6-18-2022

AI Fairness at Subgroup Level – A Structured Literature Review

Luis Lämmermann
University of Bayreuth, luis.laemmermann@uni-bayreuth.de

Patrick Richter
University of Bayreuth, patrick.richter@uni-bayreuth.de

Amelie Zwickel
University of Bayreuth, amelie.zwickel@uni-bayreuth.de

Moritz Markgraf
FIM Research Center, University of Augsburg, moritz.markgraf@fim-rc.de

AI FAIRNESS AT SUBGROUP LEVEL – A STRUCTURED LITERATURE REVIEW

Research Paper

Luis Lämmermann, FIM Research Center, University of Bayreuth, Bayreuth, Germany, Project Group Business & Information Systems Engineering of the Fraunhofer FIT, Bayreuth, Germany, luis.laemmermann@uni-bayreuth.de

Patrick Richter, University of Bayreuth, Bayreuth, Germany, patrick.richter@uni-bayreuth.de

Amelie Zwickel, University of Bayreuth, Bayreuth, Germany, amelie.zwickel@uni-bayreuth.de

Moritz Markgraf, FIM Research Center, University of Augsburg, Augsburg, Germany, Project Group Business & Information Systems Engineering of the Fraunhofer FIT, Augsburg, Germany, moritz.markgraf@fim-rc.de

Abstract

AI applications in practice often fail to gain the required acceptance by stakeholders due to unfairness issues. Research has primarily investigated AI fairness on individual or group levels. However, increasing research indicates shortcomings in this two-fold view. Particularly, the non-inclusion of the heterogeneity within different groups leads to increasing demand for specific fairness consideration at the subgroup level. Subgroups emerge from the conjunction of several protected attributes. An equal distribution of classified individuals between subgroups is the fundamental goal. This paper analyzes the fundamentals of subgroup fairness and its integration in group and individual fairness. Based on a literature review, we analyze the existing concepts of subgroup fairness in research. Our paper raises awareness for this primary neglected topic in IS research and contributes to the understanding of AI subgroup fairness by providing a deeper understanding of the underlying concepts and their implications on AI development and operation in practice.

Keywords: artificial intelligence, fairness, subgroup, literature review.

1 Introduction

The importance of the application areas of artificial intelligence (AI) has grown in recent years. Especially in areas such as medical support, human resources, and manufacturing, AI supports humans in decision making (e.g., Bayer et al., 2021; Jarrahi, 2018; Kasie et al., 2017; Magrabi et al., 2019). The steadily increasing performance of statistical models makes data-driven decisions surpass human intuition (Dawes et al., 1989). Research has identified various advantages of AI-enabled decision-making, such as absolute reliability (e.g., no human-like signs of fatigue) or incorporation of a wide variety of factors exceeding human cognitive capacity (Mehrabi et al., 2019). Consequently, AI is increasingly responsible for decision-making and, thus, influences our human society (Agrawal et al., 2018). Nonetheless, data-driven predictors, natural language models, or rule-based classifiers for decision support also have their downsides (Feuerriegel et al., 2020). Relying on AI in decision-making may be disastrous since it can easily be compromised by biases in the data and hence, promote discrimination or unfairness that potentially harms its users and the other stakeholders (Feuerriegel et al., 2020). Therefore, it is of great importance that AI-enabled decision-making ensures fairness in terms of ethical, social, and legal compliance (Larsson et al., 2019).

To create fair AI, it is of utmost importance to consider fairness as early as the first spark of the idea including AI. For this, though, it is essential to establish a common understanding of the actual meaning and implementation of fairness-compliant AI. The goal of such a "fair AI" is to provide "decision support that prevents disparate harm (or benefit) to different [groups]" (Feuerriegel et al., 2020, p. 379). In doing so, fair AI aims to provide systems with the ability to both quantify bias and mitigate discrimination against diverse groups. However, within society, the term *fairness* is perceived differently across individuals (Dwork and Ilvento, 2018b) and groups (Dwork and Ilvento, 2018a) distinguishing the independent, equal treatment of individuals and the equal distribution of desired outcomes. Consequently, there is no universal fairness and it is impossible to fulfill all (mathematical) expressions of fairness simultaneously. Thus, it is indispensable to identify the part of society that needs protection (Feuerriegel et al., 2020; Kleinberg et al., 2016; Mehrabi et al., 2019). So far, research has focused on the consideration of fairness at the individual or the group level (Mehrabi et al., 2019). However, Foulds et al. (2020), Mehrabi et al. (2019), and Kearns et al. (2017), among others, emphasize that this two-level perspective is too short-sighted. Considering the heterogeneity created by the specific characteristics of the individual subgroups can significantly bias underlying data and, therefore, cause unfair and inaccurate predictions (Mehrabi et al., 2019). Here, subgroups are "defined by a set of functions G of the protected attributes" (Kearns et al., 2019a, p. 101). When comparing two occupational groups, such as physicians and administrative staff, in the job application process, a fair distribution of individuals also includes equal representations of men, women, and others, for example. When considering multiple dimensions (e.g., gender and occupational groups) simultaneously, the population can appear evenly distributed on group level at first sight indicating that there is an equal number of physicians and administrative staff, as well as men and women. Only when analyzing at the subgroup level an imbalance becomes apparent, wherein, for example, women in administration and men in medical professions have higher chances of hiring. Consequently, it is crucial to examine subgroup fairness and internalize its implications for understanding fairness as a whole. As of now, there are various understandings of subgroup fairness as a theoretical concept complicating straightforward progress. To structure the field of research on AI fairness at the subgroup level, point out its theoretical analogies and differences, and derive a basis for common understanding, we pose the following research question.

How does existing research conceptualize AI fairness at the subgroup level from a technical perspective? What are its implications for socio-technical research?

To answer our research questions, we conduct a structured literature review to analyze the prevailing concepts of subgroup fairness in AI. We create a broader understanding of the importance of considering subgroup fairness and bringing previously disparate research together. We contribute to the important discussion around AI fairness and derive a basis for future research in terms of a research agenda.

The remainder of this paper is structured as follows. First, we introduce the theoretical concepts concerning the fundamental terms and definitions of fairness (i.e., section 2). Further, we shed light on the potential sources for AI (un-)fairness by following AI applications' conceptual data processing procedures (i.e., section 2). Thus, since there is a multitude of possible definitions and views, we introduce the basic concepts of fairness to create a basis for a common understanding. In section 3, we briefly present our research methodology and continue, in section 4, with our results of the structured literature review on the current state of AI fairness at the subgroup level. In section 5, we discuss our findings and conclude our paper with an indication of a possible research agenda.

2 Theoretical Background

2.1 Fairness in Artificial Intelligence

The concept of AI fairness and unfairness depends on the individual's objectives (e.g., developer, user) and the overall fairness perspective. To provide a relevant theoretical foundation and raise awareness for the relevance of fairness in AI, we set out outlining critical points for (un-)fairness along the AI

pipeline. Therefore, we divide the AI pipeline into three phases *data pre-processing*, *training*, and *deployment*, as suggested by Hummer et al. (2019).

