Approaches to Improve Fairness when Deploying AI-based Algorithms in Hiring – Using a Systematic Literature Review to Guide Future Research

Jonas Rieskamp
Paderborn University, Germany
jonas.rieskamp@upb.de

Lennart Hofeditz
University of Duisburg-Essen, Germany
lennart.hofeditz@uni-due.de

Milad Mirbabaie
Paderborn University, Germany
milad.mirbabaie@upb.de

Stefan Stieglitz
University of Duisburg-Essen, Germany
stefan.stieglitz@uni-due.de

Abstract

Algorithmic fairness in Information Systems (IS) is a concept that aims to mitigate systematic discrimination and bias in automated decision making. However, previous research argued that different fairness criteria are often incompatible. In hiring, AI is used to assess and rank applicants according to their fit for vacant positions. However, various types of bias also exist for AI-based algorithms (e.g., using biased historical data). To reduce AI's bias and thereby unfair treatment, we conducted a systematic literature review to identify suitable strategies for the context of hiring. We identified nine fundamental articles in this context and extracted four types of approaches to address unfairness in AI, namely pre-process, in-process, post-process, and feature selection. Based on our findings, we (a) derived a research agenda for future studies and (b) proposed strategies for practitioners who design and develop AIs for hiring purposes.

Keywords: fairness in AI, SLR, hiring, AI implementation, AI-based algorithms

1. Introduction

Human resource (HR) departments face vast amounts of applicants. To process the candidates' data, the use of artificial intelligence (AI)-based algorithms has become an established tool for an increasing number of organizations (Black & van Esch, 2020; Fernández-Martínez & Fernández, 2020; Marks, 2022; Mayer, Anne-Sophie; Strich, Franz; and Fiedler, 2020). The use of AI-based algorithms can reduce the workload of the employees that are involved in the recruiting process (Hofeditz et al., 2022). For example, an AI can screen résumés and examine data of applicants to reduce the amount of data for manual review (Marks, 2022; Raghavan et al., 2019). Hereinafter, we define AI-based algorithms according to Berente et al. (2021) as "the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems".

However, the use of AI-based algorithms can have multi-faceted consequences: First, decisions made by algorithms can be perceived as less fair and less trustworthy than decisions suggested by a human which is a phenomenon known as algorithm aversion (Mahmud et al., 2022). Second, the decisions made by algorithms in hiring directly affect the applicants since some of them are rejected through the system. Thus, an AI-based algorithm that treats all candidates fairly is demanded by both the human recruiters and the candidates. While the formers are interested in being perceived as fair, the latter are concerned about being treated fairly. In this sense, a tool is considered fair if it predicts performance accurately without disturbance due to candidates' belonging to a minority group (Alder & Gilbert, 2006). Although fairness may not come to mind when first thinking about the issues of modern AI-based algorithms because it is prominently thought about as being more objective than its human counterpart but recent studies showed otherwise: Teodorescu et al. (2021) reported gender imbalances in job advertisement targeting by algorithms, and Zehlike et al. (2017) also found gender discrimination in the search algorithm on the website XING[1].

Unfairness is often rooted in cognitive biases like the home bias, similarity bias, or stereotypes that cause recruiters to favor particular applicants over others (Liang, Chen; Hong, Yili; Gu, 2018; Soleimani et al., 2021). Unfairness of AIs, in turn, is due to their dependency on data that contains past decisions' biases and prejudices (Zhong, 2018). This happens because an AI primarily learns from observations but if the provided observations are contaminated with biases, the AI adopts this perspective (Leavy, 2018). Often, the biases in data pertains attributes like sex, ethnicity, or age which harms certain groups or individuals if an AI learns to disfavor them (Chakraborty et al., 2021). Teodorescu et al. (2021) argued that it is impossible to

[1] https://www.xing.com/

URI: https://hdl.handle.net/10125/102654
978-0-9981331-6-4

design general fairness criteria for AI-based algorithms as different fairness criteria are often incompatible and need human intervention. Thus, we consider it even more important to seek out for up- and downstream approaches that alter the in- or output of an AI-based algorithm and to guide humans using such algorithms to improve fairness for candidates.

Motivated by the aforementioned issues of unfairness in the hiring processes caused by AI-based algorithms, we aim to identify the most common, practical, and promising approaches to increase fairness in AI. Previous research mostly focused on a metalevel of fairness in AI as for instance Mirbabaie et al. (2022). However, an overview of possibilities for the technical implementation of fair AI-based algorithms in the application field of hiring does not exist but is urgently needed to improve fairness in hiring. Hence, we state the following research question: *Which approaches to improve fairness in AI-based algorithms exist in literature in the context of recruiting and hiring?*

To address this question, we performed a systematic literature review according to vom Brocke et al. (2015), focusing on features in AI classification and ranking. By answering this question, we provide an outline of suitable measures regarding the examined issue which can guide scholars as well as practitioners in designing intelligent applications for HR. Additionally, our work is embedded into the sociotechnical view of algorithmic fairness (Dolata et al., 2022). Specifically, our work addresses the technical sub-system which is said to be the "origin of discrimination" and seldom yields more than conceptual results (Dolata et al., 2022). Thus, our contributions comprise (a) a research agenda for future studies to address the issue properly, and (b) strategies for practitioners who design and develop AIs for hiring purposes.

