UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Fair AI

Mujahid Pasha Mohsin Pasha

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

MENTOR :

Mevrouw Elisavet KOUTSOVITI-KOUMERI

2022

2023

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Fair AI

Mujahid Pasha Mohsin Pasha

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Koenraad VANHOOF

MENTOR :

Mevrouw Elisavet KOUTSOVITI-KOUMERI

Acknowledgements

In the culmination of this thesis, I stand on the threshold of completing a remarkable journey in pursuit of a master's degree in management, specializing in Business Process Management at Hasselt University, Belgium. The university entrusted me with the captivating research topic of "Navigating Fairness Definitions in AI", a subject that has ignited my intellectual curiosity and fueled my dedication. Through the journey of crafting this dissertation, I aim to shed light on the vital intricacies of this research area and contribute to its evolving body of knowledge.

At this juncture, I wish to extend my deepest gratitude to Prof. Dr. Koen Vanhoof, my unwavering supervisor, whose profound guidance and unwavering support have been the cornerstone of this successful endeavor. Moreover, I am profoundly indebted to Miss Lisa Koutsoviti Koumeri, my thesis mentor, whose steadfast encouragement and unwavering support have been a beacon of inspiration throughout this research odyssey. Her insightful comments, invaluable suggestions, and the wealth of knowledge she shared have been instrumental in shaping the trajectory of this work.

In this moment of reflection, I am compelled to acknowledge the unwavering support of my cherished family and friends. Their belief in my potential has been a driving force that fueled my determination.

Finally, I reserve my most heartfelt appreciation for my parents, whose unwavering assistance, encouragement, and endorsement have carried me through both triumphant and challenging moments. Your unwavering love and support are the bedrock upon which my journey stands, and for that, I am forever grateful.

Thank you!

Mujahid Pasha Mohsin Pasha

Abstract

Our study delves into fair AI design for employee management within organizations, addressing biases in decision-making amid rapid AI integration. We identify challenges posed by algorithmic bias, emphasizing potential consequences for employees, customers, and overall outcomes. Analyzing industry-accepted fairness definitions, we emphasize diverse dimensions of fairness. Employing a systematic review approach, we highlight a lack of uniformity in fairness definitions, urging a coherent design agenda. Our insights enable inclusive decision-making, interdisciplinary collaboration, and alignment with legal and ethical frameworks. While limitations exist, our research empowers organizations to embrace responsible AI practices, ensuring equitable outcomes amid AI's transformative potential.

Keywords: fair AI design, algorithmic bias, decision-making, workplace scenarios, organizational outcomes, fairness definitions, systematic review, interdisciplinary collaboration, legal and ethical alignment, responsible AI practices.

Contents

1. Introduction

In the current world we are seeing an extensive use of Artificial Intelligence (AI) in almost every field. AI is driving innovation and lot of businesses are relying on decision support systems for making important decisions based on previously collected data and running them on probabilistic algorithms which provide predictions for unseen data under some uncertainty (Feuerriegel et al., 2020).

AI is taking critical decisions that humans usually made in the past to determine if a person gets a loan or not (Verma & Rubin, 2018), whether someone gets hired for a job through an AI hiring tool (Chhabra et al., 2021), or someone's college application getting accepted (Chhabra et al., 2021).

Since these tools are dependent on the historic data that is being used to predict, they are not error free or free from bias, as the dataset used could be prone to some kind of bias and systematic unfairness whereby individuals or whole groups are treated disparately (Feuerriegel et al., 2020).

AI bias pertains to the structured and inequitable inclinations displayed by an artificial intelligence system during its decision-making, resulting in outcomes that unfairly advantage particular individuals or groups (Pi, 2021).

Due to enormous boom in research in the field of AI many technologies and uses are arising for these tools like algorithmic moderation to solve new problems arising due to the technology driven society, one such example being the use of algorithmic moderation being used to perform hash matching and prediction by governments and firms to classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown). However, this could create more problems as the moderation tools are prone to bias for example, hate speech classifiers designed to detect violations of a platform's guidelines could be disproportionally flagging language used by a certain social group, thus making that group's expression more likely to be removed (Gorwa et al., 2020).

The use of AI will continue but there is a need to look at the consequence of using these algorithms that is bias and discrimination against individuals or entire groups (Feuerriegel et al., 2020).

This is where the absence of fairness is to be noted in all the algorithms being used to solve different problems, and research in the past has shown the lack of fairness in the models. Fair AI is a probabilistic decision support system that guards against disparate harm (or benefit) to various groupings and fair AI could also be regression-based, decision rule-based, and distance-based etc., (Feuerriegel et al., 2020). The topic of algorithmic fairness is gaining so much attention that it is being added in the upcoming EU AI act which will make it mandatory for business using AI to make sure that the algorithms are fair (Madiega., 2019). It becomes important for managers who are going to deal with AI solutions in their businesses for certain product and service offerings to comply with the upcoming regulations. Legislative bodies around the world are putting into effect laws that prohibit differential treatment in algorithmic decision-

making. For example, the US Fair Lending Act penalizes algorithmic biases in risk scoring, while the General Data Protection Regulation in the EU enforces accountability for AI (GDPR).

Hence it is important to choose the correct fairness definition for making the algorithm fair. Figuring out the right fairness definition is important. Different definitions have been put forward that explain fairness in AI mathematically. They can be grouped into notions of fairness, group fairness and individual fairness (Feuerriegel et al., 2020). But choosing the right fairness definitions depends on the result we are trying to achieve and eliminating bias, where bias is characterized as a systematic deviation from the true value of an estimated parameter (Feuerriegel et al., 2020). To ensure that a classifier is fair, one has to decide the notion of fairness one wants to adopt, and this will help in identifying which notion of fairness has been used in solving which type of software discrimination and make the decision making simpler (Verma & Rubin, 2018). (Verma & Rubin, 2018) show the most prominent definitions of fairness for solving algorithmic classification problems and also demonstrate how certain cases can be considered fair according to some definitions and unfair according to others, there are still no literature reviews which cover industry-specific cases where fair treatment is brought about. In this literature review, we try to address this by compiling a list of cases where fairness was brought about in different industries and reviewing which fairness definitions were considered to do so.

