encourage diversity in your candidate pool" [233]. This assessment of body language appeared to contradict the statement that evaluation did not consider "how they look when they say it" [233]. The whitepaper may be outdated.

Nonetheless, myInterview allowed candidates to demonstrate their abilities and qualifications through accessible, diverse assessment methods. Interview questions could be tailored to assess professionalism, basic technical skills, interest in the program, and English proficiency, as evidenced in a case study on Merit America [242]. In another case study about Ark, myInterview improved the situation by avoiding the overlooking of potential talent, especially among those with substantial experience working with young people, but lower academic achievements [243]. These examples supported the evaluation of candidates' opportunities to perform. Regarding the subdimension of the Reconsideration opportunity, no information was found about whether candidates could access their evaluation results. However, HR personnel were able to view the transcript of video interviews for review and analysis [233].

The subdimension of consistency seemed to be largely fulfilled based on a statement from myInterview, which claimed, "improved consistency by standardising the recruitment process for all participants" [244]. However, it remained uncertain whether this standardisation especially the evaluation would actually lead to consistent results. Additionally, the adherence to fairness standards could not be evaluated due to insufficient information.

Regarding the dimension of **Explanation**, specifically the subdimension of Feedback, there was no available information to assess whether timely, informative, and understandable feedback was provided to users. For the subdimension of Selection information, myInterview provided instructions and allowed candidates to conduct a test run [245]. However, information regarding the test's validity was missing. Furthermore, no information was found concerning the subdimension of Honesty.

For the dimension of **Interpersonal treatment**, specifically the subdimension of Interpersonal effectiveness, it was unclear whether the tool used inclusive, respectful language and tone in communication with candidates. In theory, the content should be customisable. Although asynchronous interviews were not designed for Two-way communication, the tool did support two-way communication to some extent. Firstly, it offered "24/7 Built in Support", presumably for technical issues [232]. Secondly, it allowed users to "proceed with live interviews from myInterview“, enabling suitable candidates to engage in two-way communication with HR [238]. For the asynchronous part, candidates can provide input and have their opinions considered during the selection process.

Regarding the subdimension of Ease of faking answers under the **Additional rules** dimension, no information was available to determine whether the tool had implemented mechanisms to minimise the potential for candidates to fabricate or manipulate their responses.

For the dimension of **Continuous improvement**, myInterview's policy stated that "The Company's employees shall conduct an internal risk assessment, following a defined procedure, before the development or use of new AI tools, services, or systems, to ensure they comply with the principles set out in this policy, applicable laws, and mitigates the potential risks" [234]. However, this policy did not comprehensively address all dimensions and lacked details regarding ongoing improvement efforts. The policy stated, "The development team will monitor and evaluate the performance and impact of the AI tools and systems on an ongoing basis and report any issues or concerns to the Company's AI point of contact promptly" [234]. Despite this commitment, the statement remained vague, providing no clear indication of the extent or frequency of such evaluations. Nonetheless, it did demonstrate a commitment to monitoring and assessment.

Assuming the information provided was accurate, myInterview achieved on average a good score in the dimensions for which information was available. Missing information was primarily identified in the dimension of Human agency and oversight, Technical robustness and safety, Societal and environmental well-being, Accountability, Explanation and Additional rules. Specific subdimensions such as Reconsideration opportunity and Interpersonal effectiveness were found to be lacking in information. Based on this evaluation, additional improvements are needed in Transparency and Continuous improvement. These aspects should be clarified with the provider, particularly concerning the dimensions prioritised by the organisation considering the use of this service.

Framework Evaluation

A literature review was conducted to identify the existing analysis of myInterview, to determine whether the developed artefact could detect issues that had been previously published. Brandner et al. also criticised the lack of transparency on the myInterview website, highlighting that although the website stated its machine learning models were trained by a team of diverse psychologists from around the world, it did not provide details about the specific training, demographic, or geographic backgrounds of these psychologists [246]. This observation aligned with the evaluation of the Stakeholder Participation subdimension. Similarly, Raghavan et al. noted a lack of information about the training data used by myInterview, such as whether it incorporated employer-specific data, was qualitatively tailored to employers without data, or relied on pre-built models [214]. The validation process was also unclear; vendor websites did not specify whether their models were validated, the methodologies used for validation, the criteria for selecting validation data, or whether validation procedures were customised for individual clients [214]. In their research, the recorded phrase found on vendor's websites addressing bis concerns was "compliance" [214].

At the time of the current evaluation, myInterview claimed a commitment to minimising bias by stating that the company "strives to use the most bias-free analysis techniques available to ensure fair shortlisting" [233]. However, specific details about these techniques were not disclosed. As previously discussed, relying solely on analysis techniques is insufficient. Drage and Mackereth's article further suggested that myInterview might be oversimplifying the complexities of bias [225]. They argued that even if the AI disregarded explicit attributes such as gender and race, bias could still manifest in other ways [225]. They also stated that tools like myInterview illustrated a commitment to uniformity, often overlooking the complexities of race and gender [225]. This approach risked neglecting the specific needs required for fair evaluations of different groups [225]. At least, myInterview's recent AI policy indicated a commitment to addressing and mitigating bias issues, also mentioning that myInterview shall review potential biases in datasets and algorithms [234].

The evaluation conducted in this thesis was based on the assumption that the publicly available statements were accurate. However, it is crucial to critically assess these statements and verify their reliability. For example, Schellmann and Wall found that myInterview assigned a candidate a high score for English proficiency, even though she spoke only in German [247]. This incident also showed that the algorithm assessed the candidate not based on the content of her answers, but rather on personality traits interpreted from her voice [247]. Such issues were also documented in the AI incident database and belong to the validity aspect, which is indirectly addressed within the job relevance and selection information dimensions of the framework [248]. The evaluation in this thesis also identified a lack of information about validity. Drage and Mackereth also challenged the assumption that AI tools assessing candidates based on the Big Five personality traits were neutral or objective, as AI tools can make questionable and irrelevant associations between candidates and their personalities [225]. Their article also implied that AI hiring tools may perpetuate existing cultural and social biases by prioritising candidates who fit predefined norms. This process reinforced the idea that the "best fit" was someone who closely resembled the current workforce, potentially limiting diversity and innovation [225].

Certain behaviours could be more comprehensively examined through application testing rather than relying solely on publicly available information. Given that the primary objective of this evaluation was to demonstrate the practicality of the framework, only publicly accessible data were used. As previously discussed, it is advisable to revisit the evaluation or at least reassess the critical dimensions after conducting real-world tests and engaging with the vendor, to enhance the validity of the findings. The evaluation of myInterview demonstrated that the artefact was useful in identifying issues.

6.2.4 Comparison

The following section compares three AI recruitment tool vendors, each specialising in different application areas. As the specific gaps in dimensions have been analysed in the respective case studies, this section does not provide a detailed discussion of each dimension. Instead, it focuses on key findings across the broader spectrum and offers illustrative examples.

Based on the publicly available information, myInterview provided the most information, followed by Impress.AI and CVVIZ. Since the evaluation relied solely on public sources, this factor partially influenced the star ratings given to each vendor. One notable similarity among the vendors is their emphasis on security and privacy, which received considerable attention and high scores. This focus could be driven by regulatory factors such as GDPR, as non-compliance could result in severe consequences.

However, other relevant dimensions should also be addressed. There was a noticeable lack of information regarding societal and environmental well-being among the vendors, despite the importance of environmental, social, and governance (ESG) principles. Accountability, another critical dimension, was often not addressed. In terms of diversity, non-discrimination, and fairness, none of the vendors provided a definition of fairness. The universal design aspect was often underrepresented. Additionally, the descriptions of stakeholder participation were vague. It is possible that vendors, from a marketing perspective, chose to focus on publishing other aspects, which may explain the absence of information on these topics. Clarification is certainly needed. Regarding fairness, most vendors only referred to bias removal. However, eliminating bias does not resolve the full spectrum of fairness-related issues. The impression conveyed by the vendors seemed to be that removing bias from the hiring process would automatically lead to a more diverse workforce. In reality, this is a more complex issue which will be further explored in chapter 7.