Considering the data pre-processing, if the data is historically biased, for example, the AI will mirror this bias, leading to undesired biases and unfairness. A famous example is the Amazon AI recruiting system described in Dastin (2018) favoring men over women, which illustrates that it is crucial to pay attention to the validity (construct, external, internal), the data quality (sparsity, noise, representativeness), the temporal variations (of populations and behaviors), and to the sources that harm different points of the pipeline like biases (Bellamy et al., 2018; Mehrabi et al., 2019; Olteanu et al., 2019; Suresh and Guttag, 2019). Particularly, biases as "systematic deviation of an estimated parameter from true value" (Feuerriegel et al., 2020, p. 381) even affect the complete AI pipeline and can lead to unfairness and discrimination through some unwanted or socially unfavorable outcomes. Different biases may even correlate with each other, and mitigation is possible only to a limited extent at the current state. Furthermore, it is essential to take the impossibility of complete objectivity in data-driven systems into account. As soon as the underlying data includes vulnerable constructs (e.g., social, ethical constructs) such as race or gender, bias becomes an inevitable characteristic of data collection from human processes (Dignum, 2019). Due to the continuous attempt to capture human behavior and values in AI applications, it is almost unavoidable that no single questionable human perspective finds its way into data representation (Zehlike et al., 2020).

Regarding training, especially on skewed data, it is well-known that the accuracy of an AI algorithm might not be a sufficient metric. Instead, optimization of both fairness and accuracy metrics may be conflicting, as investigated by Haas (2019). Furthermore, there are plenty of other metrics and even such more suitable for skewed data, for instance, yet those still usually neglect fairness. Choosing the appropriate metrics among the rich set of performance and quality metrics is a fuzzy task with very few guidelines to adhere to. As a result, developers may ground metrics selection on unfair or biased assumptions.

In the final stage of the AI pipeline, it is important to consider the actual use in production. As a matter of AI's data-driven nature, it retrieves patterns in the data and, thus, mirrors previous behavior. The focus on decision-making, which is based on decision data from the past, is an established procedure also present in the field of jurisdiction – e.g., in the United States of America – called *common law*, in which precedents play a significant role (Walter, 2015). However, there are also different approaches such as *civil law*, in which previous judgments play an insignificant or rather no role – predominant in Europe (Tetley, 1999). In terms of fairness, it is controversial what is "more" fair. Regarding the effect on the deployment phase of AI, it is important to be aware of these contrary concepts and to consciously brief the users to enable fair interpretation of the AI output. Otherwise, users may (un-)intentionally misinterpret AI outputs promoting unfairness (Ferrer et al., 2021).

2.2 Fairness and Discrimination

In the context of fairness considerations, the underlying literature refers either to the technical aspects of bias and fairness in AI or to the socio-ethical, legal implications of discrimination caused by AI systems (Caton and Haas, 2020). In accordance with the existing body of knowledge, our paper analysis AI subgroup fairness primarily from a technical perspective yet aims to derive implications for the socio-technical context, particularly by addressing gaps and avenues for future IS research. A considerable amount of research attempts to define the broad topic of *fairness* quantitatively and qualitatively in its entirety and all its various nuances (e.g., Barocas et al.; Barocas and Selbst, 2016; Dunkelau and Leuschel, 2020; Dwork et al., 2012; Friedler et al., 2018; Hardt et al., 2016; Suresh and Guttag, 2019). The enormous growth of this relatively new field of research has led to inconsistent terminology, making it increasingly difficult to compare and classify the different definitions. Yet, aiming for a greater overview, Corbett-Davies and Goel (2018) present a well-founded classification by distinguishing between three formal types of definitions for fairness: Anti-classification, classification parity, and calibration. *Anti-classification* refers to protected attributes like gender or race that are only used implicitly or not at all for decision-making. This includes the fairness concept named unawareness

(Kusner et al., 2017). *Classification parity* describes the equality of frequently used measurement criteria (e.g., true-positive rate and false-positive rate) across groups regarding the protected attribute. This means that each group has the same predictive probability of being assigned to the desired outcome (Caton and Haas, 2020). These particularly include concepts such as group fairness to which research refers as *statistical parity* (e.g., demographic parity) or *equal acceptance* (Dwork et al., 2012; Mitchell et al., 2021) and *predictive parity* (Chouldechova and G'Sell, 2017) which in turn includes predictive equality, equality of opportunity, and equalized odds (Hardt et al., 2016). *Calibration* (or test-fairness) describes the independence of the outcomes from the protected attributes due to risk assessments (Chouldechova and G'Sell, 2017). The risk perspective serves as a theoretical counter-design to the desired outcomes approach since subsequent decisions are often influenced by risk estimates that quantify the consequences of future decisions (Corbett-Davies and Goel, 2018). Additional important concepts not yet covered in the three-part division are individual fairness (or fairness through awareness) (Dwork et al., 2012) and causality(including counterfactual fairness (Kusner et al., 2017). For more detailed explanations of the fairness concepts, we refer to the research of Dunkelau and Leuschel (2020).

Unfairness as the opposite concept is also important to mention as it leads to various forms of discrimination. Direct discrimination or disparate treatment is caused by a different treatment of individuals based on their sensitive attributes (Dunkelau and Leuschel, 2020; Romei and Ruggieri, 2014). Indirect discrimination or disparate impact, on the other hand, leads to non-favorable outcomes due to implicit effects, which are caused by correlations of seemingly neutral attributes with the protected characteristics. The correlation of the zip code/neighborhood with race, for instance, can lead to indirect discrimination in AI applications (Zhang et al., 2017). Another type of discrimination is systematic discrimination that disadvantages through perpetuated behavior and rules which are no longer contemporary (Mehrabi et al., 2019). Furthermore, statistical discrimination by using average group statistics for individual cases can lead to misjudgment of individuals. Overall, hard-to-record characteristics are often overlooked, even though they are an important part of the decision-making process (Mehrabi et al., 2019).