Several researchers have proposed mathematical definitions of fairness to bring about fairness in the algorithm, however, it is unclear for industries using these algorithms to choose the right definition of fairness as fairness is not a purely technical construct, it has social, political, philosophical and legal facets (Foulds et al., 2020). Hence there is a need for an interdisciplinary analysis of fairness in AI and its relationship to society, civil rights, and social goals which are to be achieved by using the mathematical definitions of fairness. For instance, Foulds et al., in their paper propose intersectional AI fairness criteria that perform a comprehensive, interdisciplinary analysis of their relation to the concerns of diverse fields which include humanities, law, privacy, economics, and statistical machine Learning, however their analysis and proposal are motivated by intersectionality wherein they take civil rights and feminism, together simultaneously. However, it is still not clear which definition of fairness, that industries can use to bring about fairness. (Feuerriegel et al., 2020) in their paper suggest several challenges and opportunities for information system research, among them, one is the perception of Fair AI by people, it highlights how some sensitive attributes are considered easily like race whereas attributes like Christian are vaguely defined, and when moving into domain-specific cases it becomes vaguer.

This lack of perception of what is fair AI leads to a problem that industries face to determine the right fairness definition to choose from to make their algorithm fair. As a result, there is a lack of industry-specific literature reviews that ensure fair treatment, there are not many publications available that highlight this shortage, although research has been previously done to address fairness which is grouped by domain-specific and fairness definitions (Mehrabi et al., 2021), similarly (Verma & Rubin, 2018) show the most prominent definitions of fairness for solving algorithmic classification problems and also demonstrate how certain cases can be considered fair according to some definitions and unfair according to others, there are still no literature reviews which cover industry-specific cases where fair treatment is brought about. In this literature review,

we try to address this by compiling a list of cases where fairness was brought about in different industries and reviewing which fairness definitions were considered to do so.

This paper tries to answer the following research questions:

RQ1: What are the definitions that are accepted by the industry for expressing fairness in AI?
RQ2: How do definitions change for different applications & purposes?
RQ3: Is it possible to have a framework of definitions of AI Fairness based on industry usage?

Through a systematic literature review, this study examines seemingly divergent perspectives of definitions of fairness. In particular, research that addresses the issue of selecting the appropriate fairness definitions for diverse instances of bias and discrimination in various industries is examined in-depth in this article's extensive review of research on definitions of fairness. The article's foundation is a study of fairness mitigation studies that were published in scholarly journals between 2012 and 2022. These articles cover a wide range of topics, including finance, business, social science, engineering, and medical research. The study examines several dimensions of fairness for group fairness, individual fairness, and the fairness mitigation achieved via them by analyzing the gathered publications. The pursuit of these research questions is fueled by the recognition of the existing gap between theoretical fairness definitions and their practical implementation in real-world business scenarios. While the field of fairness in clustering has seen significant interest, there remains limited evidence of its actual application and efficacy in practical business contexts. Thus, this thesis seeks to bridge the gap between theoretical understandings and real-world implications of fairness in AI, offering practical insights that can drive positive change in various industries. In the realm of this study, the literature review is currently limited in scope due to the nascent nature of the field. The scarcity of publications in this domain has resulted in a concise review of existing literature. It is important to note that the relative scarcity of available resources in this field might impact the breadth and depth of the literature review, reflecting the early stages of research and exploration in this area.

2. Research method

This study conducted a systematic literature analysis to find articles that clarify fairness definitions have improved fairness in a variety of fields, including business, social science, engineering, and medical research. The thesis does not include debiasing in the literature review. This methodology

consists of five steps: selecting a review topic, locating the literature, assessing and synthesizing the literature, writing the review, and creating a list of pertinent developing references. The fairness definitions of AI used in various businesses will be used as a starting point for the literature search.

2.1. Search strategy

The review identified relevant articles which enabled a transparent, documented research process with criteria for including and excluding articles. The systematic review involved the following steps: state research questions, develop guidelines for collecting literature, decide on inclusion and exclusion criteria, develop a comprehensive search plan for finding literature, developing a narrative review and describing literature, and synthesize the literature. The present study explores the various ways in which fairness was brought about and how the selection of definition varies between different fields has been defined in the literature in order to determine whether the perspectives differ in their definitions. The main search strategy identified research articles that defined the concept of fairness in AI. In order to capture this, inclusion and exclusion criteria were developed. The initial inclusion criteria were broad to ensure that all relevant articles were identified, were peer-reviewed empirical or conceptual articles, were published in English, and were publicly accessible and that the main focus including papers that mentioned industry or business where fairness definitions were applied. Papers that purely explore theoretical notions of fairness were excluded. Existing literature reviews on fairness definitions were also assessed in the scope of this thesis. The inclusion criteria for the review also involved selecting papers that discuss fairness definitions applied in industry or business cases. Papers concentrating solely on theoretical aspects of fairness are excluded from consideration within this scope. Moreover, the assessment of existing literature reviews within this context involves analyzing the number of business cases they incorporate. Due to the shortage of published articles especially when searching for articles related to "Fairness definitions in AI in Industry" certain arXiv papers were also included. A database search of google scholar was conducted to find articles that contained the following terms in their abstract, title, or keywords: "Fairness Definitions in AI" and "Definitions of fairness in AI in industry" and "Group Fairness in AI in industry" and "Fairness in AI surveys". This research employs a narrative approach that amalgamates methodologies from prior scholarly endeavors to actively foster transparency, impartiality, and ethical contemplation during the implementation of advanced technologies. The initial phase involves identifying pertinent articles. To ensure comprehensive coverage, searches encompass databases like Scopus, IEEE Xplore, Web of Science, ACM, ABI/INFORM, EBSCO, JSTOR, ScienceDirect, and the University of Hasselt Library. The study primarily considers papers from 2015 onwards due to the increasing prominence of fairness concepts in recent years. This approach excludes pre-2015 publications as more recent papers integrate numerous findings from their predecessors. Scholarly works from ICIS, CHI 2020 proceedings, and unreviewed arXiv papers are also integrated due to the expansive nature of fairness. Considering the complexity and diversity of fairness, the research includes a database search of google scholar, was conducted to find articles that contained the following terms in their

abstract, title, or keywords: “Fairness Definitions in AI” and “Definitions of fairness in AI in industry” and “Group Fairness in AI in industry” and “Fairness in AI surveys”. The scope of the search was not limited to any particular field, subject of research, or journal so that a full overview of AI fairness research could be obtained. This results in a dataset of 70 articles. The next step involves screening articles' titles, keywords, and abstracts to exclude those not conceptually or contextually relevant. Relevance is determined by the article's primary focus on the keywords related to fairness and fairness definitions mentioned earlier. The dataset is further refined by excluding articles lacking substantial insights, resulting in 45 remaining articles for analysis. Technical papers centered on mathematical solutions for bias detection and mitigation in AI models are also excluded, given their computational focus falls outside this study's scope. The final dataset comprises 30 relevant papers. An in-depth analysis of these 31 papers reveals common themes, even though wording may vary.The scope of the search was not limited to any particular field, subject of research, or journal so that a full overview of AI fairness research could be obtained.