2. Background

2.1. Use of AI-based algorithms in hiring

AI-based algorithms have become indispensable in the contemporary hiring process (Black & van Esch, 2020). Since AI-based algorithms can improve productivity (Mayer, Anne-Sophie; Strich, Franz; and Fiedler, 2020), it is employed in HR departments to assist the hiring process (Fernández-Martínez & Fernández, 2020). It is able to scan applications and score candidates' fit for the position (Savage & Bales, 2017). Recent trends led to AI-assisted video interviews in which an AI-based algorithm analyzes the candidates' voice and face to evaluate verbal, paraverbal, and nonverbal information (Fernández-Martínez & Fernández, 2020).

Another of AI's tasks in hiring is the screening of job applications. Companies that apply an AI-based algorithm for this task can reduce their application process time and time-to-hire significantly (Black & van Esch, 2020). Based on collected features, those AI-based algorithms rank the candidates according to their fit for the position (Mujtaba & Mahapatra, 2019). Apart from candidate ranking, classification of résumés is used to categorize candidates according to whether they match certain criteria based on textual properties (Habous & El Habib, 2021). Human recruiters are still needed to monitor the automated process, and they finally decide on whether to hire a candidate (Black & van Esch, 2020; Goretzko & Israel, 2022). In summary, AI-based algorithms are used to assist human recruiters but are not yet replacing them.

2.2. Fairness in hiring

Defining fairness is nontrivial itself. However, a definition that is appropriate to the context at hand is to apply and interpret the same rules to each agent objectively and consistently (Hooker, 2005). In alignment with extant literature (Pessach & Shmueli, 2022), we use the terms related to *unfairness*, *bias*, and *discrimination* interchangeably in a similar sense.

Apart from using AI-based algorithms, unfair treatment is not a new issue in hiring. Although discrimination is prohibited by law in some countries, Pager et al. (2009) found discrimination to the disadvantage of ethnical minorities. In a similar vein, biases in favor of a specific gender were found to influence the hiring process, but they partly vanished when employers gained experience with the process (Chan & Wang, 2018; Harvie et al., 1998). In addition, studies showed that even seemingly small biases can lead to substantial consequences in terms of both discrimination and productivity loss (Hardy et al., 2022). Yet discrimination does not pertain ethnicity and gender only, also properties like age, sexual orientation, and disabilities cause unequal treatment in the hiring process (Bendick & Nunes, 2012). Those so-called protected attributes can divide candidates into groups of privileged and unprivileged candidates where the unprivileged group is substantially underrepresented and thus discriminated against (Pessach & Shmueli, 2022). In extant literature, two types of discrimination are distinguished: Firstly, *disparate treatment* is the intentional different treatment of candidates on the grounds of the belonging to a certain group. Secondly, *disparate impact* denotes negative consequences by apparently objective rules. The types are known as direct and indirect discrimination respectively (Pessach & Shmueli, 2022).

Wolgast et al. (2017) observed that an unstructured hiring process facilitates discrimination resulting in unfairness. Thus, they concluded that structured processes, focusing on skills led to greater equity for the candidates. Consul et al. (2021) agreed to the structure approach and emphasized that it must comprise unambiguous selection criteria. In addition, they suggested dedicated personnel to oversee the diversity of the candidates considered.

2.3. Fairness of AI-based algorithms in hiring

Fairness in AI-based algorithms is especially important if it performs crucial decisions that affect humans' lives like in hiring. We must expect an AI-based algorithm to not systematically discriminate against anyone. Thus, protected attributes should not have any influence on the selection of suitable candidates to avoid unfair treatment.

There are different reasons why AI-based algorithms that are applied in hiring could have some form of bias towards a specific group of people. Mehrabi et al. (2021) created a detailed list of the different sources of bias that can occur in AI and can cause unfair treatment. Some of the most prevalent in the context of recruiting are following: *Measurement bias* refers to general errors in the gathering of data, where some metrics that are used to assess a candidate's suitability are not measured in a precise enough way (Suresh & Guttag, 2021). The *representation bias* describes a situation where the data is not showing a true depiction of the relevant population. Certain groups could be underrepresented and thus assessed less accurately (Suresh & Guttag, 2021). *Algorithmic bias* refers to biases that are caused solely by the algorithm itself. These biases could occur due to specific optimization functions or the use of biased estimators (Baeza-Yates, 2018). Lastly, *historical biases* occur if existing trends or biases from the real world are transferred to the AI-based algorithm. If a specific group is underrepresented in a specific profession, an AI-based algorithm could interpret this data to omit this group in the selection process (Suresh & Guttag, 2021). This bias could also refer to the generation of the test labels that could be biased in themselves. Lakkaraju et al. (2017) referred to this specific problem as selective labels problem. To create a truly unbiased AI, one would need objectively true labels which are often not available. Hence, the utilized labels contain the biases and prejudices of a human person.