The defined Dimensions are intertwined. Categorising certain aspects precisely within a single dimension can be challenging in some cases. However, the intention is not to restrict each aspect to a specific category, but to generate awareness and initiate further meaningful discussion. Some dimensions may become inapplicable in certain contexts; for example, in CV screening, the Two-way communication perspective is not intended for this task, unlike in interviews. Nonetheless, these perspectives can still provide inspiration, such as how to enhance candidates' experiences and improve the company's image. Therefore, all defined dimensions are retained in the final artefact.

MyInterview has published its own AI Policy, which addresses perspectives similar to those outlined by the High-Level Expert Group on Trustworthy AI. This could be a preparatory measure in anticipation of regulatory frameworks like the recently enacted EU AI Act. It can be expected that companies will need to establish their own AI policies and measures to comply with the law. This also means that the artefact should evolve through continuous improvement and may require adaption to align with up-to-date regulatory requirements. Additionally, the organisation's values and policies should also be considered.

Overall, the artefact has proven valuable for identifying potential issues and can be applied across different tools with varying focuses. The dimensions can be prioritised, using an appropriate scoring system, and certain customisations may be necessary, as previously noted. It is also recommended to conduct a reevaluation following actual tool testing to ensure accuracy.

CHAPTER 7

Discussion and Conclusion

This thesis aims to support fairness in AI-assisted recruitment, particularly in selection processes, by developing an assessment tool with key questions grounded in fairness-promoting requirements derived from literature using a Design Science Research (DSR) approach. This tool provides guidance for identifying critical aspects, ensuring transparency and supporting better decision-making in the selection of fairer AI tools for recruitment. The following section concludes the answers to each research question (RQ), followed by implications for theory and practice, limitations and future work, and recommendations.

7.1 Summary of Answers to the Research Questions

RQ1: What are the criteria for a fair recruiting process?

The definition of fairness varies depending on stakeholders' perspectives. This thesis elaborates on both the objective and subjective aspects of fairness. From the subjective perspective, it focuses on candidates' perception of fairness. A well-known field of study addressing these in the workplace is called organisational justice [57]. Various scholars in the past have defined their framework. Among those, Gilliand derived a model summarising influencing factors of applicants' perceived fairness in organisational selection systems, differentiating between procedural and distributive justice rules, to which this thesis refers [56].

Procedural rules are classified into three categories: formal characteristics of the selection system, explanations provided during the process and interpersonal treatment [56]. Formal characteristics comprise job relatedness, opportunity to perform, possibility for reconsideration and consistency of administration [56]. Explanation includes feedback, selection information and honesty [56]. Interpersonal treatment, which is also called interactional justice by some scholars, involves interpersonal effectiveness of administrator, two-way communication and propriety of questions [56]. Additional rules such as invasiveness of questions concerning privacy or ease of falsifying answers also apply[56]. Distributive justice rules are based on equity, equality and needs [56]. Additionally, other factors may influence applicants' perceived fairness, such as past application experience [56]. For the purpose of this thesis, the focus is on procedural rules.

Concerning AI recruitment, the fairness of the AI tool itself should be considered. Therefore, existing concepts addressing the responsible design and governance of AI were examined; see the answer to RQ3.

RQ2: How can bias arise in AI applications for recruitment?

Bias can occur at any stage of the AI decision-making process. Human bias may already exist during the initial problem definition and requirements gathering, even before the AI system is implemented. The data collected to build the AI model might contain biases, and algorithms trained on such data can perpetuate or even amplify these biases [71]. Even if the data is unbiased, algorithms can introduce systematic errors that unfairly discriminate against certain individuals or groups while favouring others [55]. The biased outcomes of AI systems affect users' decision-making, and even if the AI-generated results are unbiased, the users' final decisions might still be biased [71]. Additionally, these biased results could be used to train future algorithms, creating a feedback loop that continues to generate biases unless actions are taken to address the issue [71]. These are also relevant for AI applications for recruitment selection.

Chapter 4.2.2 provides examples of various types of bias in data, algorithms, user interaction, and how they occur. It also provides examples of where bias can occur in different AI applications for recruitment and the problems it might cause. Furthermore, it offers examples of metrics to assess algorithmic fairness and examples of how to mitigate bias through, for example, pre-processing, in-processing, and post-processing. It is important to generate awareness, not only for the developers but also for decision-makers using such tools.

RQ3: Which concepts addressing responsible design and governance of AI exist already?

Over 63 AI guidelines were found, and the dimensions of most cited guidelines that fell under research conditions (described in Chapter 3.2) were categorised and summarised (for details, see Chapter 4.2.3). Although the description of each aspect varies among guidelines, the overall concept shows similarity. The categories include but are not limited to 1.Human rights, agency, oversight, 2.Technical robustness, safety, security, 3.Transparency, 4.Privacy, data governance, 5.Diversity, non-discrimination, fairness, 6.Societal, environmental wellbeing, 7.Accountability, responsibility. These findings show similarity with previous research such as [51, 154].

The issue is that these guidelines are mostly very abstract, leaving space for interpretation. Although some industries are mentioned, they lack deeper dives into their concrete implications. This makes it difficult for people to navigate. It is recommended to establish unified international standards as a foundational framework, upon which additional country-specific requirements can be layered. While there are existing standards covering certain dimensions, they often do not comprehensively address all aspects of AI governance. As mentioned in Chapter 5, for example, ISO/IEC 42001:2023 for AI management systems mentions dimensions like transparency and human oversight. To facilitate effective implementation, regulations can directly recommend officially recognised standards, simplifying the compliance process. Consolidating individual standards into a broader framework can create a more cohesive set of guidelines that promote trustworthy AI. Industry-specific guidelines should be derived. This also shows the need and importance of the contribution of this thesis, by defining requirements for AI applications in recruitment. Guidelines on the organisational level should be defined as well which might have to consider each specific value and vision, etc.

The guidelines on trustworthy AI from HLEG were not only the most cited but also the most comprehensive ones with dimensions closely intertwined. As the goal of this thesis is also to focus on the EU region, this was chosen. Together with the concepts of perceived fairness from Gilliland, these build the dimensions of the artefact. For more reasoning and comparison, see Chapter 6.1. Chapter 5 details the derived requirements and introduces 31 key questions, with each question linked to a specific dimension.

RQ4: How suitable is the created artefact consisting of fairness-enhancing requirements for the evaluation task of identifying critical aspects?

The artefact was able to identify dimension-related issues in accordance with the given issues and identified additional issues across three cases (CV screening - CVViz, Chatbot - impress.ai, Video interview - myInterview), based on publicly available information. For details see Chapter 6.2. Just give some examples of evaluation: One notable similarity among the vendors is their emphasis on security and privacy, which received considerable attention and high scores. This focus could be driven by regulatory factors such as GDPR, as non-compliance could result in severe consequences. Furthermore, it is important to have regulations to foster ethical behaviours. Despite the significance of environmental, social, and governance (ESG) principles, vendors lacked public information regarding societal and environmental well-being.

The outlined dimensions of the artefact are interrelated, which could make it challenging to assign specific aspects to a single category. The primary goal is to promote awareness and encourage further meaningful discussions. Certain dimensions might not be applicable in all contexts—for instance, the Two-way Communication perspective doesn't apply to tasks like CV screening as it does to interviews. Nevertheless, these perspectives can still serve as inspirations e.g., in improving the overall candidate experience and enhancing the organisation's image. The dimensions can be prioritised using an appropriate scoring system. Customisations may be required as discussed in Chapter 6.2.4. It is also advisable to perform a reevaluation after actual tool testing to ensure accuracy.