3. Literature Review

3.1 Concepts in AI Fairness

Algorithmic fairness has garnered significant attention from researchers in AI, Software Engineering, and Law circles, leading to the emergence of numerous fairness definitions in recent times. The absence of unanimous agreement on the suitable definition for different contexts is a prevailing challenge. To address this, S. Verma & J. Rubin (2018), in their work extensively explore the principal fairness definitions pertinent to the algorithmic classification problem. The paper not only clarifies the rationale underpinning each definition but also provides an illustrative case study that demonstrates their practical application. This comprehensive examination sheds light on the intricate considerations encompassing fairness in the domain of algorithmic decision-making (Verma & Rubin, 2018).

3.2 Fair AI

Fair AI refers to the ethical and principled development and implementation of artificial intelligence systems to ensure fairness and equity in their decision-making processes. It involves the use of mathematical concepts and statistical methods to quantify and monitor the level of fairness in AI-driven information systems over time. The ultimate objective of Fair AI is to reduce biases and discrimination inherent in human decision-making by designing AI algorithms to be fair and unbiased. This concept presents unprecedented opportunities for individuals, organizations, and society as a whole, offering the potential to create more inclusive and just systems.

The use of machine learning algorithms is becoming more evident in current times. Algorithms are making decisions that were usually made by expert individuals or committees, and algorithms are suggesting us recommendations on what products to buy, and whom to date (Mehrabi et al., 2021). They are being increasingly used in taking high-stake scenarios like allotment of loans and hiring decisions (Mehrabi et al., 2021). The algorithms used for making these decisions usually are probabilistic algorithms that make inferences by learning existing patterns from data and, after deployment give predictions for unseen data under some uncertainty. As a result, they are susceptible to biases and system unfairness which was existing in the data used to make these decisions and this leads to discrimination wherein individuals or whole groups are treated disparately(Feuerriegel et al., 2020). These machine learning technologies have "found dark skin unattractive", claimed that "black people re-offend more" (Caton & Haas, 2020), and these instances are just a few examples where dependence on these technologies has brought to light the prejudice that exists in the data. This can have severe consequences for individuals and groups affected by the use of this technology. It emphasizes the need for fair AI, wherein it can help quantify bias and mitigate discrimination against subgroups (Feuerriegel et al., 2020).

Businesses and organizations are inevitably going to use AI tools to innovate and the services they offer as a result of deploying these tools could expose them to substantial legal risk.

The algorithms demonstrate the discrimination in the system, and one such example is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which measured the risk of a person recommitting another crime, and an investigation into the software found that there was bias against African-American offenders (Mehrabi et al., 2021). Identifying the sources of unfairness is important in order to use the algorithms in a good way

The fairness definitions needed to address the issue of unfairness in AI require one to understand all the sources of unfairness that are to be considered while choosing the right definitions of fairness. The sources of unfairness generally are bias and discrimination.

Feuerriegel, Dolata, Schwabe, and Schwabe (2020) delve into the challenges and opportunities of Fair AI in their article. They emphasize its significance for information systems researchers and practitioners, especially as AI becomes more powerful and pervasive, raising concerns about potential biases. The authors underscore the necessity of developing tools for fair AI, anticipating that fairness in decision support systems will likely be regulated through legal initiatives in the future. Businesses and organizations without a clear strategy for achieving fair AI may face financial and reputational risks, given the potential consequences of violating fairness laws.

The concept of fairness in AI is multifaceted, and its impact on societies is substantial. Therefore, building information systems that can detect and address unfairness in an appropriate manner is crucial. Adopting mathematical notions of fairness represents a step in the right direction, enabling

a systematic and principled approach to ensuring fairness in AI-driven decision-making processes (Feuerriegel et al., 2020).

3.3 Importance of Enforcing Fairness

Enforcing fairness in AI is of utmost importance due to several compelling reasons. First and foremost, it necessitates extensive research to understand user perceptions, particularly regarding sensitive attributes. Sensitive attributes within artificial intelligence pertain to traits or personal characteristics of individuals that possess the capacity to be discriminatory or to introduce bias into decision-making processes (Brinkmann et al., 2022). This understanding is crucial to ensure that AI systems do not perpetuate discriminatory practices. Second, enforcing fairness is vital in adapting fair AI to real-world applications, as it addresses challenges related to transparency and decision support. By reconciling transparency with fairness, we can build trust in AI systems, making them more accountable and reliable.

Transparency in the context of fairness in AI pertains to the clarity and understandability of how an AI system arrives at its decisions. It involves making the decision-making process of AI algorithms more accessible and interpretable to humans. Achieving transparency ensures that the inner workings and reasoning behind AI-generated outcomes are not obscured or hidden. This concept aligns with the idea of creating machine learning models that can be comprehended by human experts, thereby enabling accountability, trust, and identification of potential biases in the system (Doshi-Velez & Kim, 2017).

Moreover, fair AI introduces a fairness-performance trade-off, wherein certain subgroups may experience reduced prediction performance. In the realm of artificial intelligence (AI), prediction performance refers to the level of accuracy exhibited by an AI system in generating accurate predictions. It indicates the degree to which the AI system's estimations correspond with real-world results. For example, let's take a medical AI designed to forecast whether a patient has a specific illness. If the AI's prognoses closely correlate with the actual medical diagnosis, it is deemed to have a high prediction performance. Conversely, if the AI frequently produces incorrect forecasts, its performance is regarded as subpar (Li et al., 2022). Addressing this trade-off is critical to ensure equitable outcomes for all users (Feuerriegel et al., 2020).

The economic implications of fairness in AI should not be overlooked, as it directly impacts management decisions and industry adoption. Fair AI tools are essential for organizations to comply with legal initiatives that enforce fairness in decision support systems. This compliance not only mitigates financial and reputational risks for businesses but also aligns with societal expectations for ethical and unbiased AI. Ultimately, fair AI presents unprecedented opportunities for organizations and society as a whole by promoting fairness by design. By employing mathematical notions of fairness, practitioners can statistically quantify fairness levels in AI systems, thereby fostering continuous monitoring and improvement over time. Most importantly, fair AI can significantly reduce biases that often plague human decision-making, making AI systems more reliable and equitable for all individuals (Feuerriegel et al., 2020).

A significant area of research in algorithmic fairness centers around individual fairness (IF)

methods, which stem from the principle of "similar treatment," aiming to treat similar individuals alike. IF employs distance metrics to assess individual similarity, and proponents argue it provides an accurate fairness definition and should be prioritized. However, this perspective is challenged by several issues. Counterexamples reveal that similarity-based treatment isn't always fair. IF's learning of similarity metrics risks encoding human bias. The reliance on prior moral judgments weakens IF's suitability as a fairness guide. Furthermore, the incongruity of moral values makes similarity metrics impractical for numerous tasks. Therefore, individual fairness might not be a comprehensive fairness definition, but rather a tool amidst others to mitigate algorithmic bias (Fleisher, 2021). Some researchers contend that fairness is not easily expressible mathematically or quantifiable due to its inherent complexity and the multitude of contextual factors involved (Fleisher, 2021).