2.3 Individual and group fairness and the needed connection in between

Analyzing existing research, we recognized some weaknesses within the common fairness conceptualizations concerning anti-classification, classification parity, and calibration. The main feature of the anti-classification principle is also the most significant point of criticism: In many cases, it is necessary to look at the explicitly protected attribute to make appropriate risk assessments and predictions (Corbett-Davies and Goel, 2018). The generally low prospect of equality poses the problem of the calibration concept. It is relatively easy to establish test fairness by strategically misclassifying individuals and applying standards like those used when considering groups or majorities. Classification parity, on the other hand, often suffers from the statistical limitation of inframarginality (Corbett-Davies and Goel, 2018). The *inframarginality* principle refers to a fair and equal society and, therefore, considers differences between groups in data and predictive algorithms as legitimate (Foulds et al., 2020). However, if the real risk distributions vary between the groups, the actual error rates will also differ (Corbett-Davies and Goel, 2018). If the various groups and subgroups were treated equally, the non-inclusion of heterogeneity generated by subgroups' specific characteristics would lead to unfairness (Mehrabi et al., 2019). Existing research mainly distinguishes between group and individual fairness and only includes subgroup fairness to a limited extent. Individual fairness describes the group-independent equal treatment of individuals who can be assumed to be similar based on common measurement criteria (Gillen et al.; Kearns et al., 2019b; Lahoti et al., 2019). However, the concept of individual fairness assumes an ideal feature space in which the similarity between individuals can be calculated and optimally extracted from the available data (Binns, 2020; Feuerriegel et al., 2020). The counterpart group fairness describes the probability of being assigned to the desired outcome, which should be equally distributed across the privileged and unprivileged groups. In addition, there is also the distinction between within groups (women vs. men) and between groups (Caton and Haas, 2020). Group fairness is part of the classification parity and is therefore limited by inframarginality. A promising concept to address the shortcomings from above is the concept of intersectionality casting a different

angle on the concepts of fairness. Notably, through the research of Kearns et al. (2017) and Hébert-Johnson et al. (2017), intersectionality received renewed attention. It postulates that the different risk distributions are often influenced by unjust social processes (Foulds et al., 2020). As a result, the heterogeneity of the particular subgroups is usually overlooked. To address this potential cause of unfairness and discrimination, it may be beneficial to explicitly consider fairness at the subgroup level. As a consequence, increasing research demands the extension of individual fairness and group fairness by subgroup fairness (e.g., Dwork and Ilvento, 2018a; Foulds et al., 2020; Kearns et al., 2017, 2019a).

3 Method

We conduct a literature review to identify and acquire relevant research publications and to get a comprehensive overview of the research that has been done in our relevant area of AI fairness at the subgroup level. The acquired body of literature forms the foundation for the development of a conceptual overview of research on AI subgroup fairness. We follow the general approach of Webster and Watson (2002), extended with methodological elements from Siddaway et al. (2019) regarding search string design.

Based on the knowledge gained from an initial unstructured literature search, we identified keywords in the context of fairness, artificial intelligence, and the subgroup perspective as the focus of the paper. We deliberately take different synonyms, spellings, classification terms into account to ensure coverage of relevant articles in our selected research area (Siddaway et al., 2019). Based on the focus areas, fairness, subgroup level, and artificial intelligence, we formulated different keywords to operationalize our research question. The keywords within the first search term aim to ensure focus on fairness as our overall research topic. The second search term further specifies the focus on articles investigating fairness at the subgroup level and its related concepts. Since we observed that researchers mostly derive subgroup fairness directly from individual and group fairness concepts, we included all three fairness concepts to avoid blind spots. The third search term defines our technical scope around artificial intelligence in which we aim to investigate subgroup fairness. In table 1, we present the final search string we used for the literature search. Our searched databases are the Association for Information Systems eLibrary (AISeL) for the information systems (IS) perspective, Web of Science (WoS) database for the broader scope, and Arxiv for the computer science perspective. While publications retrieved from AISeL and WoS are mostly peer-reviewed, we note that Arxiv includes pre-prints that have neither been peer-reviewed nor published elsewhere yet. However, the focus of the CS research community on publishing pre-prints parallel to peer-review processes offers the opportunity to include the latest research in our literature search. Accordingly, we deliberately included Arxiv in our literature search. Furthermore, we did not limit the time frame of our search.

Table 1. Search string of the literature search

[Title]		**[All Fields/Topic]**		**[All Fields/Topic]**
(discriminat* OR disparit* OR fair* OR unfair* OR equal* OR inequal* OR bias* OR classifi* OR measur*)	**AND**	("subgroup fairness" OR "individual fairness" OR "group fairness" OR "statistical parity" OR "intersectionality" OR "gerrymandering" OR "subgroup" OR "sub-group" OR "subpopulation")	**AND**	(algorithm* OR "artificial intelligence" OR "machine learning" OR "data processing" OR "decision support")

Our initial searches resulted in an aggregated number of 680 papers; thereof, 309 identified papers in WoS; 51 identified papers in AISeL; 220 identified papers in Arxiv. After removing duplicates, we manually screened each paper in a two-stage exclusion process to see if it matched our relevant research question. In this process, it was possible to identify ambiguities of keywords in titles and abstracts, leading to irrelevant results that we excluded from further consideration. Furthermore, we excluded

papers from further consideration if the titles and abstracts did not explicitly consider the concepts of fairness as a central object of research or if they were too focused on a specific domain. In doing so, we screened the titles reducing the literature sample to 63 papers and then the abstract leaving 24 papers for the subsequent full-text search. Throughout the full-text search, we excluded another nine papers. Predefined exclusion criteria were applied in the individual steps of the screening process. Due to the ambiguity of some keywords, some articles were explicitly excluded. Furthermore, we excluded articles dealing with a too specific research focus, for instance, on medical, legal, and philosophical topics, or missing focus on artificial intelligence. Thus, we excluded papers solely focusing on fairness at the group and/or individual levels without addressing subgroup-relevant concepts. To be considered relevant (i.e., inclusion criteria), the studies had to be consistent with our content framework on the socio-technical aspects of subgroup fairness and its related concepts. Included research had to explicitly investigate the fairness at an intersectional or subgroup level or explicitly derive subgroup fairness as an extension building on related concepts like individual or group level (e.g., group-overlapping, consideration of several sensitive attributes simultaneously). Overall, we identified 15 papers as relevant through our regular search; thereof eight papers from WoS, one paper from AISeL, six papers from Arxiv. Through forward and backward searching, we additionally identified another seven specific articles to integrate into our portfolio. The final sample of literature search comprises 22 publications. Since the Arxiv database includes many pre-prints that may have been published elsewhere in the meantime, we finally checked whether the relevant papers had been published at conferences or journals. In doing so, we avoid the risk of potential duplicates in our final set through backward and forward-searching.