7.2 Implications for Theory and Practice

This thesis contributed to both theory and practice. Starting with the theoretical perspectives: First, it adapts generic guidelines of Trustworthy AI to the AI recruitment domain promoting fairness. As discussed in Chapter 6.1, no particular framework outlines the criteria for promoting fairness in AI recruitment tools comprehensively. Most research on fairness in AI recruitment has centred on addressing issues like discrimination and bias, as well as algorithmic fairness (e.g., [179, 5]). Fairness is a complex topic and is deeply connected with other dimensions. The artefact also includes perspectives on perceived fairness in recruitment. Several papers such as [55] also mention the importance of perceived fairness in the area of AI recruitment.

Second, it generates an overview and awareness about AI recruitment: e.g., where AI is currently applied in recruitment, how bias could arise, and AI guidelines for a more responsible design and governance. Over 63 AI guidelines were found and the dimensions of the most cited guidelines were categorised and summarised. These dimensions show similarity to Jobin et al.'s research, which analysed 84 global principles and guidelines concerning ethical AI and to Hagendorff's work, which mapped 21 major AI ethics guidelines into categories [51, 154].

Third, it evaluates AI recruitment tools through case studies and identifies issues that should be addressed to promote fairness. Using the publicly available information, it identified deficiencies in dimensions and was able to uncover more issues than the previously recognised (details and comparison with existing issues see Chapter 6.2. Beyond that, it also finds that overall, the tool providers seemed to have or rather create a simplified and nearly idealistic image in AI recruitment such as ensuring unbiased hiring by hiding biasing personally identifiable information, focusing on automated personality assessment to encourage diversity, and eliminating bias from the hiring process leads to a more diverse workforce [191, 206, 233]. These could be seen as marketing advertising perspectives, but in reality, this is a more complex topic. Bias may arise at several stages not only in the data but also in algorithms and through user interaction. AI tools can make questionable and irrelevant associations between candidates and their personalities.

Similar findings were also found by Drage and Mackereth: They mentioned the point that by evaluating candidates based on predefined cultural and behavioural norms derived from past data, AI systems could reinforce existing biases, making hiring decisions that favour candidates who resemble the current workforce [225]. This perpetuates a cycle where the "best fit" is defined by how well a candidate aligns with historical patterns, rather than truly fostering diversity or innovation within the organisation [225]. Therefore, it is recommended to maintain a critical mindset, and the outcomes of this thesis promote this mindset.

From the practical perspective: The developed artefact provides practical guidance for selecting AI-assisted recruitment tools, addressing a critical gap in the literature concerning fairness in AI recruitment and selection processes [6]. It holds significant relevance not only for technical roles such as data scientists and software engineers but also for recruiters and decision-makers who intend to select and use these tools. The knowledge gap makes it challenging for stakeholders to effectively evaluate AI recruitment tools, particularly concerning fairness criteria. By identifying key dimensions that incorporate ethical and legal considerations such as references to the EU AI ACT, the artefact assists in preventing potential problems and associated costs. This approach enhances trust in the software and enables a more transparent selection process for a fairer tool, ultimately promoting fairness in recruitment and selection. It also demonstrates the evaluation of these criteria in practice via case studies. Even for the tool providers, it is very useful to reflect on and improve their products according to the dimensions.

7.3 Limitations and Future Work

In the following, the limitations and future work of this research are discussed. First, regarding the evaluation: the evaluation was based on the assumption that the claims of the AI recruitment tool providers were accurate: e.g., if they said that they are 100% GDPR compliant, then this would hold. In reality, to obtain a more valid image, one can ask e.g., proof of audits or standards and certifications related to GDPR. As the evaluation was based on the public information available, there might be cases that the dimensions achieved lower scoring because the information was just not public, which does not mean that they do not have measures. The evaluation is a snapshot of a stage. It is recommended to redo the evaluation after testing the tool. Issues should be clarified with providers.

The evaluation aims to show the practicability and that it could identify potential issues. Future research could evaluate more tools and compare them with others for the same use case and observe deeper after real testing of the tools. This could include critically examining the claims and products of AI recruitment tool vendors. Despite aims and claims of neutrality, AI tools can still perpetuate discrimination. The evaluation system could be improved as well. As each organisation has its values, vision, strategic goals and governance policies, the prioritisation of the dimensions and its final evaluation would have to consider that. For example, one could choose to use the arithmetic average score or weighted score to get the end evaluation of the whole tool to compare with other tools for the same use case. It would be helpful to define the minimum degree per dimension to reach. This also depends on several factors such as evolving regulations, corresponding standards and region. This is beyond the scope of this thesis but could be improved in the future by giving more suggestions and examples.

Second, usability plays an important role in real-world applications. However, the goal of this thesis is to define the requirements and key questions that promote fairness in tool selection. The subsequent step could involve enhancing its usability. To achieve this, a study allowing recruiters and organisational decision-makers to test the artefact would be beneficial. For better usability, this could be developed into an online questionnaire or possibly a chatbot with LLM and integrate these contents to support people's decision-making.

Third, the defined dimensions are based on the current knowledge base and should also evolve with time. Each country has its applicable laws and regulations, which are also subject to change, such as the recently published EU AI ACT. Although the artefact already covers certain aspects of the EU AI ACT due to the intersection of HLEG's Guidelines with it, further adaptation of the dimensions and concretisation of key questions could be necessary. These adaptions can be customised based on several factors, such as the use case, laws and regulations, and the organisation's values and policies. In the current artefact, only the main questions have been created as examples. However, these should not be used as standalone. They are intended to initiate discussion, serve as a starting point, and help decision-makers shortlist solutions. Developing more in-depth questions might aid further evaluation. However, usability should also be considered. Too many or overly complex questions could result in users either not using the tool or using it inappropriately.

Finally, promoting fairness in AI recruitment is a complex challenge. This thesis aims to raise awareness and assist organisational decision-makers in selecting tools that better promote fairness. However, audits and experts with specialised knowledge in this field may still be required for guidance. It is recommended to develop industry standards and certifications that vendors can adhere to, to ensure a baseline and foster trust.

7.4 Recommendations

Developing and using AI in hiring practices requires a more comprehensive approach. For example, addressing broader systemic inequalities rather than only focusing on correcting individual instances of bias [225]. Industry practitioners should not assume their AI systems are inherently neutral or harmless. They might reflect the biases of the people who make them or the society they are used to. Even if AI developers remove obvious markers like race or gender, the system could still pick up on proxy variables (such as education history or word choices) that correlate with those categories and perpetuate discriminatory patterns. Therefore, AI developers and practitioners should critically examine the underlying assumptions in their systems. Beyond ensuring technical fairness, it is recommended to consider the broader societal impact of their technology, such as on marginalised and underserved groups [249]. Other dimensions mentioned in the artefact should be considered as well. The EU AI Act emphasises that responsibility spans the entire AI value chain, including providers, deployers, importers, and distributors. Beyond that, each stakeholder can contribute to promoting ethical AI.

HR professionals need to critically think about how AI affects power dynamics in their field [225]. When selecting an AI recruitment tool, HR professionals and decision-makers should critically evaluate claims and demand transparency from vendors. Apart from the dimensions mentioned in the artefact, they should not assume that AI will automatically solve hiring challenges or diversity and inclusion issues. Drage and Mackereth warned that "diversity tools" might mask deeper, structural problems within organisations [225]. These underlying issues, such as systemic barriers to representation and inclusive culture, could remain unaddressed if organisations rely too heavily on AI tools without tackling the root causes of underrepresentation [225]. By becoming more aware of the strengths and risks of AI, HR professionals can make more informed decisions about using these tools. If they use them, they also need suitable oversight monitoring mechanisms for assessing the system's performance and impact.