Machine learning (ML) has been applied to critical issues with societal implications, like predicting prisoner recidivism, bank loan disbursement, job applications, and college admissions. However, ML models trained on biased data can amplify biases in high-impact applications. Examples include Microsoft's chatbot, which adopted racist language from biased tweets, and the COMPAS tool, which unfairly predicted criminal behavior based on race. Ensuring fairness in ML is crucial to address these challenges (Chhabra et al., 2021).

In the above section it is noteworthy to highlight that both the authors (Chhabra et al., 2021) and (Feuerriegel et al., 2020) emphasize the significance of enforcing fairness in AI and machine learning. They stress the need for extensive research to understand user perceptions and the impact of sensitive attributes on AI systems. They acknowledge the challenges in reconciling transparency with fairness and the trade-off between fairness and prediction performance. Both authors also recognize the economic implications of fairness in AI, highlighting the importance of compliance with legal initiatives and industry adoption. Moreover, they both highlight the potential of fair AI to reduce biases in decision-making processes, making AI systems more reliable and equitable for all individuals.

3.4 Bias

Understanding the root of bias is important when using the algorithms, since most algorithms are data-driven and require data to be trained upon, data becomes an important aspect in the functionality of these algorithms and systems. (Mehrabi et al., 2021) in their paper explain how bias could linger in the data used to train the algorithm, and as a result the algorithms also learn this bias and perpetuate them in their results, making the predictions obtained from them biased.

There are various types of biases that could lead to unfairness in different downstream learning tasks, it becomes difficult for a manager of a firm or company who is trying to address the issue of unfairness. Unfairness can come from various sources of bias like data to the algorithm, algorithm to user bias, and user to data bias, these are some broad types of biases that are addressed in research by researchers in the field of computer science (Mehrabi et al., 2021).

A manager unaware of these technical aspects of bias mentioned will have a difficult time choosing the right fairness definition, so understanding the sources of unfairness is vital.

3.5 Fairness definitions in AI

In the field of AI, researchers have come up with different mathematical definitions of fairness. These definitions are categorized into two main types: group-level fairness and individual fairness. Group-level fairness focuses on attributes like race, gender, or disability that should not lead to discrimination. It evaluates how prediction errors are distributed among different groups like protected and unprotected groups. On the other hand, individual fairness emphasizes treating similar individuals in a similar way, irrespective of their group membership. For instance, when it comes to loan applications, individual fairness means that people with similar financial attributes should receive similar treatment in terms of loan approval and interest rates (Verma & Rubin, n.d.).

The exploration of fairness definitions in AI has attracted significant attention from AI, Software Engineering, and Law communities, resulting in the proposal of more than twenty fairness notions (Verma & Rubin, n.d.). However, achieving a consensus on the appropriate definition for each situation remains a challenge. To address this issue and provide clarity, researchers have gathered and presented the most prominent fairness definitions for the algorithmic classification problem (Verma & Rubin, n.d.).

The definitions encompass the following aspects of fairness:

- Fairness through Awareness: Ensuring that similar individuals receive similar classification outcomes based on a defined distance metric (Verma & Rubin, 2018).
- Well-Calibration: Aligning predicted probabilities for both protected and unprotected groups with the true probability of belonging to the positive class (Hardt et al., 2016).
- Counterfactual Fairness: Ensuring that the predicted outcome does not depend on the descendants of the protected attribute in the causal graph (Kusner et al., 2017).
- No Unresolved Discrimination: Ensuring that the causal graph lacks paths from the protected attribute to the predicted outcome, except through a resolving variable (Kilbertus et al., 2017).
- No Proxy Discrimination: Ensuring that the causal graph lacks paths from the protected attribute to the predicted outcome that are blocked by a proxy variable (Kilbertus et al., 2017).
- Fair Inference: Classifying paths in the causal graph as legitimate or illegitimate to ensure fair decision-making (Zafar et al., 2017).
- Distributive Fairness: Distributive fairness refers to the equitable allocation of resources, benefits, or outcomes among individuals or groups. It ensures that the distribution is

based on relevant criteria without favoring one group over another. One prominent framework for distributive fairness is Rawls' theory of justice, which emphasizes fair distribution of societal goods to maximize the welfare of the least advantaged (Rawls, 1971).

- Procedural Fairness: Procedural fairness focuses on the fairness of the processes and procedures used to make decisions. It emphasizes transparency, consistency, and inclusiveness in decision-making processes. Research has shown that even when outcomes are not favorable, individuals are more likely to accept decisions if they perceive the procedures as fair (Lind & Tyler, 1988).
- Interactional Fairness: Interactional fairness, also known as interpersonal or informational fairness, pertains to the fairness of the communication and treatment individuals receive during decision-making processes. It emphasizes respectful and considerate treatment, providing explanations for decisions, and showing empathy. Interactional fairness can significantly impact individuals' perceptions of the overall fairness of an organization (Bies & Moag, 1986).

Causal Reasoning definitions are based on causal graphs, where attributes and their relationships influence the outcome. This approach includes Counterfactual Fairness, where the predicted outcome should not depend on descendants of the protected attribute in the causal graph. No Unresolved Discrimination ensures that there is no path from the protected attribute to the predicted outcome in the causal graph, except through a resolving variable. Similarly, No Proxy Discrimination ensures that there is no path blocked by a proxy variable.

These definitions offer different perspectives on fairness, with causal reasoning providing a comprehensive approach to building fair algorithms based on causal relationships between attributes and outcomes (Verma & Rubin, 2018).

Fairness through Awareness focuses on ensuring that similar individuals receive similar classification outcomes based on a defined distance metric. Well-Calibration, on the other hand, requires predicted probabilities for both protected and unprotected groups to match the true probability of belonging to the positive class.

Fair Inference classifies paths in the causal graph as legitimate or illegitimate, aiming to ensure fair decision-making based on causal relationships between attributes and outcomes.

To address this question comprehensively, the study draws upon the foundational work by Chhabra, Ė, and Mohapatra (2021) titled "An Overview of Fairness in Clustering," which provides valuable insights into fairness in the context of clustering algorithms.