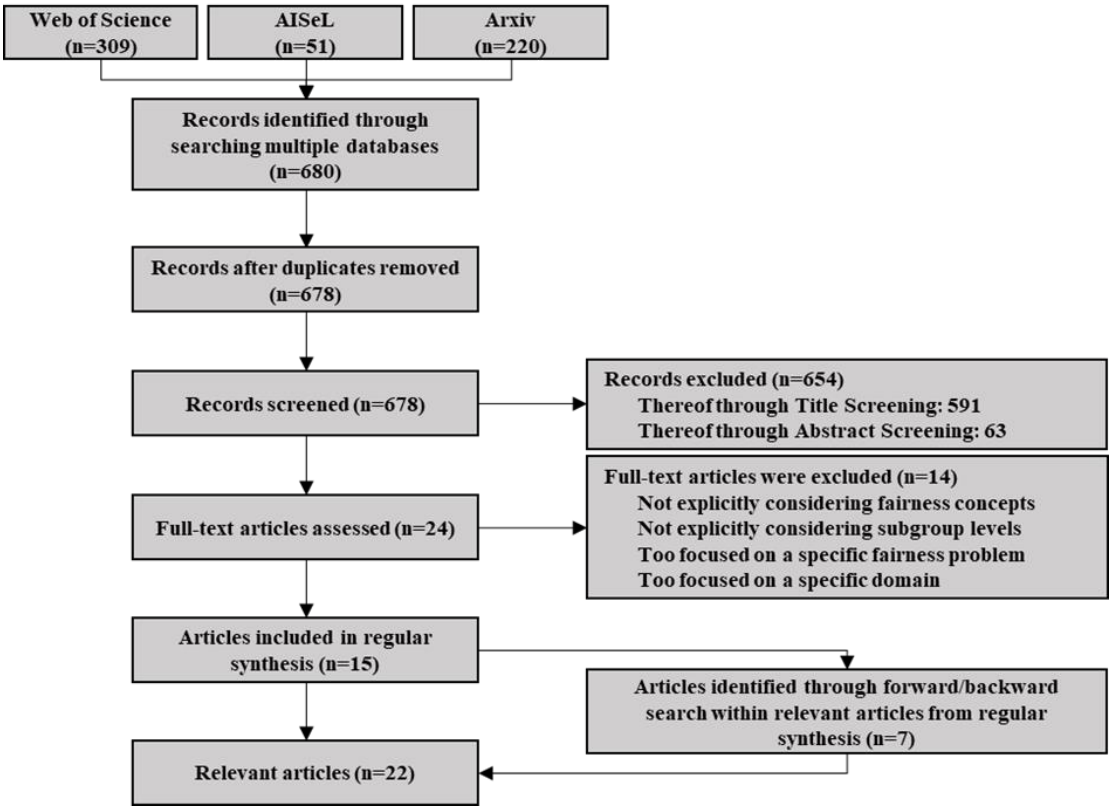

Figure 1. Literature screening process adapted from Tricco et al. (2018)

4 Research Results

4.1 Descriptive Analysis

This study analyzes 22 articles published between 2016 and 2021. Even though researchers have investigated algorithmic fairness on group or individual attribute levels long before that time (Dwork et al., 2012; Kamishima et al., 2011), the publication dates of our set of relevant articles show that research on AI fairness at the subgroup level has begun relatively late with 2016. Furthermore, we recognize that the research stream of AI subgroup fairness is mainly driven by the computer science domain. Analyzing the research fields, 21 out of 22 articles stem from the computer science field. Only the study of Teodorescu et al. (2021) published in the special issue on *Managing AI* of *MIS Quarterly* clearly refers to the information systems research domain. Furthermore, literature analysis reveals that 17 out of 22 papers have initially been published as pre-prints at Arxiv and ten out of the 17 pre-prints have been published by peer-reviewed journals and conferences in the meantime. At the time of writing, the set of 22 relevant papers divides into the following research outlets: journal (5), conference (8), pre-print (9).

4.2 Conceptual analysis of AI fairness at the subgroup level

To display the current state of research on subgroup fairness, table 1 summarizes the identified 22 papers addressing subgroup fairness. Illustrating the awareness of subgroup fairness in research, we classify the papers based on their degree of conceptualization of subgroup fairness. 13 papers deal with the issue of subgroup fairness and call for the need to consider it independently of the group and the individual fairness. The remaining nine papers, though, consider subgroup fairness as a separate concept of fairness but do not primarily discuss it.

Analyzing the conceptualization, we identified three different streams distinguishing between ‘mathematical notions’ (7), ‘practical algorithms and frameworks in subgroup fairness’ (7), and ‘counter mechanism for practice’ (9). Mathematical notions are concrete quantitative assessments of fairness and create comparability and evaluation opportunities. Practical algorithms and frameworks provide a working basis that can be used in subgroup fairness and implemented in practical applications, e.g., to highlight subgroups. Counter-mechanisms for practice are differentiated approaches to fairness treatment and discrimination detection, both for general fairness and subgroup fairness.

Table 2. Concept matrix

Articles			**Concepts**		
	Dedicated focus on the conceptualization of subgroup fairness	Contextual derivation of subgroup fairness as an extension for group and individual fairness	Mathematical notions	Practical algorithms and frameworks	Counter mechanism for practice
Binns, 2020		x			x
Cabrera et al., 2019	x			x	
Caton and Haas, 2020	x				x
Chouldechova and G'Sell, 2017	x			x	
Dunkelau and Leuschel, 2020		x			x
Dwork and Ilvento, 2018a	x		x		
Foulds et al., 2020	x		x		
Hébert-Johnson et al., 2017	x		x		
Joseph et al., 2017	x			x	
Kearns et al., 2019a	x		x		
Kearns et al., 2017	x		x		x
Kim et al., 2018	x		x		
Mehrabi et al., 2019		x			x
Miron et al., 2020		x			x
Mitchell et al., 2021		x			x
Pastor et al., 2021	x		x		
Rahmattalabi et al., 2019		x		x	
Raji and Buolamwini, 2019	x				x
Saleiro et al., 2018		x		x	
Teodorescu et al., 2021		x			x
Zehlike et al., 2020		x		x	
Zhang and Neill, 2016	x			x	

Mathematical notions aim to formalize subgroup fairness based on the findings of the individual and group fairness concepts. Dwork and Ilvento (2018a) point out weaknesses of group fairness formalizations, distinguishing among competition between related tasks, unrelated tasks, and group fairness under composition without competition. Competition between related tasks describes the conflict of classifiers satisfying conditional parity in isolation but no longer under competition. When considering the competition between unrelated tasks, statistical parity does not consider all protected subgroups but may discriminate subgroups equally. Classifiers can be trained to iteratively improve the algorithm with respect to conditional parity. However, it does not follow that general inequality is implemented equally for all subgroups. Composition without competition concentrates on the outcome of an OR-formulation either without or under conditional parity. In both mentioned cases (elements treated equally and not treated equally), failures for conditional parity arise for individuals. Kearns et al. (2017), Kearns et al. (2019a), Hébert-Johnson et al. (2017), and Kim et al. (2018) consider the constraints for subgroups resulting from the existing fairness definitions. Kearns et al. (2017) for instance, describe the occurrence of biases caused by evaluating statistical notions of fairness that are