While numerous principles and guidelines address AI ethics, they are often general and abstract. These frameworks should be tailored to specific industries, each of which has unique characteristics. For example, in the context of fairness in recruitment selection, a key factor to consider is the job relevance of the assessment criteria. Often, the existing guidelines leave significant room for interpretation. In the context of AI recruitment tools, specific regulations are scarce. For instance, current legislation, such as the Artificial Intelligence Video Interview Act, addresses only certain aspects of AI use in hiring processes [225]. More comprehensive efforts might be needed, particularly from AI ethicists, regulators, and policymakers [225].

Many guidelines lack consequences. Mechanisms to enforce compliance with critical values and principles are required, such as the recently enacted EU AI Act. Apart from the need for the technical development of measures for better fulfilment of each dimension of the artefact, it is recommended to foster a culture of third-party audits for AI applications as well as standards and certifications. The regulations can list recognised standards that are compliant, making actual implementation easier and also providing codes of practice. Guidelines and regulations need to be continuously updated to reflect the evolving environment (e.g., knowledge, techniques, feedback) with authorities's oversight. Incentivising relevant research and generating public awareness are also important [103].

Last but not least, AI regulation can open new markets while promoting public trust and ethical standards. However, it should strike a careful balance. Strong safeguards can prevent abuse and build confidence, but overly strict rules risk, for example, slowing innovation, discouraging investment, and causing a loss of top talent, which could affect global competitiveness in the AI sector.

Overview of Generative AI Tools Used

The following AI tools were used to correct grammar, spelling, and punctuation. Suggestions regarding better word choice, sentence structure, and translation were taken into consideration when I deemed them useful.

- Grammarly
- ChatGPT

Übersicht verwendeter Hilfsmittel

Die folgenden KI-Tools wurden verwendet, um Grammatik, Rechtschreibung und Zeichensetzung zu korrigieren. Vorschläge bezüglich einer besseren Wortwahl, des Satzbaus sowie der Übersetzung wurden berücksichtigt, wenn ich sie als sinnvoll erachtet habe.

- Grammarly
- ChatGPT

List of Figures

List of Tables

Bibliography

[1] Ideal. "AI For Recruiting: A Definitive Guide For HR Professionals." ideal.com. Accessed: July 20, 2021. [Online]. Available: https://ideal.com/ai-recruiting/

[2] E. T. Albert, "Ai in talent acquisition: a review of ai-applications used in recruitment and selection," *Strategic HR Review*, 2019.

[3] J. Dastin. "Insight - Amazon scraps secret AI recruiting tool that showed bias against women." reuters.com. Accessed: July 20, 2021. [Online]. Available: https://www.reuters.com/article/uk-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUKKCN1MK08K?edition-redirect=uk

[4] L. Yarger, F. C. Payton, and B. Neupane, "Algorithmic equity in the hiring of underrepresented it job candidates," *Online Information Review*, 2019.

[5] D. F. Mujtaba and N. R. Mahapatra, "Ethical considerations in ai-based recruitment," in *2019 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2019, pp. 1–7.

[6] S. Nugent *et al.*, "Recruitment ai has a disability problem: Questions employers should be asking to ensure fairness in recruitment," 2020.

[7] European Commission: Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[8] OECD. "Recommendation of the Council on Artificial Intelligence." legalinstruments.oecd.org. Accessed: April 4, 2023. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[9] Beijing Academy of Artificial Intelligence. "Beijing AI Principles." Ai-ethics-and-governance.institute. Accessed: April 4, 2023. [Online]. Available: https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/

[10] M. Hagenbuchner, "The black box problem of ai in oncology," in *Journal of Physics: Conference Series*, vol. 1662, no. 1. IOP Publishing, 2020, p. 012012.

[11] M. Beer, B. A. Spector, P. R. Lawrence, D. Q. Mills, and R. E. Walton, *Managing human assets.* Simon and Schuster, 1984.

[12] P. Boxall and J. Purcell, *Strategy and human resource management.* Macmillan International Higher Education, 2011.

[13] J. Bratton and J. Gold, *Human resource management: theory and practice.* Palgrave, 2017.

[14] M. Armstrong and S. Taylor, *Armstrong's handbook of human resource management practice.* Kogan Page Publishers, 2020.

[15] S. Köszegi, *Human Resource Management and Leadership(VO).* TU Wien, 2016.

[16] H.-S. Shih, L.-C. Huang, and H.-J. Shyur, "Recruitment and selection processes through an effective gdss," *Computers & Mathematics with Applications*, vol. 50, no. 10-12, pp. 1543–1558, 2005.

[17] S. D. Rozario, S. Venkatraman, and A. Abbas, "Challenges in recruitment and selection process: An empirical study," *Challenges*, vol. 10, no. 2, p. 35, 2019.

[18] S. Newell, "Recruitment and selection," *Managing human resources: Personnel management in transition*, pp. 115–147, 2005.

[19] A. E. Barber, *Recruiting employees: Individual and organizational perspectives.* Sage Publications, 1998.

[20] P. Kuryło, A. Idzikowski, J. Cyganiuk, and R. Paduchowicz, "Recruitment, selection and adaptation of staff in enterprise," *System Safety: Human-Technical Facility-Environment*, vol. 1, no. 1, 2019.

[21] B. R. Dineen and S. M. Soltis, "Recruitment: A review of research and emerging directions," *APA Handbook of Industrial and Organizational Psychology*, 2011.

[22] J. A. Breaugh and M. Starke, "Research on employee recruitment: So many studies, so many remaining questions," *Journal of management*, vol. 26, no. 3, pp. 405–434, 2000.

[23] J. R. Mueller and B. Baum, "The definitive guide to hiring right," *Journal of Applied Business and Economics*, vol. 12, no. 3, pp. 140–153, 2011.

[24] T. Thebe and G. Van der Waldt, "A recruitment and selection process model: The case of the department of justice and constitutional development," 2014.

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

[26] A. B. Holm, "E-recruitment: towards an ubiquitous recruitment process and candidate relationship management," *German Journal of Human Resource Management*, vol. 26, no. 3, pp. 241–259, 2012.

[27] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," *AI magazine*, vol. 27, no. 4, pp. 12–12, 2006.

[28] S. E. Dilsizian and E. L. Siegel, "Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment," *Current cardiology reports*, vol. 16, no. 1, p. 441, 2014.

[29] C. Stryker and E. Kavlakoglu. "Artificial Intelligence (AI)." Ibm.com. Accessed: Jan. 9, 2021. [Online]. Available: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

[30] D. Yingying, "Optimization of labor value and relationship distribution under the blockchain," in *2020 International Conference on Data Processing Techniques and Applications for Cyber-Physical Systems*. Springer, 2021, pp. 1131–1136.

[31] A. Kayid, "The role of artificial intelligence in future technology," 2020.

[32] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers." *IEEE Access*, 2024.

[33] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7327–7347, 2021.

[34] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[35] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Manchester, 2000, pp. 29–39.

[36] European Commission: Directorate-General for Communications Networks, Content and Technology, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020.

[37] Stanford University and McKinsey&Company. "Artificial intelligence (AI) adoption worldwide 2022, by industry and function." statista.com. Accessed: September 10, 2024. [Online]. Available: https://www.statista.com/statistics/1112982/ai-adoption-worldwide-industry-function/

[38] D. Collier and C. Zhang, "Can we reduce bias in the recruiting process and diversify pools of candidates by using different types of words in job descriptions?" 2016.

[39] Y. K. Dwivedi *et al.*, "Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, p. 101994, 2021.

[40] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *International conference on machine learning*. PMLR, 2018, pp. 1115–1124.

[41] M. Broussard, *Artificial unintelligence: How computers misunderstand the world*. MIT Press, 2018.

[42] S. U. Noble, "Algorithms of oppression: How search engines reinforce racism," in *Algorithms of oppression*. New York university press, 2018.

[43] C. Maloney, "Weapons of math destruction: How big data increases inequality and threatens democracy," *Journal of Markets & Morality*, vol. 20, no. 1, pp. 194–197, 2017.