Addressing the lack of consensus on fairness definitions is vital to promoting fairness by design in AI systems, which has direct implications for various fields such as economics, law, and technology

(Verma & Rubin, n.d.). As AI becomes more powerful and pervasive, fairness in decision support systems will be enforced by legal initiatives like EU AI ACT (European Commission, 2022) and Canada's AIDA (Government of Canada, 2022), making the development of fair AI tools crucial for industry adoption and compliance. European Union's AI ACT emphasizes ethical guidelines and regulatory measures for artificial intelligence, aiming to ensure responsible and transparent AI development and deployment (European Commission, 2022). On the other hand, Canada's AIDA, as proposed in the Artificial Intelligence and Data Act, seeks to establish a comprehensive framework to govern AI and data usage, fostering innovation while safeguarding privacy and ethical considerations (Government of Canada, 2022). Although (Feuerriegel et al., 2020) suggest in there paper that achieving fairness in AI can reduce biases inherent in human decision-making, leading to more equitable outcomes and societal benefits, this notion remains disputable among academics. Fairness in AI has significant implications for society, technology, and organizations. It is crucial to approach this topic holistically and scientifically (Feuerriegel et al., 2020).

3.6 Legal definitions on which the Fairness Definitions are based

Romei and Ruggieri (2014) explore discrimination, focusing on its implications within the context of human rights and anti-discrimination laws. The discussion reveals two primary approaches: formal equality, which prioritizes merit-based treatment while disregarding irrelevant attributes, and substantive equality, which aims to attain fair outcomes by accounting for individual differences. Various discrimination measures such as risk difference, risk ratio, and selection rate are highlighted, used across different countries. The authors stress the importance of data collection and statistical evidence in discrimination cases, with statistical conclusions serving as initial evidence. Prima facie evidence pertains to the evidence that, at first glance, is satisfactory for establishing a fact or assertion unless contradicted or refuted (Chellasamy et al., 2014).

3.7 Fariness Definitions in Industry

3.7.1 Human Resources (industry

In organizations that prioritize fairness, instances of unfairness can still occur, leading to a need for redress approaches to address the consequences of unfairness. Restorative justice and retributive justice are two such approaches that can be employed to deal with unfairness in organizational settings (Bradfield & Aquino, 1999; Darley & Pittman, 2003; Wenzel et al., 2008). However, the focus on achieving justice and redressing unfairness through AI systems in organizations has been limited. To address this gap, it is crucial for designers in the HR industry to

consider how AI systems can identify and redress instances of unfairness. AI systems should be equipped to determine whether unfairness has occurred and provide employees with a pathway for redress through restorative or retributive justice (Robert et al., 2020).

Restorative justice and retributive justice are two distinct approaches utilized to address instances of unfairness within organizational contexts. Restorative justice involves bringing together the individuals affected by unfair actions to engage in open communication and seek resolutions that restore relationships and foster understanding (Bradfield & Aquino, 1999; Darley & Pittman, 2003; Wenzel et al., 2008). This approach aims to mend the social fabric and rebuild trust among individuals involved in the unfair situation.

On the other hand, retributive justice emphasizes the punishment or consequences that wrongdoers should face for their unfair actions. It is centered on the principle of proportionality, where the severity of the punishment is aligned with the severity of the offense (Bradfield & Aquino, 1999; Darley & Pittman, 2003; Wenzel et al., 2008). Retributive justice seeks to provide a sense of accountability and deterrence by imposing penalties that are perceived as just and fitting for the unfair behavior.

3.7.2 Fairness in Healthcare

The current assessment appears incomplete, as critical risks associated with AI in healthcare, notably algorithmic bias and inequality, have been overlooked. Despite limited research on AI fairness in medical applications, a recent notable study evaluated state-of-the-art deep neural networks using extensive chest X-ray datasets, examining patient attributes including sex, age, race, and insurance type, indicative of socioeconomic status (Seyyed-Kalantari et al., 2020). Findings demonstrated that models trained on large datasets do not inherently ensure equality of opportunity, potentially leading to care disparities if deployed without adjustments. The study employed true positive rates (TPR) to gauge fairness, though the literature offers alternative fairness metrics such as statistical parity, group fairness, equalized odds, and predictive equality (Barocas et al., 2017). Addressing fairness concerns in healthcare AI is vital for equitable patient care and positive societal impact.

3.7.3 Fairness in Industries

In the realm of addressing fairness in machine learning (ML) within the industry, researchers have sought to understand the perspectives of practitioners actively engaged in this endeavor. To gain insights, a study by Holstein et al. (2019) employed snowball sampling to interview members of teams whose ML-driven products had previously garnered media attention due to biases and unfairness. By targeting such practitioners, the study aimed to tap into their motivation to tackle fairness issues.

The research findings shed light on the dedication exhibited by interviewees towards enhancing fairness in their products, even when faced with challenges and lack of support from their team or company leadership. Many participants reported investing substantial time and effort to enhance fairness, underscoring their commitment to rectifying biases. The study's outcomes can be

perceived as representative of the hurdles that industry ML practitioners encounter while striving to promote fairness, even with strong intrinsic motivation (Holstein et al., 2019).

In essence, this investigation provides valuable insights into the barriers and challenges faced by industry practitioners in their pursuit of fairness improvement within ML systems. These insights serve to highlight the complexities involved in addressing fairness concerns in practical applications, shedding light on potential areas for improvement in the industry's approach to ensuring fairness in machine learning technologies (Holstein et al., 2019).

3.7.4 Fairness in Financial Industry

In the field of financial institutions, assessing fairness in artificial intelligence (AI) models is a key concern. To evaluate this, specific metrics are employed:

Statistical Parity Difference: This metric gauges the discrepancy in favorable outcomes between different groups, like privileged and unprivileged individuals. When the value is 0, it suggests equal benefits. Negative and positive values signify higher benefits for the privileged and unprivileged groups, respectively (Hardt et al., 2016).

Equal Opportunity Difference: This metric focuses on the variation in true positive rates among groups. When the value is 0, it signifies balanced benefits. Negative and positive values indicate greater benefits for the privileged and unprivileged groups, respectively (Hardt et al., 2016).

Disparate Impact: This metric assesses the ratio of positive outcome probabilities across groups. A value of 1 indicates parity in benefits. Values below 1 or above 1 indicate higher benefits for the privileged or unprivileged groups, respectively (Feldman et al., 2015).

In the financial sector, the equal opportunity measure holds particular importance. For instance, it ensures that loan applicants, regardless of characteristics like age, gender, or ethnicity, are treated fairly and have an equal chance of loan approval.