To approach fairness in AI, different methods exist in literature: *Fairness through unawareness* states that fairness can be achieved by depriving the AI of protected attributes. Nonetheless, this approach fails if there is any correlation between protected and unprotected attributes (M. Teodorescu et al., 2021). According to *demographic parity*, every identified group should make up an equal proportion in their occurrence in the selected population. However, this measure assumes every group to be equally suitable, and if there is indeed a difference between any two groups, the group with the better preconditions would be at a disadvantage compared to their actual suitability (Pessach & Shmueli, 2022). This problem is addressed by the measures *equalized odds* and *equal opportunity*, that calculate the rates that candidates were selected correctly and incorrectly for every group (Hardt et al., 2016). Hardt et al. (2016) emphasized that the latter is often easier to achieve computationally, but it represents a weaker fairness criterion.

The problem is, again, that for these measures to work, ground-truth labels are needed which are often unavailable (Lakkaraju et al., 2017). While all these measures refer to the concept of group fairness, it is important to consider individual fairness, that is, treating similar candidates equally. However, this raises the issue of creating a proper method to assess candidates' similarity because the design of such similarity measure could be biased itself depending on which properties are considered for the comparison (Fleisher, 2021). In addition to the difficulties of assessing fairness, increased fairness is frequently accompanied with a decrease of the AI's accuracy (Pessach & Shmueli, 2022).

Overall, however, existing research on fairness and AI mainly focus on a rather philosophical metalevel (e.g., Teodorescu et al., 2021) or on single criteria or specific cases (e.g., Mirbabaie et al., 2021). An example of the latter is a study focusing on perceived fairness of AI-based algorithms that are in charge of management tasks compared to a human counterpart (Lee 2018). However, this can hardly by transferred to the hiring context in which a human decision maker is supported by an AI-based algorithm.

A literature review can be used to identify problems and highlight the relevance of addressing them (vom Brocke et al., 2015). It would therefore be highly valuable for research and practice to have a structured review of the existing literature on fairness in AI in hiring at the algorithm level, to guide future research and practice on AI-based algorithms in hiring from a more technical than a philosophical perspective. Consequently, we analyze papers proposing specialized methods which we synthesize into a categorical structure to uncover approaches to address fairness in AI-based hiring.

3. Methods

In order to understand which approaches can be used to improve fairness in AI-based algorithms in hiring while maintaining AI's accuracy, a systematic literature review (SLR) was conducted as literature reviews are essential for any type of research (Webster & Watson, 2002). For structuring our SLR, we followed the recommendations by vom Brocke et al. (2015) and determined a process, specific sources, the coverage and the selected techniques. Our aim was to gain knowledge from existing literature on the topic that describes different aspects of the design or the data for machine-learning algorithms and other AI-based algorithms with a specific focus on the recruiting context. For this, we chose a sequential process (vom Brocke et al., 2015).

As sources for the literature search, we selected the online databases Scopus, IEEE Xplore, ACM Digital Library, and AIS eLibrary. With the selected databases, we aimed to cover the most important works on fairness in AI-based in hiring across disciplines. The final selection of papers should act as a representative selection (vom Brocke et al., 2015) to highlight the current state of fairness-sensitive AI algorithms in the field of recruiting.

Based on the comprehensive databases and the aimed representative coverage, all the literature that was considered in this review was obtained via keyword search within the above-mentioned databases and an additional backwards search. These steps were also adopted from vom Brocke et al. (2015).

After an initial search that was used for scanning popular literature to get a sense of the important aspects and to check the terminology that is used in this field, we collected suitable search parameters. We conducted different test searches and repeated this process until the results mostly seemed to fit our criteria while still being broad enough to include suited papers that do not include very specific keywords. As inclusion criteria, we specified that (1) a method had to be described to improve the fairness of AI-based algorithms and related ML algorithms and (2) the paper should have a specific focus on the HR topic area (if an AI-based system was tested on a dataset, it should include specific HR data). Accordingly, we also excluded all papers in which hiring was only mentioned as one of several application areas.

We created a search string that contained one keyword from each of the three areas of AI-based algorithms, fairness, and recruiting. We restricted the results to articles containing the keywords "feature*" or "classifier*" in the full text. The final search string that we used for this research was:

ABS (("artificial intelligence" OR algorithm OR "machine learning") AND (fair* OR discriminat* OR ethics) AND (hiring OR recruiting OR "Human Resource*")) AND ALL (feature* OR classifier*)*

Other terms such as HRM, AI or further synonyms were not used, because our initial search had already indicated that relevant articles always covered at least one other keyword from the search string. The search string was modified to fit the respective search options that the databases offer. For AIS eLibrary, our query did not yield any results. Thus, we excluded this source from the subsequent description. Across the three remaining databases, the search resulted in 101 results. After excluding the duplicate papers that occurred in the results of more than one database, we worked with a total of 84 different papers.