[44] J. Iivari and J. Venable, "Action research and design science research - seemingly similar but decisively dissimilar," *17th European Conference on Information Systems, ECIS 2009*, pp. 1642–1653, 01 2009.

[45] A. R. Hevner, "A three cycle view of design science research," *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4, 2007.

[46] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision support systems*, vol. 15, no. 4, pp. 251–266, 1995.

[47] P. Johannesson and E. Perjons, *An introduction to design science*. Springer, 2014.

[48] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[49] P. Palvia, D. Leary, E. Mao, V. Midha, P. Pinjani, and A. F. Salam, "Research methodologies in mis: an update," *The Communications of the Association for Information Systems*, vol. 14, no. 1, p. 58, 2004.

[50] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," Technical report, Ver. 2.3 EBSE Technical Report. EBSE, Tech. Rep., 2007.

[51] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[52] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS quarterly*, pp. 75–105, 2004.

[53] S. Baskarada, “Qualitative case study guidelines,” *Baškarada, S.(2014). Qualitative case studies guidelines. The Qualitative Report*, vol. 19, no. 40, pp. 1–25, 2014.

[54] P. Baxter and S. Jack, “Qualitative case study methodology: Study design and implementation for novice researchers,” *The qualitative report*, vol. 13, no. 4, pp. 544–559, 2008.

[55] A. Köchling and M. C. Wehner, “Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development,” *Business Research*, vol. 13, no. 3, pp. 795–848, 2020.

[56] S. W. Gilliland, “The perceived fairness of selection systems: An organizational justice perspective,” *Academy of management review*, vol. 18, no. 4, pp. 694–734, 1993.

[57] J. Greenberg, “Organizational justice: Yesterday, today, and tomorrow,” *Journal of management*, vol. 16, no. 2, pp. 399–432, 1990.

[58] S. E. Wolfe, J. Rojek, V. M. Manjarrez Jr., and A. Rojek, “Why does organizational justice matter? uncertainty management among law enforcement officers,” *Journal of Criminal Justice*, vol. 54, pp. 20–29, 2018.

[59] J. S. Adams, “Inequity in social exchange,” ser. Advances in Experimental Social Psychology, L. Berkowitz, Ed. Academic Press, 1965, vol. 2, pp. 267–299.

[60] J. Thibaut and L. Walker, “Procedural justice: A psychological analysis.” 1975.

[61] R. G. Folger and R. Cropanzano, *Organizational justice and human resource management.* Sage, 1998, vol. 7.

[62] G. S. Leventhal, “What should be done with equity theory?” in *Social exchange.* Springer, 1980, pp. 27–55.

[63] R. J. Bies and J. F. Moag, “Interactional justice: Communication criteria of fairness,” *Research on negotiation in organizations*, vol. 1, pp. 43–55, 1986.

[64] D. S. Conner, “Socially appraising justice: A cross-cultural perspective,” *Social Justice Research*, vol. 16, no. 1, pp. 29–39, 2003.

[65] M. Chan, “Organizational justice theories and landmark cases,” *the international journal of organizational analysis*, 2000.

[66] M. K. Lee, “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management,” *Big Data & Society*, vol. 5, no. 1, p. 2053951718756684, 2018.

[67] A. L. Hunkenschroer and C. Luetge, “Ethics of ai-enabled recruiting and selection: A review and research agenda,” *Journal of Business Ethics*, vol. 178, no. 4, pp. 977–1007, 2022.

[68] M. Langer, C. J. König, and M. Papathanasiou, “Highly automated job interviews: Acceptance under the influence of stakes,” *International Journal of Selection and Assessment*, vol. 27, no. 3, pp. 217–234, 2019.

[69] C. Kaibel, I. Koch-Bayram, T. Biemann, and M. Mühlenbock, “Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderators,” in *Academy of Management Proceedings*, vol. 2019, no. 1. Academy of Management Briarcliff Manor, NY 10510, 2019, p. 18172.

[70] J. Ochmann and S. Laumer, “Fairness as a determinant of ai adoption in recruiting: An interview-based study,” 2019.

[71] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[72] C. A. Readler, “Local government anti-discrimination laws: Do they make a difference,” *U. Mich. JL Reform*, vol. 31, p. 777, 1997.

[73] M. Mercat-Bruns, D. B. Oppenheimer, and C. Sartorius, *Comparative Perspectives on the Enforcement and Effectiveness of Antidiscrimination Law: Challenges and Innovative Tools.* Springer, 2018, vol. 28.

[74] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado, “Bias and discrimination in ai: a cross-disciplinary perspective,” *IEEE Technology and Society Magazine*, vol. 40, no. 2, pp. 72–80, 2021.

[75] European Commission. “Non-discrimination.” ec.europa.eu. Accessed: April 20, 2022. [Online]. Available: https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en

[76] U.S.Department of Labor. “Ethnic/National Origin.” dol.gov. Accessed: May 7, 2022. [Online]. Available: https://www.dol.gov/general/topic/discrimination/ethnicdisc#:~:text=Title%20VII%20of%20the%20Civil,religion%2C%20sex%20or%20national%20origin.

[77] S. Barocas and A. D. Selbst, “Big data’s disparate impact,” *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[78] EU General Data Protection Regulation (GDPR). “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data

and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." eur-lex.europa.eu. Accessed: May 7, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

[79] P. Hacker, E. Wiedemann, and M. Zehlike, "Towards a flexible framework for algorithmic fairness," *arXiv preprint arXiv:2010.07848*, 2020.

[80] A. Pena, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal ai: Testbed for fair automatic recruitment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 28–29.

[81] European Commission. "AI Act enters into force." commission.europa.eu. Accessed: August 1, 2024. [Online]. Available: https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

[82] M. Vasconcelos, C. Cardonha, and B. Gonçalves, "Modeling epistemological principles for bias mitigation in ai systems: an illustration in hiring decisions," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 323–329.

[83] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–9.

[84] M. Cummings, "Automation bias in intelligent time critical decision support systems," in *AIAA 1st intelligent systems technical conference*, 2004, p. 6313.

[85] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on information systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[86] S. Akter, G. McCarthy, S. Sajib, K. Michael, Y. K. Dwivedi, J. D'Ambra, and K. N. Shen, "Algorithmic bias in data-driven innovation in the age of ai," p. 102387, 2021.

[87] D. Pessach and E. Shmueli, "Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings," *Expert Systems with Applications*, vol. 185, p. 115667, 2021.

[88] Equal Employment Opportunity Commission, "Uniform guidelines on employee selection procedures," U.S. Department of Labor, Tech. Rep., 1978.

[89] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[90] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[91] K. Makhlouf, S. Zhioua, and C. Palamidessi, “On the applicability of machine learning fairness notions,” *ACM SIGKDD Explorations Newsletter*, vol. 23, no. 1, pp. 14–23, 2021.

[92] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *arXiv preprint arXiv:1609.05807*, 2016.

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

[94] L. H. Nazer *et al.*, “Bias in artificial intelligence algorithms and recommendations for mitigation,” *PLOS Digital Health*, vol. 2, no. 6, p. e0000278, 2023.

[95] K. Orphanou, J. Otterbacher, S. Kleanthous, K. Batsuren, F. Giunchiglia, V. Bogina, A. S. Tal, A. Hartman, and T. Kuflik, “Mitigating bias in algorithmic systems—a fish-eye view,” *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[96] I. Žliobaitė and B. Custers, “Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models,” *Artificial Intelligence and Law*, vol. 24, pp. 183–201, 2016.

[97] R. Fu, Y. Huang, and P. V. Singh, “Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications,” in *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*. INFORMS, 2020, pp. 39–63.

[98] House of Lords et al., “Ai in the uk: ready, willing and able?” *Retrieved August*, vol. 13, p. 2021, 2018.

[99] IEEE, “Ethically aligned design - a vision for prioritizing human well-being with autonomous and intelligent systems,” *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2019.