3.7.5 Fairness in Regulations across the Globe

In the context of the Fairness in the EU AI Act, the concept of explainability plays a crucial role. As highlighted by Madiega (2019), explainability encompasses the provision of explanations for algorithmic decision-making systems. This is a response to the challenge posed by the inherent complexity of machines and algorithms, often resulting in a lack of transparency regarding their behavior and processes. This opacity creates a "black box" effect, where AI systems generate outcomes without fully understandable explanations for their decisions. This lack of transparency raises concerns about fairness, as it can obscure potential biases and hinder the identification of discriminatory patterns in training data.

Explainability becomes a paramount requirement to ensure that AI systems are accountable and just. It encompasses not only shedding light on the technical operations of AI systems but also clarifying the human decisions made in line with EU guidelines. For fairness to be upheld, it is

imperative that the process by which AI systems influence decision-making, their design principles, and the underlying rationale for their deployment are clearly explained and accessible (MADIEGA, 2019).

In essence, the inclusion of explainability within the EU AI Act aligns with the broader goal of addressing fairness concerns and potential biases. By making AI systems more transparent and understandable, explainability promotes accountability, allows for the identification and rectification of discriminatory outcomes, and supports the overarching aim of ethical AI deployment in the European Union (MADIEGA, 2019).

3.7.6 AI Tool Kits

AI fairness toolkits are essential for addressing the challenges of fairness in AI. With a multitude of fairness definitions and bias handling algorithms, understanding how and when to use them can be challenging even for experts in algorithmic fairness. To provide clarity and guidance, the AI Fairness 360 (AIF360) toolkit has been developed. AIF360 is an non-industry specific toolkit widely being used. AIF360 is an extensible open-source toolkit designed to detect, understand, and mitigate unwanted algorithmic bias. Its primary goals are to promote a deeper understanding of fairness metrics and mitigation techniques, create a common platform for fairness researchers and industry practitioners to share and benchmark their algorithms, and facilitate the integration of fairness research into real-world industrial applications (Bellamy et al., n.d.).

FairTest is a toolkit that plays a significant role in the realm of fairness in machine learning ("FairTest", n.d.). It offers a general methodology to examine potential biases and feature associations in datasets and identifies regions where algorithms might exhibit higher errors. Another relevant toolkit is THEMIS, which focuses on black-box decision-making procedures and automatically generates test cases to explore possible group-based or causal discrimination. Additionally, fairness measures provide evaluation metrics for specific algorithms to assess their fairness (Friedler et al., 2019).

3.7.7 Fairness in the Legal Industry

The concept of non-discrimination is firmly enshrined in key United Nations human rights treaties (United Nations Legislation, 2012). However, anti-discrimination laws have taken distinct paths in common law and civil law countries. In common law nations like the U.S., U.K., and Australia, the legal development lacks systematic structure, leading to laws formed on a case-by-case basis (Schiek et al., 2007). In contrast, European Union Legislation (2012) and its member states follow a principled approach, encompassing a comprehensive range of discrimination grounds (Romei & Ruggieri, n.d.).

Various legal resources delve into this subject matter, such as Lerner (2003) for international group rights, Ellis (2005) for E.U. laws, and Bamforth et al. (2008) for U.S. laws. The discourse on discrimination contrasts formal equality and substantive equality (Barnard and Hepple, 2000). Formal equality dictates treating similar cases alike, focusing on individual merit and excluding irrelevant attributes (Romei & Ruggieri, n.d.). Substantive equality, in contrast, addresses differential treatment based on the circumstances of disadvantaged groups, aiming to achieve fair outcomes. Actions like affirmative measures and combating indirect discrimination stem from the distributive principle of justice, seeking substantive equality (Romei & Ruggieri, n.d.).

This perspective on discrimination highlights the nuanced approaches taken by different legal systems, each navigating the balance between formal and substantive equality in the pursuit of fairness (Romei & Ruggieri, n.d.).

3.7.8 Frameworks to address AI Fairness

Following are 2 such frameworks that address AI fairness:

- IEEE's Software and Systems Engineering Vocabulary Service: The Software and Systems Engineering Vocabulary service by the Institute of Electrical and Electronics Engineers (IEEE) offers a standardized repository of terms, definitions, and concepts relevant to the fields of software and systems engineering. It aims to foster clear and consistent communication among professionals in this domain, enhancing mutual understanding and minimizing potential ambiguities (IEEE, n.d.).

- ISO (International Organization for Standardization): The International Organization for Standardization (ISO) is an independent global organization known for creating and disseminating widely recognized standards. In the realm of software engineering and related disciplines, ISO produces guidelines, best practices, and specifications that promote quality and consistency in software development, management, and quality assurance. These standards, formulated through the collaboration of industry experts, serve as valuable resources to ensure the reliability, interoperability, and efficiency of software-related processes on a global scale (ISO, n.d.).

4. Discussion

The discussion section of this study seeks to address the research questions concerning fairness definitions accepted by the industry in the context of AI, the variations of these definitions for different applications and purposes, and the possibility of establishing a framework of AI definitions tailored for industry usage. The work by Robert et al. (2020), serves as a foundational reference for this discussion, offering insights into the current state of AI fairness literature.
In essence, this fundamentally provides insights into the prevailing state of the AI fairness

literature. However, it also brings to light significant shortcomings, which primarily include a lack of differentiation among types of fairness and insufficient consideration for the organizational context that surrounds AI systems and influences fairness. Additionally, they highlight that there is minimal coverage addressing how to effectively redress instances of unfairness once they have occurred.

As evident from the literature review, only a limited number of articles explicitly discussed distributive, procedural, and interactional fairness (Lind & Tyler, 1988; Bies & Moag, 1986; Rawls, 1971). However, despite this limitation that fairness is categorised as such the practical scenarios where it is being used in industries does not take all three into account, most of the articles were successfully categorized into the theoretical framework employed in this study. This suggests that while there is progress in addressing fairness in AI, there are still gaps and shortcomings in the existing literature.

One significant gap identified in the AI fairness literature is the lack of differentiation among types of fairness (Friedler et al., 2019). The research acknowledges that different fairness definitions might be more suitable for distinct applications and contexts. It is essential to recognize that fairness is a multifaceted concept, and different industries and applications may require specific fairness criteria. Therefore, it is crucial to delve deeper into the industry's perspectives on fairness definitions to establish a more context-aware framework (Verma & Rubin, 2018).

5. Limitations and Future Research Directions

The limitations section of this thesis acknowledges certain shortcomings in the research that warrant consideration for future work. One notable limitation is the lack of coverage on how to redress instances of unfairness that may occur in AI systems. While the study successfully identifies the importance of addressing unfairness like procedural unfairness and the need for rectification, it does not delve into specific methodologies or guidelines for implementing redress mechanisms. As a result, future work in this area could focus on developing practical approaches to address unfair outcomes in AI systems effectively. By exploring and implementing redress strategies, future research can contribute to the advancement of fairness in AI, promoting equitable decision-making and fostering a more inclusive and just AI-driven world (Robert et al.,2020).