In the next step, the search results were filtered to exclude papers that did not match our inclusion and exclusion criteria. Therefore, two independent coders rated the results regarding fairness being further discussed in the paper according to an abstract check. We excluded articles where both coders decided that they were not relevant to the research question. Among these papers were some that only used the recruiting context as an example for AI-based algorithms, some that claimed to only give an ethical view of the subject, and some that had an entirely different topic. After this step, 45 papers could be selected.

These 45 papers were then examined for further criteria that were relevant as context for this literature review: The papers should describe the architecture or other specifics that affect AI-based algorithms and were introduced to improve some fairness metrics in the full text. Additionally, the papers should have a specific focus on the recruiting context. Especially the second criteria resulted in the removal of additional articles from the selection.

After this selection step, nine fundamental papers remained and were examined and analyzed. The entire selection process including the number of papers that were considered in each step is depicted in Figure 1.

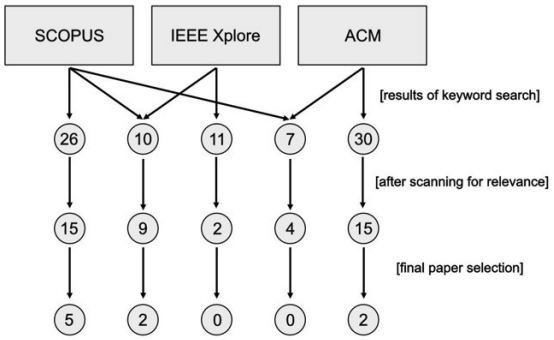

Figure 1: Number of articles at each step in the systematic selection process.

4. Findings

One task of a literature review is balancing the amount of time with an appropriate level of coverage (vom Brocke et al., 2015). We reduced the initial number of 101 articles to nine representative works and classified these nine articles, which were left after systematically filtering, into four approaches, namely pre-process, in-process, post-process, and feature selection. The classification was based on a screening of the title and the abstract of each paper. We draw the process-oriented approaches from extant literature where they were commonly used to categorize techniques that address fairness (Bellamy et al., 2018; D'Alessandro et al., 2017): *Pre-processing* alters the data before being processed by the AI-based algorithm. This method either removes the protected attributes in the dataset completely or just reweighs or modifies them to mitigate bias or prejudice (Mujtaba & Mahapatra, 2019). *In-processing* refers to an approach, where the AI-based algorithm itself is altered to fit fairness standards (Biswas & Rajan, 2020). Details on the modification depend on the concrete algorithm that is changed. *Post-processing*—as the name implies—denotes techniques that intervene after the AI-based algorithm has already processed the data. They are meant to achieve fairness through correction of the prediction results. For the most part, those algorithms change the odds at which certain groups are favored or disfavored by the AI and replace results to even out those biases (Bellamy et al., 2018).

Table 1. Final Papers and approach of fairness.

Articles	Approaches			
	Pre-Process	In-Process	Post-Process	Feature Selection
Booth et al. (2021)	X			X
Burke et al. (2021)			X	
Geyik et al. (2019)			X	
Goretzko and Israel (2022)				X
Hemamou et al. (2021)		X		X
Kappen and Naber (2021)				X
Mehrotra and Celis (2021)	X			
Mujtaba and Mahapatra (2019)	X	X	X	
Pessach and Shmueli (2021)	X	X		

We further derived *feature selection* from the pre-processing approach. The reason we consider this a separate approach is to increase the ability to discriminate between the manipulation of features and their simple exclusion. Thus, feature selection addresses fairness by excluding protected attributes from the data used to train an AI-based algorithm. Thereby, the AI-based algorithm itself is not modified, instead it is not provided with information that can be used to discriminate applicants (Grgic-Hlača et al., 2016).

Table 1 shows the categorization of the articles to the approaches for fairness in AI-algorithmic hiring.

Pre-Process. The articles with focus on this approach used additional algorithms to reweigh protected attributes. In this regard, Booth et al. (2021) applied normalization for each protected group separately, that is, feature values among each men and women are normalized. Similarly, Pessach and Shmueli (2021) proposed a method that aligns the distribution of values between the privileged and unprivileged group to eliminate across-group differences. Even though their methods increased fairness substantially, the fairness came at high costs in terms of accuracy. Mehrotra and Celis (2021), on the other hand, argued that the fairness of an AI-based algorithm decreases due to the noise contained in real-world data. To address this problem, they suggest a model to denoise protected attributes in order to mitigate gender and ethnicity bias. Mujtaba and Mahapatra (2019) did not develop a specific approach. Instead, they offered definitions for different notions of fairness (e.g., demographic parity, accuracy parity, predictive rate parity, individual fairness, and counterfactual fairness). Further, they outlined existing tools that are useful for developers aiming to design fair AIs.