[100] Future of Life Institute. “Asilomar AI Principles.” FutureofLife.org. Accessed: Mar. 9, 2024. [Online]. Available: https://futureoflife.org/open-letter/ai-principles/

[101] Google. “AI at Google: Our Principles.” Google.com. Accessed: Mar. 9, 2024. [Online]. Available: https://blog.google/technology/ai/ai-principles/

[102] Université de Montréal. “Montréal Declaration for a Responsible Development of Artificial Intelligence.” DeclarationMontreal-IAResponsable.com. Accessed: Mar. 9, 2024. [Online]. Available: https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl_IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

[103] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, “Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations,” *Minds and machines*, vol. 28, pp. 689–707, 2018.

[104] B. Shneiderman, “Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[105] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. Fuso Nerini, “The role of artificial intelligence in achieving the sustainable development goals,” *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[106] S. Lo Piano, “Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward,” *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–7, 2020.

[107] E. Neri, F. Coppola, V. Miele, C. Bibbolino, and R. Grassi, “Artificial intelligence: Who is responsible for the diagnosis?” pp. 517–521, 2020.

[108] S. Thiebes, S. Lins, and A. Sunyaev, “Trustworthy artificial intelligence,” *Electronic Markets*, vol. 31, no. 2, pp. 447–464, 2021.

[109] M. Kuziemski and G. Misuraca, “Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings,” *Telecommunications policy*, vol. 44, no. 6, p. 101976, 2020.

[110] B. Mittelstadt, “Principles alone cannot guarantee ethical ai,” *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.

[111] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information fusion*, vol. 58, pp. 82–115, 2020.

[112] B. Mittelstadt, “Ai ethics–too principled to fail,” *arXiv preprint arXiv:1906.06668*, 2019.

[113] K. Siau and W. Wang, “Artificial intelligence (ai) ethics: ethics of ai and ethical ai,” *Journal of Database Management (JDM)*, vol. 31, no. 2, pp. 74–87, 2020.

[114] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices,” *Science and engineering ethics*, vol. 26, no. 4, pp. 2141–2168, 2020.

[115] S. Umbrello and I. Van de Poel, “Mapping value sensitive design onto ai for social good principles,” *AI and Ethics*, vol. 1, no. 3, pp. 283–296, 2021.

[116] P. Nemitz, “Constitutional democracy and technology in the age of artificial intelligence,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180089, 2018.

[117] J. M. Puaschunder, J. Mantl, and B. Plank, “Medicine of the future: The power of artificial intelligence (ai) and big data in healthcare,” *Available at SSRN 3607616*, 2020.

[118] O. J. Erdélyi and J. Goldsmith, “Regulating artificial intelligence: Proposal for a global solution,” *Government Information Quarterly*, p. 101748, 2022.

[119] M. Sutrop, “Should we trust artificial intelligence?” *Trames: A Journal of the Humanities and Social Sciences*, vol. 23, no. 4, pp. 499–522, 2019.

[120] L. Floridi, “Translating principles into practices of digital ethics: Five risks of being unethical,” in *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, 2021, pp. 81–90.

[121] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong *et al.*, “Toward trustworthy ai development: mechanisms for supporting verifiable claims,” *arXiv preprint arXiv:2004.07213*, 2020.

[122] I. Munoko, H. L. Brown-Liburd, and M. Vasarhelyi, “The ethical implications of using artificial intelligence in auditing,” *Journal of Business Ethics*, vol. 167, no. 2, pp. 209–234, 2020.

[123] N. A. Smuha, “The eu approach to ethics guidelines for trustworthy artificial intelligence,” *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019.

[124] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi, “The ethics of algorithms: key problems and solutions,” *AI & SOCIETY*, vol. 37, no. 1, pp. 215–230, 2022.

[125] S. Gerke, T. Minssen, and G. Cohen, “Ethical and legal challenges of artificial intelligence-driven healthcare,” in *Artificial intelligence in healthcare*. Elsevier, 2020, pp. 295–336.

[126] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, “Co-designing checklists to understand organizational challenges and opportunities around fairness in ai,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[127] H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux, “Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns,” *Big Data & Society*, vol. 6, no. 1, p. 2053951719860542, 2019.

[128] F. J. Zuiderveen Borgesius, “Strengthening legal protection against discrimination by algorithms and artificial intelligence,” *The International Journal of Human Rights*, vol. 24, no. 10, pp. 1572–1593, 2020.

[129] S. Wachter, B. Mittelstadt, and C. Russell, "Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai," *Computer Law & Security Review*, vol. 41, p. 105567, 2021.

[130] U. Gasser and C. Schmitt, "The role of professional norms in the governance of artificial intelligence," in *The oxford handbook of ethics of AI*. Oxford University Press Oxford, 2020, p. 141.

[131] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: the design and governance of ethical ai and autonomous systems [scanning the issue]," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 509–517, 2019.

[132] B. Shneiderman, "Human-centered artificial intelligence: three fresh ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109–124, 2020.

[133] D. Schiff, J. Biddle, J. Borenstein, and K. Laas, "What's next for ai ethics, policy, and governance? a global overview," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 153–158.

[134] A. Winfield, "Ethical standards in robotics and ai," *Nature Electronics*, vol. 2, no. 2, pp. 46–48, 2019.

[135] B. W. Wirtz, J. C. Weyerer, and B. J. Sturm, "The dark sides of artificial intelligence: An integrated ai governance framework for public administration," *International Journal of Public Administration*, vol. 43, no. 9, pp. 818–829, 2020.

[136] Y. Zeng, E. Lu, and C. Huangfu, "Linking artificial intelligence principles," *arXiv preprint arXiv:1812.04814*, 2018.

[137] M. Taddeo and L. Floridi, "How ai can be a force for good," *Science*, vol. 361, no. 6404, pp. 751–752, 2018.

[138] S. Mohamed, M.-T. Png, and W. Isaac, "Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence," *Philosophy & Technology*, vol. 33, no. 4, pp. 659–684, 2020.

[139] B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial intelligence and the public sector—applications and challenges," *International Journal of Public Administration*, vol. 42, no. 7, pp. 596–615, 2019.

[140] W. Xu, "Toward human-centered ai: a perspective from human-computer interaction," *interactions*, vol. 26, no. 4, pp. 42–46, 2019.

[141] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building ethics into artificial intelligence," *arXiv preprint arXiv:1812.02953*, 2018.

[142] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, "Artificial intelligence & human rights: Opportunities & risks," *Berkman Klein Center Research Publication*, no. 2018-6, 2018.

[143] Y. Wang, M. Xiong, and H. Olya, "Toward an understanding of responsible artificial intelligence practices," in *Proceedings of the 53rd hawaii international conference on system sciences.* Hawaii International Conference on System Sciences (HICSS), 2020, pp. 4962–4971.

[144] A. F. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, 2018.

[145] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and machines*, vol. 30, no. 3, pp. 411–437, 2020.

[146] B. C. Stahl and D. Wright, "Ethics and privacy in ai and big data: Implementing responsible research and innovation," *IEEE Security & Privacy*, vol. 16, no. 3, pp. 26–33, 2018.

[147] L. Delponte and G. Tamburrini, *European Artificial Intelligence (AI) leadership, the path for an integrated vision.* European Parliament, 2018.

[148] F. McKelvey and M. MacDonald, "Artificial intelligence policy innovations at the canadian federal government," *Canadian Journal of Communication*, 2019.

[149] K. H. Keskinbora, "Medical ethics considerations on artificial intelligence," *Journal of clinical neuroscience*, vol. 64, pp. 277–282, 2019.

[150] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi, "The chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation," *AI & society*, vol. 36, no. 1, pp. 59–77, 2021.