only applicable across a small number of subgroups, known as gerrymandering. To address this, the introduced *false-positive subgroup fairness* based on statistical parity assigns positive labels at equally probable assignments. This may allow the identification of subgroups with different classifications. (Caton and Haas, 2020). However, this approach has limitations because there are endless subgroups when considering all notions of statistical fairness at the same time. Thus, it can be not feasible or rather lead to a state of overfitting – every individual is their subgroup. To address this issue, a subsequent study (Kearns et al., 2019a) focuses on the practical application by introducing an iterative heuristic that continually adjusts the most disadvantaged subgroup to its favor. Another detached approach is *multicalibration* as a fairness notion requiring that any value predicted by the algorithm should be approximately equally distributed between actual positively labeled individuals and a subset of individuals predicted by the algorithm (Hébert-Johnson et al., 2017). Thus, multicalibration offers predictions not only for the average of a population but also for subsets addressing a common weakness among calibration. This approach analyses the goal of "perfect predictions" that Kim et al. (2018) extend in their work by focusing on error rates. Additionally, this implementation extends by the metric fairness condition of Dwork et al. (2012), a statistical condition from the subgroup level. Foulds et al. (2020) favor another approach contrasting the opposing principles of intersectionality and inframarginality to develop *differential fairness*. It seeks to enhance the *statistical parity subgroup fairness* (Kearns et al., 2017) as it lacks to properly include minorities. Particularly, minorities are worthy of protection and thus, the newly introduced criterion considers multidimensional and intersectional categories. By this, it aims to achieve similar probabilities for desired outcomes "regardless of the composition of the protected attributes" (Foulds et al., 2020, p. 1919).

Further seven papers conceptualize subgroup fairness by means of **practical algorithms and frameworks**. The concept of practical algorithms and frameworks refers to papers that investigate algorithmic approaches or code-based frameworks (e.g., toolkits, etc.) for practice in order to accomplish subgroup fairness. Joseph et al. (2017), for instance, build upon John Rawls' notion of *fair equality of opportunity* (Rawls, 2009) and theoretically and experimentally evaluate an algorithm to analyze the distribution of individuals within the group. Complementary, Rahmattalabi et al. (2019) explore robust graph covering problems with fairness constraints to identify the misaligned subgroups and maximize group fairness. Zhang and Neill (2016) highlight the auditing phase and focus on the classification of algorithms to avoid discrimination of undefined subgroups. Their work focuses on single models evaluating the estimated event probabilities under the influence of certain characteristics. Chouldechova and G'Sell (2017) provide another approach aiming not to avoid the potentially misleading overall performance. They elaborate a framework for automatic subgroup detection by extracting relevant variables of interest (i.e., fairness properties). Those variables form the basis for distinguishing subgroups that enable to validate the fairness among the resulting homogeneous subgroups. Saleiro et al. (2018) develop a practical approach to an open-source audit toolkit for biases and fairness which considers standardized AI metrics, especially for subgroup fairness. Relatedly, Cabrera et al. (2019) introduce a visual analytics system that can identify and examine subgroups in datasets as well as suggest homogeneous subgroups.

The remaining nine papers propose **counter-mechanisms** concerning subgroup unfairness. Whereas some counter-mechanisms are specific to subgroups, others are more general at the fairness level, yet, with explicit applicability to the subgroup level. For a general overview, Mehrabi et al. (2019) provide a taxonomy classifying the different fairness and bias types, which enables identifying approaches specialized in different kinds of subgroups, for example. Dunkelau and Leuschel (2020) and Caton and Haas (2020) follow in line, yet, extend an overview of fairness and bias with various fairness-enhancing and discrimination-detecting algorithms for application in the specific phases of the data pipeline. Additionally, Miron et al. (2020) show, based on juvenile criminal recidivism, that static demographic characteristics are the source of inequalities in group fairness metrics. Moreover, they point out that equalizing the outcomes for one protected group does not automatically result in equitable outcomes for the respective subgroups. This can lead to positive as well as negative discrimination, which ultimately falsely leads to an aggregated *fair* assessment. Mitchell et al. (2021) elaborate on fairness-related terminology and assumptions as well as on various metrics for assessing model fairness. As a result, the

authors suggest using prior information to increase the utility of a decision system, thus avoiding the assumption that benefits and harms are equally distributed across decisions. This prior information can lead to a more realistic representation of the actual relationships, increasing fairness and contributing to subgroup formation. As a specific example, Raji and Buolamwini (2019) investigate disparities based on gender and skin color and show that using an auditing process leads to greater fairness. At last, Binns (2020) examines the conflict between individual and group fairness by defining the use of inconcrete fairness measures as the relevant cause rather than the concepts themselves. This shows that in the practical application case, various fairness measures must first be checked for suitability, and subgroup fairness should be adequately evaluated in the overall concept of individual and group fairness.

5 Discussion

Based on the predominant fairness approaches in the literature (i.e., statistical parity subgroup fairness (Kearns et al., 2019a), multi-calibration (Hébert-Johnson et al., 2017), differential fairness (Foulds et al., 2020), and the various characteristics of subgroup fairness identified in the literature, we compile several interesting insights.

One of the most important goals of subgroup fairness is the protection of minorities. Assuming that the basic potential does not vary substantially across groups, AI applications in production environments should strive for equal treatment between groups. Furthermore, all combined values of the protected attributes should be protected by the fairness definition. Nevertheless, the protection of individual attribute values should still be ensured. This means that combinations such as ‘black woman’ must be protected just as the individual values ‘black’ or ‘woman’ (Foulds et al., 2020). Additionally, we recognize that the composition of the groups remains challenging since a harmonized balance is necessary. Subgroups need to be as small as possible but still robust against gerrymandering (Caton and Haas, 2020; Kearns et al., 2017). Analyzing the subgroup fairness concepts with respect to individual and group fairness to a broader realm, we expect that considering AI fairness at the subgroup level as a complementing cornerstone may further improve overall fairness.

In terms of practical implications, we present a selection of actions that may further improve subgroup fairness along the AI pipeline if successfully implemented. In the **data pre-processing** phase (i.e., data preparation, data acquisition), the appropriate subgroup selection is one of the crucial first steps. As the presented examples illustrate, single selected subgroups are fragile, and thus, a variety of possible subgroups is rather recommendable. Another approach favors neglecting sensitive attributes to unbias the dataset (Valentim et al., 2019). This approach is a seemingly proven approach to ensure fairness, yet indirect discrimination can occur through attributes associated with sensitive attributes, leading to bias in evaluating subgroup fairness. Benthall and Haynes (2019) propose an unsupervised machine learning process, as existing discriminations are based on human actions and assumptions (Ochmann and Laumer, 2019). In practice, for each classification of input data conducted by a person, information of the author, the time stamp, and the circumstances should be documented. In this way, later analyses may reveal whether particular motivations are embedded in the process, indicating intrinsic behavior, leading to unfairness. Furthermore, in databases with underrepresented groups, it is difficult to facilitate a fair algorithm (Caton and Haas, 2020; Mehrabi et al., 2019). Multitask learning, however, offers a solution enabling the simultaneous learning of multiple tasks (Dwork et al., 2017; Oneto et al., 2019). This allows a wide variety of subgroups to be identified and more accurate models to be developed (Caton and Haas, 2020; Zehlike et al., 2020). For a high acceptance of AI applications, transparency of implemented classifiers that affect decision-making is critical (Raji and Buolamwini, 2019). In the **training** phase, a fair classifier on discriminatory datasets can be beneficial (Lohia et al., 2018). Potential methods are adjusted learning algorithms, subgroup fairness constraints in the optimization function, and training of an individual classifier for each subgroup (Caton and Haas, 2020; Dunkelau and Leuschel, 2020; Fitzsimons et al., 2019; Miron et al., 2020). In the **deployment** phase, relabeling could lead to steady improvements with each further iteration. However, Kearns et al. (2017) correctly note that precise adjustment for subgroups does not always lead to a general improvement in fairness but can reduce it again. Overall, the successful integration of AI subgroup fairness may not be seen as an isolated fairness