In-Process. Hemamou et al. (2021) built a neural network to infer hireability from monologue videos. They argue that latent representations of the protected attributes gender and ethnicity exist in the model. In order to address those, they designed an adversarial network to mitigate the gender and ethnicity bias. Although they found statistical evidence that their approach increases fairness, it reduces the model's accuracy. Pessach and Shmueli (2022) modified logistic regression and a neural network by adding regularization terms for fairness. These regularization terms address fairness by penalizing the use of protected attributes. For both approaches a decrease in accuracy came along with the increase of fairness.

Post-Process. Burke et al. (2021) built their approach upon a ranking procedure. Their approach re-ranked the already processed data to receive candidates from protected groups as well. Geyik et al. (2019) addressed the ranking problem, too. To mitigate the gender bias, they combine the high ranked candidates

with a gender distribution over the qualified candidates to obtain a new and fair ranking. Simulations revealed that the approach did not harm utility and performance.

Feature Selection. Booth et al. (2021) deleted features that contain information about the candidates' gender which, in turn, led to an substantial reduction of the bias (greater reduction than their pre-process approach). Even though the work by Goretzko and Israel (2022) was of a theoretical nature, they emphasized essential considerations regarding the design of fair AI: It is essential to clearly define the criteria for the feature selection process and that the selection must not include any protected attributes or attributes that are related to or correlated with them. For instance, the zip code can be related to ethnicity in some locations. Kappen and Naber (2021) on the other hand evaluated their approach of reducing bias in the algorithmic assessment of candidates' motivation during job interviews. They demonstrated that the use of training labels which the candidates assigned themselves lead to better performance than the assessment of professional raters. At first glance, this might not seem to approach fairness through a change of feature, but, looking at it from a more conceptual level, it is: The decision to use a dataset gained through self-assessment instead of using one, which is based on professional raters, is not a major alteration to the feature, nevertheless it still changes the foundation on which the algorithm works.

5. Discussion

Prior research increasingly addressed the need for fairness in AI-based algorithms in hiring (M. Teodorescu et al., 2021), however, there has been insufficient attention to how fair AI can be applied and implemented in algorithms. We therefore provide an overview of technical approaches to implement and further examine fairness in AI-based algorithms in hiring while maintaining accuracy.

As the first category that we identified in our SLR, in-process approaches need to be considered more critically, which was addressed by Hemamou et al. (2021) and Pessach and Shmueli (2021) apart from the solely conceptual work of Mujtaba and Mahapatra (2019). This should not be considered surprising or worrying at all, since these approaches draw on the algorithm itself and even change it in fundamental ways. The research focused on this approach thus tends to have a larger scope than this specific topic of recruiting. It is to be noted that the adversarial method used by Hemamou et al. (2021) is used quite commonly in similar research (e.g., Zhang & Zhang, 2018). Here, the ML algorithm also tries to predict the protected attribute, which is supposed to be as inaccurate as possible. That way, ML algorithms cannot treat a group unfairly if it does not know how to predict a group.

Taking a look at the findings concerning post-process approaches, we see that a large part of the process is identical to standard algorithms. As it is part of the definition of post-process methods, the metric that determines how suited a candidate is, is calculated as it would be in the case of an application that is not sensitive to fairness. Subsequently, the AI-based algorithm's output is modified to match the fairness criteria defined previously (Burke et al., 2021; Mujtaba & Mahapatra, 2019). Both papers that conducted a practical utilization of their proposed method aimed at selecting an equal or proportional number of candidates from both the protected and the unprotected group in the final selection (Burke et al., 2021; Geyik et al., 2019). This fully satisfies the fairness metric of demographic parity (Pessach & Shmueli, 2022)**,** which comes with its own set of problems as we already mentioned previously. To overcome these problems the actual features and possible biases would have to be considered during the process of elimination bias, which can be accomplished by using pre-process approaches. But even in this case, most methods rely on demographic parity as a fairness metric, because ground-truth labels that would be needed for the equalized odds metric are simply not available in most cases (Lakkaraju et al., 2017). The rating of suitable candidates without any form of bias cannot be achieved by a human being to give reliable labels for the algorithm to work with. The main difference between pre-processing and post-processing approaches is the procedure to reach demographic parity. While post-process approaches often try to force demographic parity on existing computations, pre-process approaches can try to remove the biases themselves from the data. Booth et al. (2021) employ a normalization on the data sets of protected and unprotected groups, which can reduce differences concerning the relevant features. These differences could be caused by measurement or historical biases, which could be removed from the calculations, but the results of the study indicated that this approach could not remove bias efficiently.

The pre-process approach used by Mehrotra and Celis (2021) actually does not improve selection fairness in itself. Instead, they focus on denoising the protected attributes if they have to be derived from other attributes. These attributes are much needed for methods based on fairness through awareness. How large the impact can be if the protected attributes are not given with high accuracy is shown by a study by Ramezanzadehmoghadam et al. (2021). Here, an algorithm could not produce reliable results for a minority group since it had problems processing their names. Finally, Hemamou et al. (2018) used methods to

transform data contained in a video to not imply protected information that could also be used without the applied in-process method. In this case, the approach would show similarities to the concept of fairness through unawareness, where the protected information is removed from the calculation.