Another critical aspect that the existing literature fails to extensively address is how to redress instances of unfairness after they occur. Redressing procedural unfairness is of paramount

importance when AI systems deviate from established procedures or demonstrate biases in practice. It is crucial to develop methodologies and guidelines for addressing and rectifying instances of unfairness to ensure equitable outcomes.

Moreover, the literature review also highlights a lack of consideration for the organizational context that surrounds AI systems and impacts fairness. AI systems are often deployed in complex organizational settings, and their implementation can be influenced by organizational policies, norms, and power structures (Gorwa, Binns, & Katzenbach, 2020). Understanding these contextual factors is crucial for ensuring fairness in AI decision-making processes within organizations.

To answer the research questions, it is essential to examine the definitions of fairness that are currently accepted by the industry for expressing fairness in AI. Robert et al. (2020) provide valuable insights into designing fair AI for managing employees in organizations, offering a starting point for understanding the industry's perspectives on fairness. Additionally, exploring how definitions of fairness change across different applications and purposes is imperative. Different industries and contexts may prioritize different fairness criteria, making it necessary to establish a more flexible and adaptive framework of definitions for AI fairness tailored for industry usage.

In conclusion, the research questions regarding fairness definitions accepted by the industry and their variations for different applications and purposes hold significant importance in fostering fair AI practices. The literature review and the work by Robert et al. (2020) offer valuable insights into the current state of AI fairness literature and highlight the need to address the identified shortcomings. Establishing a more comprehensive framework of AI definitions that considers industry usage and organizational contexts is crucial for promoting fairness in AI decision-making processes and ensuring equitable outcomes for various stakeholders.

The central focus of this thesis revolves around two pivotal research questions, each bearing significant importance in the pursuit of fairness in AI. The first research question centers on understanding the definitions of fairness that are widely accepted and implemented within the industry. By analyzing industry-accepted definitions, the research aims to shed light on the practical implications of fairness in AI and its applications in real-world business scenarios.

In addition to exploring the industry-accepted definitions, the second research question delves into the variability of fairness definitions across different applications and purposes. It is recognized that fairness is not a one-size-fits-all concept; rather, it requires careful consideration of the specific context and objectives of each AI application. Understanding how these definitions adapt and change in diverse contexts is crucial in developing tailored and context-specific approaches to address fairness challenges in AI. By investigating how fairness definitions evolve across various

applications, the research aims to provide valuable insights for businesses and AI practitioners seeking to foster equitable decision-making processes in their respective domains.

To address the research questions effectively, the thesis employs an empirical approach that includes analyzing real-world datasets from diverse business domains such as recruiting agencies and university admission processes. By examining the datasets, the research endeavors to uncover potential biases that may arise in the clustering process and understand how these biases infiltrate the data. This understanding is vital in the development of more fair algorithms and improved fairness definitions tailored to tackle specific biases found in business applications.

Furthermore, the thesis recognizes the significance of leveraging analytical models and algorithms to induce fairness into real-world business applications. Specifically, in domains like university admissions and recruitment, the implications of biased clustering can be far-reaching, affecting the opportunities and prospects of individuals . By exploring the practical implementation of proposed fair algorithms, the research aims to showcase how such algorithms can lead to fairer outcomes in critical business processes. This has the potential to transform AI applications, making them more equitable and just, while simultaneously contributing to the ongoing discourse on responsible AI development.

In conclusion, the research questions proposed in this thesis aim to delve into the definitions accepted by the industry for expressing fairness in AI and explore their variability across different applications and purposes. By drawing upon foundational research by Chhabra et al. (2021) and examining real-world case studies, the thesis seeks to provide a comprehensive understanding of fairness in AI and its practical relevance in real-world business settings. The findings of this research have the potential to guide businesses, fairness researchers, and AI practitioners in fostering equitable decision-making processes and promoting responsible AI applications that serve the broader interests of society.

6. Conclusion

In conclusion, achieving fairness in AI is a multifaceted and critical endeavor that requires a deeper understanding of distinct fairness types and the development of a coherent design agenda. As the field of AI continues to evolve and influence various aspects of organizational functioning, ensuring that AI systems are fair and just becomes paramount for managing workers and fostering equitable outcomes (Robert et al., 2020).

In conclusion, the research on fair AI design for managing employees in organizations is of significant importance due to the potential impact on worker experiences, customer experiences, and organizational outcomes. While AI fairness has received considerable attention, there remains a lack of literature focused on developing a theoretical and systematic design agenda for fair AI.

Other reasons behind the lack of literature include the fact that companies cannot share their data due to privacy concerns, confidentiality clauses, non-disclosure agreements, etc., or they simply wish to protect their reputation. Addressing this gap is crucial for several reasons. Firstly, the current discussions often overlook the distinct types of fairness, such as fairness of outcomes, process, or interactions, leading to a risk of adopting a one-size-fits-all approach in AI fairness design. Secondly, the absence of an overarching theoretical framework hinders the organization and integration of design solutions within the broader HCI community, resulting in a fragmented understanding of the problem and design space related to AI fairness.

Regarding the research questions, RQ1 explores the definitions accepted by the industry for expressing fairness in AI. While the dissertation discusses various fairness definitions, it highlights the lack of differentiation among types of fairness, emphasizing the need to acknowledge the diverse dimensions of fairness in AI design.

RQ2 investigates how definitions of fairness change for different applications and purposes. Although the dissertation does not explicitly address this question, it points out the challenges of designing fair AI due to the complexity and variability of fairness considerations across contexts.

RQ3 inquires about the possibility of having a framework of definitions of AI based on industry usage. While the dissertation does not provide a specific framework, it underscores the importance of developing a theoretical design agenda for fair AI, which can serve as a guide for industry practitioners to address fairness concerns in AI systems effectively.

In the broader landscape, the realm of AI fairness presents a complex convergence of AI, Software Engineering, and Law. The emergence of various fairness definitions underscores the evolving discourse on fairness in algorithmic decision-making. While a unanimous agreement on the suitable definition remains a challenge, researchers like S. Verma & J. Rubin (2018) have contributed by extensively exploring principal fairness definitions and shedding light on their practical applications. In conclusion, the concise nature of the literature review can be attributed to the limited availability of publications in this emerging field of study. The scarcity of relevant papers indicates that this domain is still in its infancy, prompting challenges in sourcing extensive literature. This inherent lack of existing resources underscores the pioneering nature of this research, highlighting the need for future contributions to further enrich the understanding and knowledge base in this evolving area.