approach neglecting other fairness goals to be fulfilled. Successful AI fairness integration requires careful trade-offs both horizontally (i.e., concerning potential conflicts within the multitude of AI subgroup fairness goals) and vertically (i.e., concerning individual and group fairness goals). Appropriate measures need to be weighed on a case-by-case basis. Overall though, we would like to emphasize that individually perceived fairness relies on subjective values and expectations by the individual rather than being objectively accountable across all stakeholders. Consequently, we argue to communicate the AI application's fairness assumptions to the respective stakeholders to ensure user acceptance in real-life environments. The proper communication of AI fairness assumptions, however, is a non-trivial task as fairness may not always be clearly understandable for the different stakeholder types. Most users of real-world AI applications are domain and process experts of the underlying use case rather than academics or AI developers with in-depth fairness expertise. Therefore, we recommend considering the complexity of fairness approaches when communicating them to the stakeholders. The simplicity of fairness definitions (e.g., low granularity, few subgroups) are easier to communicate and incur a lower cognitive cost to understand. In contrast, more complex fairness definitions, or even alternative versions of definitions, might be "fairer" in terms of individual subgroups but incur a higher cognitive cost for the user to understand the underlying differences.

5.1 Limitations and Future Research

Despite carefully following established methodological standards, our paper is subject to several limitations. Our research approach relies on technology-driven literature analysis, including the major databases for Computer Science and IS research (i.e., WoS – incl. IEEE, ACM – and AISeL, Arxiv). We cover the technology-focused scientific literature by ensuing a thorough backward and forward search. However, an even greater understanding could be drawn by including other databases focusing on rather ethical, behavioral and sociological aspects. In addition, the literature review limits to a subset of publications after applying exclusion criteria. Despite deliberately paying attention not to miss essential papers within the sampling process, there still might be the residual risk of missing a relevant paper. Thus, a larger-scale analysis (e.g., using additional synonyms as keywords, additional search fields) may reveal further concepts. Nevertheless, we are confident that we incorporated the most relevant papers into our literature sample and further concepts will most likely underpin them or expand them.

Although the existing concepts of AI subgroup fairness offer a promising basis for making AI applications fairer and more accepted among stakeholders, the investigated concepts still face several shortcomings. The shortcomings occur on several levels. On a superordinate level, one relevant shortcoming is the *ambiguous integrability* of the respective fairness concepts and approaches. While some concepts can be applied simultaneously, we also recognize various concepts that conflict with each other when being combined. For instance, the fairness approaches that rely on anti-classification concepts are somewhat incompatible with fairness calibration approaches. Another shortcoming is the *challenging assignment and connectivity* of the numerous AI subgroup fairness approaches to concrete real-world fairness issues. Since there are many approaches available, AI developers may be unable to select the best concept for the given use case. Furthermore, the investigated AI subgroup fairness concepts build their fairness understanding predominantly on a *unified AI fairness perception* among the stakeholders. However, for certain cases, we recognize this view as too simplified as it neglects the sociological variations of fairness opinions among the AI applications stakeholders. Although the investigated subgroup fairness concepts face concrete shortcomings, they can offer actionable areas for future IS research: To address the ambiguous integrability, future research should analyze the identified concepts in terms of compatibility. To shed light on the challenging assignment of the existing fairness concepts to concrete fairness problems, future IS research may analyze their connectivity through a socio-technical perspective on the underlying use cases. The shortcoming of unified AI fairness perceptions is also worth being investigated by IS research. Future IS research should evaluate stakeholders' variations in fairness understandings and perceptions to develop more differentiated and specific fairness targets for AI applications.

Besides deriving future IS research from the shortcomings of our analyzed subgroup fairness concepts on a superordinate level, we also recognize the relevance for future IS on a more specific level. By this, we emphasize that further research is needed to indicate which level of granularity, i.e., which trade-off should be considered in a given use case or scenario. From an IS perspective, this raises the following research questions: What level of granularity of fairness is required by users? This research question helps to shed light on whether there is a cognitive limitation (i.e., "too detailed" granularity level) where users of the AI application start having trouble differentiating between the different levels of fairness.

The current state of research indicates that the analyzed AI subgroup fairness concepts require more validation in production environments across different industries. The concepts must be elaborated even more precisely and confirmed in validation studies. Therefore, future research can draw on our findings to validate subgroup fairness approaches in practice and to derive design principles for subgroup-fair AI development. Design principles help to promote the transfer of theoretical concepts into productive environments by capturing "the knowledge gained about the process of building solutions" (Sein et al., 2011, p. 45) for subgroup-fair AI applications. From an IS perspective, researchers should ensure a clear use-case focus taking socio-technical aspects of the underlying business case and the stakeholder requirements into account. One resulting IS research question could be: *What are the design requirements for subgroup-fair AI applications across all phases of the lifecycle from development to operation from a socio-technical perspective?* Also, it may be reasonable to investigate suitable monitoring principles for ensuring subgroup fairness throughout AI use. In doing so, future research may target how one can develop AI applications to make fair decisions at the subgroup level and, thus, what AI production environments are suitable. Monitoring principles may, thus, help to ensure fairness compliance after AI deployment by unveiling potential sources of unfairness (e.g., discrimination). To develop design and monitoring principles, it is essential to have in-depth knowledge about both the existing AI subgroup fairness concepts and the stakeholders' fairness perception. Future IS research should, therefore, address the following research question: *How does subgroup fairness perception vary among stakeholders, and how do stakeholders prioritize subgroup fairness compared to individual or group fairness concepts in trade-off situations?* Therefore, our analysis of the relevant concepts in existing literature contributes as a valuable basis for future research. Future IS research should, thus, address the issues related to hidden correlations and confusion of protected attributes at the subgroup level. In doing so, we suggest answering the following research question: *How can we systematically identify hidden correlations and confusions of clearly identifiable protected attributes with other variables of the AI dataset?* Counterfactual fairness approaches, for example, provide an opportunity to test the influence of substituting sensitive variables (Kusner et al., 2017). Incorporating causal graph approaches allows to estimate the influence of potential biases and to break them down so that the causes of unfairness can be targeted systematically even at the subgroup level considering several protected attributes simultaneously. Besides expanding the research stream horizontally with previously stated research approaches, future research could also focus on expanding the body of research vertically by deepening the insights of our conceptual findings at the subgroup level of AI fairness. Consequently, research should investigate the subconcepts underneath the subgroup concepts we identified in our literature review. For instance, researchers may investigate what different types of counter-mechanisms exist at the subgroup level and put them in context with the other concepts captured by our literature analysis.