Even prior to the pre-process approaches is the step of feature selection, which could normally be considered as more of a conceptual step than a technical one. The general aim is to select features that do not contain any form of bias. The work of Kappen and Naber (2021) shows how a different elicitation of the same variable can have a huge impact on the fairness of the model that is trained by this data. Taking self-reported values over those of potentially biased observers is a good example of mitigating the measurement bias. A careful selection and the consideration of potential sources of biases is essential and easy to implement towards fairer AI decisions. Because of possible correlations between protected and unprotected attributes, it is insufficient to discard the protected attributes as features for the AI-algorithm (M. Teodorescu et al., 2021). Only including the features with the smallest weights in a model predicting the protected attribute showed a significant effect on the fairness of the final ML algorithm (Booth et al., 2021).

The above-mentioned approaches should not be praised as a solution without downsides. One main problem is the tradeoff between fairness and accuracy that was observed by most of the discussed studies (Pessach & Shmueli, 2022). An unmodified algorithm is aimed to reach the best possible performance when trying to predict the labels of the test set. Any further change to the model would result in a deviation from this best solution. Biswas and Rajan (2020) elaborate on this relationship between fairness and accuracy: Generally, all post-process approaches can be expected to have noticeable losses in their performance, while different in-process approaches can lead to widely unpredictable and different results. However, there is a larger problem with the accuracy metric itself in the context of fair AI algorithms. Since accuracy and similar performance metrics rely on the accordance of the prediction with ground-truth labels to allow making a statement on the utility or correctness of the algorithm. If the labels themselves are biased in any form, a solely accuracy-based decision would be suboptimal since it would state the biased view as the perfect decision that the ML algorithm aims to (Kilbertus et al., 2020). An approach to mitigate this problem is presented by Schoeffer et al. (2021). Here, the algorithm is not trained to predict any labels but to make an unbiased decision. The suitability of candidates is defined by the value of certain features that have to have a linear relationship with the suitability. The assessed labels are used as guidance for the weights of the features but not as a prediction for acceptance. Having discussed this potential downside of algorithmic fairness, most of the examined approaches to building fairer algorithms make a large step forward towards a more common usage of AI in the field of human resources. When using any form of automated decision-making, some form of protection from discrimination should be added to the model. Depending on the specific requirements of the recruiting decision, different types of improvements could be used. If the different groups of applicants are expected to not have general disparities and are potentially equally suited, the most effective method with the least effort could be post-processing approaches like the Rooney Rule. These methods guarantee a defined level of demographic parity in a candidate selection while potentially having a positive impact on the actual utility (Raghavan et al., 2020). But even if the field of applicants is not homogenous, there are feasible methods that lead towards fairer ML algorithmic decisions while not overcompensating for a protected group. These could be the adversarial in-process approach by Hemamou et al. (2021) or the pre-process approach with a focus on the selected features by Booth et al. (2021).

Frequent criticisms on SLR concern the very specific and delimited focus on a topic (Boell & Cecez-Kecmanovic, 2015). Nonetheless, our goal was to study the particular intersection of fairness, hiring, and AI, making SLR a suitable method in this regard. Nonetheless, our research is limited by the amount of literature on the topic, implying that our conclusions and research agenda are drawn upon a small sample.

6. Research Agenda

On a technical level, future research could close the gaps that occur in the landscape of fair AI-based algorithms in hiring. To encourage and guide further interdisciplinary research (e.g., in Computer Science or Information Systems) on implementing fairness in AI-based algorithms in recruiting, we propose a research agenda based on the results of our SLR. We suggest a general approach and four more specific approaches to achieve fairness in AI-based algorithms in hiring and suggest exemplary research questions:

General approach (Booth et al., 2021; Geyik et al., 2019; Goretzko & Israel, 2022; Nadeem et al., 2021)**:** Gender seems to be the most prevalent attribute used in fairness analysis, followed by ethnicity. However, other attributes are relevant too (e.g., maternity status, age, religion, sexual orientation, etc.). Hence, there is the need for further studies.

Possible research questions*:(1) When implementing an AI-based algorithm, which attributes need to be*

considered to ensure fairness? (2) How do attributes such as age, religion, sexual orientation, and other characteristics need to be treated by an AI-based algorithm to be considered fair?

Pre-process (Booth et al., 2021; Pessach & Shmueli, 2022): Methods like normalization and rescaling per protected group is likely to be problematic since they do not scale very well. Thus, existing approaches are hardly extensible and require a re-computation of the features whenever a new type of bias is included, or an additional attribute is considered. Further issues arise when one of an individual's attribute has to be normalized or rescaled according to the individual's membership in multiple protected groups.

***Possible research question**: How could an extensible pre-process approach be designed to ensure fairness in datasets for AI-algorithms in hiring?*

In-process (Hemamou et al., 2021; Pessach & Shmueli, 2022): We noted that approaches of this nature are significantly more complex than up- or downstream approaches as one needs to adapt an AI-algorithm. However, if an adaption is possible, then in-process approaches offer possibilities to increase fairness and maintain the performance simultaneously.