[151] National New Generation Artificial Intelligence Governance Expert Committee. "Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence." DigiChina.Stanford.edu. Translation by Graham Webster and Lorand Laskai. Accessed: Mar. 9, 2024. [Online]. Available: https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/

[152] Executive Office of the President, National Science and Technology Council Committee on Technology. "Preparing for the Future of Artificial Intelligence." Whitehouse.gov. Accessed: Mar. 9, 2024. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

[153] R. T. Vought. "Guidance for Regulation of Artificial Intelligence Applications." Whitehouse.gov. Accessed: Mar. 9, 2024. [Online]. Available: https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/11/M-21-06.pdf

[154] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and machines*, vol. 30, no. 1, pp. 99–120, 2020.

[155] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable," in *International workshop on explainable, transparent autonomous agents and multi-agent systems*. Springer, 2021, pp. 119–138.

[156] European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act)," 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[157] S. Koffas, J. Xu, M. Conti, and S. Picek, "Can you hear it? backdoor attacks via ultrasonic triggers," in *Proceedings of the 2022 ACM workshop on wireless security and machine learning*, 2022, pp. 57–62.

[158] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 253–261.

[159] C. Chen, Z. Liu, W. Jiang, S. Q. Goh, and K.-Y. Lam, "Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations," *arXiv preprint arXiv:2408.12935*, 2024.

[160] J. Gu, "A survey on responsible generative ai: What to generate and what not," *arXiv preprint arXiv:2404.05783*, 2024.

[161] European Commission. "The EU Cybersecurity Act." ec.europa.eu. Accessed: April 18, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-act

[162] B. Wolford. "Data Protection Impact Assessment (DPIA)." gdpr.eu. Accessed: April 18, 2024. [Online]. Available: https://gdpr.eu/data-protection-impact-assessment-template/

[163] ISO. "ISO/IEC JTC 1/SC 42 Artificial Intelligence." iso.org. Accessed: April 18, 2024. [Online]. Available: https://www.iso.org/committee/6794475.html

[164] IEEE. "The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems." ieee.org. Accessed: October 6, 2024. [Online]. Available: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[165] J. Smither and K. Pearlman, "Perceptions of the job-relatedness of selection procedures among college recruits and recruiting/employment managers," in *RR Reilly (Chair), Perceived validity of selection procedures: Implications for organizations. Symposium conducted at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO*, 1991.

[166] A. Anastasi, *Psychological testing*, 6th ed. New York:: Macmillan, 1988.

[167] J. Greenberg, "Determinants of perceived fairness of performance evaluations." *Journal of applied psychology*, vol. 71, no. 2, p. 340, 1986.

[168] B. H. Sheppard and R. J. Lewicki, "Toward general principles of managerial fairness," *Social justice research*, vol. 1, pp. 161–176, 1987.

[169] R. D. Arvey and P. R. Sackett, "Fairness in selection: Current developments and perspectives," in *Personnel selection.* Jossey-Bass, 1993.

[170] T. R. Tyler and R. J. Bies, "Beyond formal procedures: The interpersonal context of procedural justice," in *Applied social psychology and organizational settings.* Psychology Press, 1990, pp. 77–98.

[171] European Union. "Equal Treatment with Nationals." Europa.eu. Accessed: Nov. 23, 2024. [Online]. Available: https://europa.eu/youreurope/citizens/work/work-abroad/equal-treatment-with-nationals/index_en.htm

[172] J. W. Lounsbury, W. Bobrow, and J. B. Jensen, "Attitudes toward employment testing: Scale development, correlates, and" known-group" validation." *Professional Psychology: Research and Practice*, vol. 20, no. 5, p. 340, 1989.

[173] A. Huffcutt, "Intelligence is not a panacea in personnel selection," *The Industrial-Organizational Psychologist*, vol. 27, no. 3, pp. 66–67, 1990.

[174] B. RJ, "Interactional justice: Communication criteria of fairness," *Research on negotiation in organizations*, vol. 1, pp. 43–55, 1986.

[175] O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders, "Truthful ai: Developing and governing ai that does not lie," *arXiv preprint arXiv:2110.06674*, 2021.

[176] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, and X. Huang, "Building guardrails for large language models," *arXiv preprint arXiv:2402.01822*, 2024.

[177] S. G. Ayyamperumal and L. Ge, "Current state of llm risks and ai guardrails," *arXiv preprint arXiv:2406.12934*, 2024.

[178] E. F. Stone and D. L. Stone, "Privacy in organizations: Theoretical issues, research findings, and protection mechanisms," *Research in personnel and human resources management*, vol. 8, no. 3, pp. 349–411, 1990.

[179] N. Tilmes, “Disability, fairness, and algorithmic bias in ai recruitment,” *Ethics and Information Technology*, vol. 24, no. 2, p. 21, 2022.

[180] J. Yu, Z. Ma, and L. Zhu, “The configurational effects of artificial intelligence-based hiring decisions on applicants’ justice perception and organisational commitment,” *Information Technology & People*, 2023.

[181] X. Zhang, Y. Zhao, X. Tang, H. Zhu, and H. Xiong, “Developing fairness rules for talent intelligence management system,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[182] E. Albaroudi, T. Mansouri, and A. Alameer, “A comprehensive review of ai techniques for addressing algorithmic bias in job hiring,” *AI*, vol. 5, no. 1, pp. 383–404, 2024.

[183] J. Dastin. Insight - amazon scraps secret ai recruiting tool that showed bias against women. [Online]. Available: https://www.reuters.com/article/idUSKCN1MK0AG/

[184] A. A. Kodiyan, “An overview of ethical issues in using ai systems in hiring with a case study of amazon’s ai based hiring tool,” *Researchgate Preprint*, pp. 1–19, 2019.

[185] A. Persson, “Implicit bias in predictive data profiling within recruitments,” *Privacy and Identity Management. Facing up to Next Steps: 11th IFIP WG 9.2, 9.5, 9.6/11.7, 11.4, 11.6/SIG 9.2. 2 International Summer School, Karlstad, Sweden, August 21-26, 2016, Revised Selected Papers 11*, pp. 212–230, 2016.

[186] CVViZ. About us. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://in.linkedin.com/company/cvviz

[187] ——. Hire top candidates faster with ai recruiting software. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/

[188] ——. About cvviz. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/career/

[189] ——. Gdpr. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/gdpr/

[190] ——. Resume screening using ai. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/product/resume-screening/

[191] ——. 3 advantages of using ai for resume screening. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/blog/ai-for-resume-screening/

[192] ——. Resume screening: How to guide on effective screening. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/resume-screening-guide/

[193] Passivern. Cvviz- ats review - the fastest applicant tracking system (ats) | passivern. youtube.com. Accessed: April 15, 2024. [Online]. Available: https://www.youtube.com/watch?v=gIKBxTP3jiA

[194] CVViZ. Privacy policy. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/privacy-policy/

[195] ——. Choose your plan for ai powered ats. cvviz.com. Accessed: April 15, 2024. [Online]. Available: https://cvviz.com/pricing/

[196] K. D. Strang and Z. Sun, "Erp staff versus ai recruitment with employment real-time big data," *Discover Artificial Intelligence*, vol. 2, no. 1, p. 21, 2022.

[197] A. incident database. Ai incident database. incidentdatabase.ai. Accessed: April 3, 2024. [Online]. Available: https://incidentdatabase.ai/

[198] O. Allal-Chérif, A. Y. Aránega, and R. C. Sánchez, "Intelligent recruitment: How to identify, select, and retain talents from around the world using artificial intelligence," *Technological Forecasting and Social Change*, vol. 169, p. 120822, 2021.

[199] impress.ai, "Conversational ai in recruitment automation," Tech. Rep. [Online]. Available: https://impress.ai/whitepapers/conversational-ai-in-recruitment-automation/

[200] B. Barghi, E. Gallardo-Gallardo, and V. Fernandez, "An overview of chatbots usage in recruitment and selection practices," 2022.

[201] impress.ai. Candidate assessments & evaluation. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/candidate-assessments-evaluation/

[202] V. Taecharungroj, ""what can chatgpt do?" analyzing early reactions to the innovative ai chatbot on twitter," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 35, 2023.