The concept of Fair AI, aimed at ethically designing and implementing AI systems to ensure equity in decision-making, holds transformative potential. As AI technologies play an increasingly

significant role in critical scenarios such as loan allocation and hiring, their susceptibility to bias underscores the urgent need for fairness in AI. Instances like the COMPAS tool's racial bias illustrate the real-world impact of unchecked algorithmic discrimination.

Enforcing fairness in AI is not only ethically imperative but also economically significant. It aligns with legal initiatives like the EU AI ACT and Canada's AIDA, providing a framework for responsible AI development and deployment. Transparency, accountability, and trust are integral aspects of achieving fairness in AI, promoting a holistic approach toward building AI systems that mitigate biases and ensure equitable outcomes.

Understanding the sources of bias, encompassing data, algorithm, and user biases, is fundamental in combating unfairness. Fairness definitions in AI, categorized into group-level and individual fairness, offer a structured approach to addressing these issues. However, the challenge lies in selecting the most appropriate definition for a given context.

Across industries like Human Resources, addressing unfairness through AI involves concepts of restorative and retributive justice. AI fairness toolkits such as AIF360, FairTest, and THEMIS provide crucial resources for detecting, understanding, and mitigating bias in AI systems.

Legal definitions and approaches to discrimination vary globally, with common law and civil law countries adopting distinct paths. Formal and substantive equality represent two contrasting paradigms that legal systems navigate while seeking fairness. This diversity in approaches underscores the multifaceted nature of fairness in AI and its intersection with legal frameworks.

In conclusion, addressing the intricacies of AI fairness necessitates interdisciplinary collaboration, continuous research, and the ethical commitment to designing AI systems that empower equitable decision-making. By striving to integrate fairness into AI technologies, industries, and societies can harness the full potential of AI while ensuring that its benefits are accessible to all.

References:

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv preprint arXiv:1810.01943.

Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An Overview of Fairness in Clustering. IEEE Access, 9, 130698–130720. https://doi.org/10.1109/ACCESS.2021.3114099.

Feuerriegel, S., Dolata, M., Schwabe, G., & Schwabe, Á. G. (2020). Challenges and Opportunities. Business & Information Systems Engineering, 62(4), 281-285. https://doi.org/10.1007/s12599-020-00650-3.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In ACM SIGKDD (pp. 259-268).

DOI: 10.1145/2783258.2783311

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An Intersectional Definition of Fairness. 2020 IEEE 36th International Conference on Data Engineering (ICDE), 1918–1921. https://doi.org/10.1109/ICDE48307.2020.00203.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data and Society, 7(1). https://doi.org/10.1177/2053951719897945.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In NIPS (pp. 3315-3323).

DOI: 10.5555/3157382.3157460

Holstein, K., Wortman, J., Daumé, H., Dudík, M., Wallach, H., & Vaughan, J. W. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? https://doi.org/10.1145/3290605.3300830

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6). https://doi.org/10.1145/3457607.

Pi, Y. (2021, August 24). Machine learning in Governments: Benefits, Challenges and Future Directions. https://scite.ai/reports/10.29379/jedem.v13i1.625.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. Information Systems Research, 31(2), 381-404. https://doi.org/10.1287/isre.2019.0898.

Romei, A., & Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 00, 0–1. https://doi.org/10.1017/S000000000000000.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton and D. Roth, A comparative study of fairness-enhancing interventions in machine learning *, https://doi.org/10.1145/3287560.3287589.

Seyyed-Kalantari, L., Moradi, M., & Ghassemi, M. (2020). Large-scale analysis of clinical notes uncovers gender and age disparities in machine learning models. arXiv preprint arXiv:2005.08303.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings - International Conference on Software Engineering, 1–7. https://doi.org/10.1145/3194770.3194776.

Bahadoran, P., Labeau, M., & Masi, M. (2019). Discrimination in Algorithmic Decision-Making: A Comprehensive Survey. arXiv preprint arXiv:1908.08151.

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, *1*, 2017.

Brinkmann, M. M., Fricke, L. V., Diedrich, L., Robra, B., Krauth, C., Dreier, M. (2022). Attributes In Stated Preference Elicitation Studies On Colorectal Cancer Screening and Their Relative Importance For Decision-making Among Screenees: A Systematic Review. Health Economics Review, 1(12). https://doi.org/10.1186/s13561-022-00394-8.

Chellasamy, B., Manogaran, G., Priyan, M. K., & Sundarasekar, R. (2014). An Efficient Secure Data Transmission Model for Cluster-Based Wireless Sensor Networks. Journal of Sensors, 2014, 1-13. https://doi.org/10.1155/2014/541057.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and Opportunities. Business and Information Systems Engineering, 62(4), 379–384. https://doi.org/10.1007/s12599-020-00650-3.

Li, C., Liu, M., Li, J., Wang, W., Feng, C., Cai, Y., … & Qu, J. (2022). Machine Learning Predicts the Prognosis Of Breast Cancer Patients With Initial Bone Metastases. Frontiers in Public Health, (10). https://doi.org/10.3389/fpubh.2022.1003976.

MADIEGA, T. A. (2019). EU guidelines on ethics in artificial intelligence: Context and implementation.

N. Lerner. Group Rights and Discrimination in International Law. Martinus Nijhoff Publishers, 2 edition, 2003.

D. Schiek, L. Waddington, and M. Bell, editors. Cases, Materials and Text on National, Supranational and International Non-Discrimination Law. Hart Publishing, 2007.

E. Ellis. EU Anti-Discrimination Law. Oxford University Press, 2005.

C. Barnard and B. Hepple. Substantive equality. The Cambridge Law Journal, 59(3):562–585, 2000.

European Union Legislation. (a) European Convention on Human Rights, 1950; (b) Racial Equality Directive, 2000; (c) Employment Equality Directive, 2000; (d) Gender Goods and Services Directive, 2004; (e) Gender Employment Directive, 2006; (f) Equal Treatment Directive (proposal), 2008, 2012. http://eur-lex.europa.eu.

Rawls, J. (1971). A Theory of Justice. Harvard University Press.

Lind, E. A., & Tyler, T. R. (1988). The Social Psychology of Procedural Justice. Springer.

Bies, R. J., & Moag, J. F. (1986). Interactional Justice: Communication Criteria of Fairness. Research on Negotiation in Organizations, 1(1), 43-55.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

FairTest. (n.d.). FairTest. https://fairtest.org/.

IEEE. (n.d.). Software and Systems Engineering Vocabulary Service. Retrieved from https://pascal.computer.org/sev_display/index.action.

ISO (International Organization for Standardization). (n.d.). ISO/IEC TR 24027:2015(en) - Information technology - Security techniques - IT security management - Guidelines for information security management systems auditing. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24027:ed-1:v1:en.