***Possible research question:** How can an existing AI-algorithm for hiring be modified to account for fairness in the hiring processes?*

Post-process (Geyik et al., 2019): The approaches belonging to this category focused on ranking tasks aiming to correct the order with respect to fairness. Since these approaches are applied after the main AI-algorithm finished its ranking, the re-ranking or correction does not affect the AI. Thus, they are not limited to any constraints given by the AI-algorithm. However, they pose a deviation from the algorithm's best ranking and consequently, they forfeit accuracy to some degree.

***Possible research question:** How can AI-algorithms outputs be corrected towards more fairness and diversity while maintaining accuracy?*

Feature selection (Booth et al., 2021; Goretzko & Israel, 2022): Removing information about applicants' protected attributes can led to a bias reduction. It is essential to define concise criteria by which features should be selected: Special attention needs to be paid when there are associations or correlations between seemingly non-protected and protected attributes. For example, the zip code can indicate ethnicity in some areas.

***Possible Research questions:** (1) What are attributes in hiring that necessarily need to be removed for feature selection of fair AI-algorithms? (2) What are criteria to assess whether an attribute needs to be removed to achieve fair AI-algorithms in recruiting?*

In addition to this agenda for future research, we aim to encourage practitioners to consider fairness in the development of AI for hiring purposes. This does not require innovative breakthrough techniques, but for instance questioning the fairness of a given dataset. An organization can evaluate whether the accuracy of an AI trained with protected attributes (e.g., gender) performs better in terms of effectiveness than an AI trained without protected attributes. Similarly, the correlation between protected attributes with the target variable can be analyzed. In regard to the actual implementation, we would like to refer to general existing methods such as IBM's AI Fairness 360 toolkit[2] or Google's Responsible AI Principles and Practices[3]. If an organization already uses AI-algorithms in hiring, we recommend checking whether their algorithms can be considered to be fair. For example, they can employ *Local Interpretable Model-agnostic Explanations* (Ribeiro et al., 2016) to examine the AI's decision process. Such tracing might reveal an unwanted behavior. In addition, future research needs to further guide algorithmic fairness by focusing on different sub-facets such as bias which could be a starting point for future literature reviews.

7. Conclusion

In this study, we contributed to fairness in AI-based algorithms in hiring by identifying and evaluating suitable approaches in this regard. The outcomes of our systematic literature review are put in four categories, namely pre-process, in-process, post-process, and feature selection. Our study revealed general as well as specific techniques to mitigate the unfairness of AIs in the hiring processes. We identified research gaps and corresponding questions which can guide scholars in their research. Overall, researchers need to determine further attributes relevant to fairness in AI-based hiring. Our findings on the pre-process approaches reveal the need for further research especially with respect to multiple, overlapping protected groups. Through an in-process lens, it occurs that Computer Science research is able to modify AI-based algorithms inherently towards increased fairness. From a post-process point of view, Information Systems research can bridge the gap between organizational research and technology to address the performance-fairness-trade-off (increase in fairness reduces performances and vice versa). Besides that, we made suggestions for practitioners to design and development fair AI-based algorithms. To begin with, an evaluation of the current degree of unfairness of the employed AI is needed to plan subsequent steps to

[2] https://developer.ibm.com/open/projects/ai-fairness-360/

[3] https://ai.google/responsibilities/responsible-ai-practices

increase fairness. For this, the four identified categories of pre-processing, in-processing, post-processing, and feature selection can be processed systematically. In this endeavor, developers should refrain from relying on one approach only since it became evident that each comes with its own drawbacks (e.g., limited extensibility or correlating attributes). However, building fair AI-based algorithms is only one side of the coin; biases are systematic in the way how people and the broader system use an AI-based algorithm. A recruiter can still reject a minority candidate deliberately even if the system presents a candidate unbiasedly. Thus, Organizational and Social Science are called upon to further examine fairness in hiring.

8. References

Alder, G. S., & Gilbert, J. (2006). Achieving Ethics and Fairness in Hiring: Going Beyond the Law. *Journal of Business Ethics 2006 68:4*, *68*(4), 449–464.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., et al. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. In *arXiv preprint arXiv:1810.01943*.

Bendick, M., & Nunes, A. P. (2012). Developing the Research Basis for Controlling Bias in Hiring. *Journal of Social Issues*, *68*(2), 238–262.

Berente, N.; Gu, B.; Recker, J.; Santhanam, R. (2021). Managing Artifical Intelligence. *MIS Quarterly*, *45*(3), 1433--1450.

Biswas, S., & Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. *ESEC/FSE 2020*, 642–653.

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, *63*(2), 215–226.

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'Systematic' in Literature Reviews in IS. *Journal of Information Technology*, *30*(2), 161–173.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. *ICMI 2021 - Proceedings of the 2021 International Conference on Multimodal Interaction*, 268–277.