[203] impress.ai. About us. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/about-us/

[204] ——. Candidate sourcing. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/solutions/candidate-sourcing/

[205] ——. Technology. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/technology/

[206] ——. Diversity, equity, and inclusion. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/diversity-equity-and-inclusion/

[207] ——. Professional hiring. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/solutions/professional-hiring/

[208] ——. Practice online case experience. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/html-widget/chat-widget/5290cce9-2748-4823-aebd-0809804e24b6/

[209] ——. Conversational virtual assistant. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/conversational-virtual-assistant/

[210] ——. Security policy. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/security-policy/

[211] ——. Privacy policy. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/privacy-policy/

[212] ——. How are ai-powered chatbots changing the landscape of resume scoring and candidate shortlisting? impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/blogs/how-are-ai-powered-chatbots-changing-the-landscape-of-resume-scoring-and-candidate-shortlisting/

[213] ——. Conversational ai. impress.ai. Accessed: April 2, 2024. [Online]. Available: https://impress.ai/conversational-al/

[214] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[215] B. Kammerer, “Hired by a robot: The legal implications of artificial intelligence video interviews and advocating for greater protection of job applicants,” *Iowa L. Rev.*, vol. 107, p. 817, 2021.

[216] L. T. Brandner, P. Mahlow, A. Wilken, A. Wölke, H. Harmouch, and S. D. Hirsbrunner, “How data quality determines ai fairness: The case of automated interviewing.”

[217] F. Alessandro, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Z. Borgesius, and A. J. Biega, “Fairness and bias in algorithmic hiring: a multidisciplinary survey,” *arXiv preprint arXiv:2309.13933*, 2023.

[218] C. Pattapu. How ai interviewing is redefining the way we hire. talview.com. Accessed: April 18, 2024. [Online]. Available: https://blog.talview.com/en/how-ai-interviewing-redefining-the-way-we-hire

[219] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, “Tensorflow-based automatic personality recognition used in asynchronous video interviews,” *IEEE Access*, vol. 7, pp. 61 018–61 023, 2019.

[220] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D’Mello, “Bias and fairness in multimodal machine learning: A case study of automated video interviews,” in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 268–277.

[221] L. Hemamou, A. Guillon, J.-C. Martin, and C. Clavel, “Don’t judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[222] S. Yan, D. Huang, and M. Soleymani, “Mitigating biases in multimodal personality assessment,” in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 361–369.

[223] R. Maurer. Hirevue discontinues facial analysis screening. shrm.org. Accessed: April 15, 2024. [Online]. Available: https://www.shrm.org/topics-tools/news/talent-acquisition/hirevue-discontinues-facial-analysis-screening

[224] L. Rhue, “Racial influence on automated perceptions of emotions,” *Available at SSRN 3281765*, 2018.

[225] E. Drage and K. Mackereth, “Does ai debias recruitment? race, gender, and ai’s “eradication of difference”,” *Philosophy & technology*, vol. 35, no. 4, p. 89, 2022.

[226] J. Fergus. A bookshelf in your job screening video makes you more hirable to ai. [Online]. Available: https://www.inverse.com/input/culture/a-bookshelf-in-your-job-screening-video-makes-you-more-hirable-to-ai

[227] European Commission, “Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts,” 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[228] R. Heilweil, “Illinois says you should know if ai is grading your online job interviews,” *Vox*. [Online]. Available: https://www.vox.com/recode/2020/1/1/21043000/artificial-intelligence-job-applications-illinios-video-interivew-act

[229] myInterview. About. myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://www.myinterview.com/about

[230] ——. About us. linkedin.com. Accessed: April 18, 2024. [Online]. Available: https://www.linkedin.com/company/myinterview

[231] ——. “Ready for your next adventure?” Myinterview.pinpointhq.com. Accessed: April 18, 2024. [Online]. Available: https://myinterview.pinpointhq.com/

[232] ——. Start video interviewing today. myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://www.myinterview.com/pricing

[233] ——. Myinterview intelligence. myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://www.myinterview.com/product-intelligence

[234] ——. Myinterview privacy policy. myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://www.myinterview.com/privacy

[235] R. Media. myinterview demos video interview software. youtube.com. Accessed: April 18, 2024. [Online]. Available: https://www.youtube.com/watch?v=Y_ps82lr4WA

[236] myInterview. myinterview - remote interviewing. youtube.com. Accessed: April 18, 2024. [Online]. Available: https://www.youtube.com/watch?v=03g-smH7vxU

[237] SSR. myinterview demo. youtube.com. Accessed: April 18, 2024. [Online]. Available: https://www.youtube.com/watch?v=OR1n1qm6Ouk

[238] myInterview. The myinterview experience. myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://www.myinterview.com/product-features

[239] ——. Purpose-built machine learning: Smart shortlisting fueled by diverse data. blog.myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://blog.myinterview.com/purpose-built-machine-learning-automated-shortlisting-fueled-by-diverse-data

[240] ——, "The definitive guide to ai for human resources," Tech. Rep. [Online]. Available: https://embed.myinterview.com/landing/media/myinterview+-+AI+for+HR+Whitepaper.pdf?_gl=1*1ffvpkj*_gcl_au*NTA1NDIyNDU0LjE3MTA3MTA4ODQ.

[241] ——. myinterview intelligence tm: How ai is paving the way for faster and more effective hiring. hubspot.net. Accessed: April 18, 2024. [Online]. Available: https://cdn2.hubspot.net/hubfs/2074952/myInterview%20-%20Intelligent%20candidate%20interviewing%20-%20Dec2019.pdf?utm_campaign=White%20Paper&utm_medium=email&_hsmi=81035347&_hsenc=p2ANqtz--b4NsKUE-QdJF_wZBJNR6LXGO-99e2qbceEsaKgrnljAxwxJM8WVh1hzlO0g2WrKHETtThqxDCwv5Vne-Db1mMm9qTsg
utm_content=81035347&utm_source=hs_automation

[242] ——. Case study myinterview merit america. hubspotusercontent-na1.net. Accessed: April 18, 2024. [Online]. Available: https://2074952.fs1.hubspotusercontent-na1.net/hubfs/2074952/myInterview%20-%20Merit%20America%20Case%20Study.pdf

[243] ——. Case study myinterview ark. hubspotusercontent-na1.net. Accessed: April 18, 2024. [Online]. Available: https://2074952.fs1.hubspotusercontent-na1.net/hubfs/2074952/myInterview%20-%20ARK%20Case%20Study.pdf

[244] ——. Redefining our recruitment process through implementing one-way video interviewing - ocado group case study. hubspotusercontent00.net. Accessed: April 18, 2024. [Online]. Available: https://f.hubspotusercontent00.net/hubfs/2074952/Ocado%20Group%20Case%20Study-%20%20Redefining%20The%20Recruitment%20Process%20with%20Video%20Interviewing.pdf

[245] ——. Setting up your interview. support.myinterview.com. Accessed: April 18, 2024. [Online]. Available: https://support.myinterview.com/en/articles/4350830-setting-up-your-interview

[246] L. T. Brandner, P. Mahlow, A. Wilken, A. Wölke, H. Harmouch, and S. D. Hirsbrunner, “How data quality determines ai fairness: The case of automated interviewing.” in *EWAF*, 2023.

[247] S. Wall and H. Schellmann, “We tested ai interview tools. here’s what we found,” 2021, 2024-04-18. [Online]. Available: https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

[248] AI incident database. Incident 344: Hiring algorithms provided invalid positive results for interview responses in german. incidentdatabase.ai. Accessed: April 18, 2024. [Online]. Available: https://incidentdatabase.ai/cite/344/

[249] P. Kalluri, “Don’t ask if artificial intelligence is good or fair, ask how it shifts power,” *Nature*, vol. 583, no. 7815, pp. 169–169, 2020.