5.2 Contribution

This paper contributes to the theoretical knowledge base of the IS research field. Our analysis of the derivations of the different fairness definitions shows the necessity to consider subgroup fairness since the current focus on individual and group fairness illustrates the insufficient deliberation of intersectional connections between groups, i.e., the connection of multiple protected attributes. The assumption of inframarginality and the disregarded heterogeneity of subgroups carries the risk of increasing fairness issues and discrimination through AI applications in production environments. Our research raises awareness for AI subgroup fairness and emphasizes its importance for the IS research community. Particularly, since we identified only one relevant publication on AI subgroup fairness

within the IS research domain through our literature analysis, our study can be a useful starting point for further IS research on AI subgroup fairness.

Furthermore, by analyzing the existing literature about the characteristics of AI fairness at the subgroup level and the delimitation of AI at individual and group levels, we foster a multi-domain perspective on subgroup fairness. As our results illustrate, research in the area of AI subgroup fairness is represented by multiple streams primarily favored by various disciplines. In order for IS research to contribute to this area in a straightforward, structured manner, we facilitate further research by creating a common understanding of the subgroup concept.

6 Conclusion

This paper reviews current research on AI subgroup fairness and analyzes its different concepts in literature. We observe the concepts from two perspectives, namely, subgroup fairness as a conceptual extension of group and individual fairness, and subgroup fairness as a stand-alone concept within the scope of universal fairness. Although the majority of the literature refers to the core papers (Dwork and Ilvento, 2018a; Kearns et al., 2017, 2019a), the main hindrance within the literature is the heterogeneous understanding of the subgroup fairness concept itself. Furthermore, the literature review underpins the necessity of considering subgroup fairness as a stand-alone concept that is not sufficiently mapped by individual and group fairness. Our reviewed concepts within the context of subgroup fairness are mathematical notions, practical algorithms and frameworks, and counter mechanisms for practice. Even though these approaches differ in their methods of addressing subgroup fairness, they all aim to protect minorities through the consideration of protected attributes. Through our research, we contribute to the existing literature by extending the conceptual understanding of subgroup fairness and indicating how future research can promote fairness in AI applications by considering fairness at the subgroup level.

Acknowledgments

We gratefully acknowledge the Bavarian Ministry of Economic Affairs, Regional Development and Energy for their support of the project “Fraunhofer Blockchain Center (20-3066-2-6-14)” that made this paper possible.

7 References

Agrawal, A., J. Gans and A. Goldfarb (2018). *Prediction machines. The simple economics of artificial intelligence.* Boston, USA: Harvard Business Review Press.

Barocas, S., M. Hardt and A. Narayanan. *Fairness and machine learning*. URL: https://fairmlbook.org/ (visited on 11/09/2021).

Barocas, S. and A. D. Selbst (2016). “Big Datas Disparate Impact” *California Law Review* 104 (3).

Bayer, S., H. Gimpel and M. Markgraf (2021). “The role of domain expertise in trusting and following explainable AI decision support systems” *Journal of Decision Systems* Forthcoming.

Bellamy, R. K. E., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney and Y. Zhang (2018). “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias”.

Benthall, S. and B. D. Haynes (2019). “Racial categories in machine learning”. In: *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*: ACM, pp. 289–298.

Binns, R. (2020). “On the apparent conflict between individual and group fairness”. In: *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency*: ACM, pp. 514–524.

Cabrera, Á. A., W. Epperson, F. Hohman, M. Kahng, J. Morgenstern and D. H. Chau (2019). *FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning*. URL: https://arxiv.org/abs/1904.05419v4 (visited on 11/15/2021).

Caton, S. and C. Haas (2020). *Fairness in Machine Learning: A Survey*. URL: https://arxiv.org/abs/2010.04053v1 (visited on 11/15/2021).

Chouldechova, A. and M. G'Sell (2017). *Fairer and more accurate, but for whom?* URL: https://arxiv.org/abs/1707.00046 (visited on 11/15/2021).

Corbett-Davies, S. and S. Goel (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. URL: https://arxiv.org/abs/1808.00023 (visited on 11/15/2021).

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on 11/10/2021).

Dawes, R., D. Faust and P. Meehl (1989). "Clinical Versus Actuarial Judgment" *Science* 243 (4899), 1668–1674.

Dignum, V. (2019). *Responsible Artificial Intelligence*. *How to Develop and Use AI in a Responsible Way*. Cham, Germany: Springer.

Dunkelau, J. and M. Leuschel (2020). *Fairness-Aware Machine Learning: An Extensive Overview*. URL: https://www3.hhu.de/stups/downloads/pdf/fairness-survey.pdf (visited on 11/09/2021).

Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. Zemel (2012). "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*: ACM, pp. 214–226.

Dwork, C. and C. Ilvento (2018a). "Group fairness under composition". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*: ACM.

Dwork, C. and C. Ilvento (2018b). "Individual fairness under composition. https://www.fatml.org/schedule/2018". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*: ACM.

Dwork, C., N. Immorlica, A. T. Kalai and M. Leiserson (2017). *Decoupled classifiers for fair and efficient machine learning*. URL: https://arxiv.org/pdf/1707.06613v1.pdf (visited on 11/15/2021).

Ferrer, X., T. van Nuenen, J. Such, M. Cote and N. Criado (2021). "Bias and Discrimination in AI: A Cross-Disciplinary Perspective" *IEEE Technology and Society Magazine* 40 (2), 72–80.

Feuerriegel, S., M. Dolata and G. Schwabe (2020). "Fair AI. Challenges and Opportunities" *Business & Information Systems Engineering* 62 (4), 379–384.

Fitzsimons, J., A. Al Ali, M. Osborne and S. Roberts (2019). "A General Framework for Fair Regression" *Entropy* 21 (8).

Foulds, J. R., R. Islam, K. N. Keya and S. Pan (2020). "An Intersectional Definition of Fairness". In: *Proceedings of the 36th International Conference on Data Engineering*: IEEE, pp. 1918–1921.

Friedler, S. A., C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton and D. Roth (2018). "A comparative study of fairness-enhancing interventions in machine learning".

Gillen, S., C. Jung, M. Kearns and A. Roth. "Online Learning with an Unknown Fairness Metric".

Haas, C. (2019). "The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness". In: *Proceedings of the 40th International Conference on Information Systems*: AIS.

Hardt, M., E. Price and N. Srebro (2016). "Equality of Opportunity in Supervised Learning".

Hébert-Johnson, Ú., M. P. Kim, O. Reingold and G. N. Rothblum (2017). "Calibration for the (Computationally-Identifiable) Masses".

Hummer, W., V. Muthusamy, T. Rausch, P. Dube, K. El Maghraoui, A. Murthi and P. Oum (2019). "ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI". In: *Proceedings of the 7th International Conference on Cloud Engineering*: IEEE, pp. 113–120.

Jarrahi, M. H. (2018). "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making" *Business Horizons* 61 (4), 577–586.

Joseph, M., M. Kearns, J. Morgenstern, S. Neel and A. Roth (2017). *Rawlsian Fairness for Machine Learning*. URL: https://arxiv.org/pdf/1610.09559v3.pdf.

Kamishima, T., S. Akaho and J. Sakuma (2011). "Fairness-aware Learning through Regularization Approach" *2011 11th IEEE International Conference on Data Mining Workshops*, 643–650.

Kasie, F. M., G. Bright and A. Walker (2017). "Decision support systems in manufacturing: a survey and future trends" *Journal of Modelling in Management* 12 (3), 432–454.

Kearns, M., S. Neel, A. Roth and Z. S. Wu (2017). "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness".

Kearns, M., S. Neel, A. Roth and Z. S. Wu (2019a). "An Empirical Study of Rich Subgroup Fairness for Machine Learning". In: *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency*: ACM, pp. 100–109.

Kearns, M., A. Roth and S. Sharifi-Malvajerdi (2019b). "Average Individual Fairness: Algorithms, Generalization and Experiments".

Kim, M. P., O. Reingold and G. N. Rothblum (2018). "Fairness Through Computationally-Bounded Awareness" *Advances in Neural Information Processing Systems (NeurIPS 2018)* 2018 (31), 4842–4852.

Kleinberg, J., S. Mullainathan and M. Raghavan (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores".

Kusner, M. J., J. R. Loftus, C. Russell and R. Silva (2017). "Counterfactual Fairness".

Lahoti, P., K. P. Gummadi and G. Weikum (2019). "iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making". In: *Proceedings of the 35th International Conference on Data Engineering*: IEEE, pp. 1334–1345.

Larsson, S., M. Anneroth, A. Felländer, L. Felländer-Tsai, F. Heintz and R. C. Ångström (2019). *Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence.* AI Sustainability Center. URL: https://lucris.lub.lu.se/ws/portalfiles/portal/62833751/Larsson_et_al_2019_SUSTAINABLE_AI_web_ENG_05.pdf (visited on 11/10/2021).

Lohia, P. K., K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney and R. Puri (2018). "Bias Mitigation Post-processing for Individual and Group Fairness".

Magrabi, F., E. Ammenwerth, J. B. McNair, N. F. de Keizer, H. Hyppönen, P. Nykänen, M. Rigby, P. J. Scott, T. Vehko, Z. S.-Y. Wong and A. Georgiou (2019). "Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications" *Yearbook of Medical Informatics* 28 (1), 128–134.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan (2019). "A Survey on Bias and Fairness in Machine Learning".

Miron, M., S. Tolan, E. Gómez and C. Castillo (2020). "Evaluating causes of algorithmic bias in juvenile criminal recidivism" *Artificial Intelligence and Law* 29, 111–147.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour and K. Lum (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions" *Annual Review of Statistics and Its Application* 8 (1), 141–163.

Ochmann, J. and S. Laumer (2019). "Fairness as a Determinant of AI Adoption in Recruiting: An Interview-based Study". In: *Proceedings of the 19th Diffusion Interest Group in Information Technology*: AIS.

Olteanu, A., C. Castillo, F. Diaz and E. Kıcıman (2019). "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries" *Frontiers in Big Data* 2.

Oneto, L., M. Doninini, A. Elders and M. Pontil (2019). "Taking Advantage of Multitask Learning for Fair Classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*: ACM, pp. 227–237.

Pastor, E., L. de Alfaro and E. Baralis (2021). "Identifying Biased Subgroups in Ranking and Classification".

Rahmattalabi, A., P. Vayanos, A. Fulginiti, E. Rice, B. Wilder, A. Yadav and M. Tambe (2019). "Exploring Algorithmic Fairness in Robust Graph Covering Problems", 15750–15761.

Raji, I. D. and J. Buolamwini (2019). "Actionable Auditing". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*: ACM, pp. 429–435.

Rawls, J. (2009). *Theory of Justice.* Cambridge, USA: Harvard University Press.

Romei, A. and S. Ruggieri (2014). "A multidisciplinary survey on discrimination analysis" *The Knowledge Engineering Review* 29 (5), 582–638.

Saleiro, P., B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa and R. Ghani (2018). "Aequitas: A Bias and Fairness Audit Toolkit".

Sein, M., O. Henfridsson, S. Purao, M. Rossi and R. Lindgren (2011). "Action Design Research" *MIS Quarterly* 35 (1), 37–56.

Siddaway, A., A. Wood and L. Hedges (2019). "How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses" *Annual review of psychology* 70, 747–770.

Suresh, H. and J. V. Guttag (2019). "A Framework for Understanding Unintended Consequences of Machine Learning".

Teodorescu, M., L. Morse, Y. Awwad and G. Kane (2021). "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation" *Management Information Systems Quarterly* 45 (3), 1483–1500.

Tetley, W. (1999). "Mixed Jurisdictions: Common Law v. Civil Law (Codified and Uncodified)" *Louisiana Law Review* 60, 677.

Tricco, A. C., E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. J. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garritty, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp and S. E. Straus (2018). "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation" *Annals of internal medicine* 169 (7), 467–473.

Valentim, I., N. Lourenco and N. Antunes (2019). "The Impact of Data Preparation on the Fairness of Software Systems". In: *Proceedings of the 30th International Symposium on Software Reliability Engineering*: IEEE, pp. 391–401.

Walter, M. K. (2015). *Verfassungsprozessuale Umbrüche. Eine rechtsvergleichende Untersuchung zur französischen Question prioritaire de constitutionnalité.* Zugl.: Bielefeld, Univ., Diss., 2013/14. Tübingen: Mohr Siebeck.

Webster, J. and R. Watson (2002). "Analyzing the past to prepare for the future: Writing a literature review" *MIS Quarterly* 26 (2), xiii–xxiii.

Zehlike, M., P. Hacker and E. Wiedemann (2020). "Matching code and law: achieving algorithmic fairness with optimal transport" *Data Mining and Knowledge Discovery* 34 (1), 163–200.

Zhang, L., Y. Wu and X. Wu (2017). "A Causal Framework for Discovering and Removing Direct and Indirect Discrimination". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935.

Zhang, Z. and D. B. Neill (2016). "Identifying Significant Predictive Bias in Classifiers".