Burke, I., Burke, R., & Kuljanin, G. (2021). Fair candidate ranking with spatial partitioning: Lessons from the SIOP ML competition. *Proceedings of the Workshop on Recommender Systems for Human Resources*.

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: why? how? what to do? *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 429–440.

Chan, J., & Wang, J. (2018). Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, *64*(7), 2973–2994.

Consul, N., Strax, R., DeBenedectis, C. M., & Kagetsu, N. J. (2021). Mitigating Unconscious Bias in Recruitment and Hiring. *Journal of the American College of Radiology*, *18*(6), 769–773.

D'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data*, *5*(2), 120–134.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, *32*(4), 754–818.

Fernández-Martínez, C., & Fernández, A. (2020). AI and recruiting software: Ethical and legal implications. *Paladyn, Journal of Behavioral Robotics*, *11*(1), 199–216.

Fleisher, W. (2021). What's Fair about Individual Fairness? *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 480–490.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2221–2231.

Goretzko, D., & Israel, L. S. F. (2022). Pitfalls of Machine Learning-Based Personnel Selection: Fairness, Transparency, and Data Quality. In *Journal of Personnel Psychology* (Vol. 21, Issue 1, pp. 37–47). Hogrefe Publishing GmbH.

Grgic-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS symposium on machine learning and the law* (p. 2).

Habous, A., & El Habib, N. (2021). Combining Word Embeddings and Deep Neural Networks for Job Offers and Resumes Classification in IT Recruitment Domain. *International Journal of Advanced Computer Science and Applications*, *12*(7), 651–658.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Hardy, J. H., Tey, K. S., Cyrus-Lai, W., Martell, R. F., Olstad, A., & Uhlmann, E. L. (2022). Bias in Context: Small Biases in Hiring Evaluations Have Big Consequences. *Journal of Management*, *48*(3), 657–692.

Harvie, K., Marshall-McCaskey, J., & Johnston, L. (1998). Gender-Based Biases in Occupational Hiring Decisions1. *Journal of Applied Social Psychology*, *28*(18), 1698–1711.

Hemamou, L., Guillon, A., Martin, J.-C., & Clavel, C. (2021). Don't Judge Me by My Face: An Indirect Adversarial Approach to Remove Sensitive Information From Multimodal Neural Representation in Asynchronous Job Video Interviews. *9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8.

Hofeditz, L., Mirbabaie, M., Luther, A., Mauth, R., & Rentemeister, I. (2022). Ethics Guidelines for Using

AI-based Algorithms in Recruiting: Learnings from a Systematic Literature Review. *HICSS*, Maui, Hawaii.

Hooker, B. (2005). Fairness. *Ethical Theory and Moral Practice*, *8*(4), 329–352.

Kappen, M., & Naber, M. (2021). Objective and bias-free measures of candidate motivation during job applications. *Scientific Reports*, *11*(1).

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., & Silva, R. (2020). The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. In R. P. Adams & V. Gogate (Eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference* (Vol. 115, pp. 616–626). PMLR.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The Selective Labels Problem. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Leavy, S. (2018). Gender bias in artificial intelligence. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16.

Liang, Chen; Hong, Yili; Gu, B. (2018). Home Bias in Hiring: Evidence from an Online Labor Market. *PACIS 2018 Proceedings*.

Marks, P. (2022). Algorithmic hiring needs a human face. *Communications of the ACM*, *65*(3), 17–19.

Mayer, Anne-Sophie; Strich, Franz; and Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quaterly Executive*, *19*(4), 6.

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), 1–35.

Mehrotra, A., & Celis, L. E. (2021). Mitigating bias in set selection with noisy protected attributes. *FAccT 2021*, 237–248.

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in Information Systems Research. *Communication of the Association for Information Systems*, *50*, 123–150.

Mirbabaie, M., Hofeditz, L., Frick, N. R. J., & Stieglitz, S. (2021). Artificial intelligence in hospitals: providing a status quo of ethical considerations in academia to guide future research. *AI & SOCIETY*, *Published online.*

Mujtaba, D. F., & Mahapatra, N. R. (2019). Ethical Considerations in AI-Based Recruitment. *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 1–7.

Nadeem, A., Marjanovic, O., & Abedin, B. (2021). Gender Bias in AI: Implications for Managerial Practices, *12896 LNCS*, 259–270.

Pager, D., Bonikowski, B., & Western, B. (2009). Discrimination in a Low-Wage Labor Market. *American Sociological Review*, *74*(5), 777–799.

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 1–44.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019). Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices. *SSRN Electronic Journal*.

Ramezanzadehmoghadam, M., Chi, H., Jones, E. L., & Chi, Z. (2021). Inherent Discriminability of BERT Towards Racial Minority Associated Data. In *Computational Science and Its Applications – ICCSA 2021*. Springer International Publishing.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. In *arXiv preprint arXiv: 1602.04938.*

Savage, D., & Bales, R. (2017). Video games in job interviews: using algorithms to minimize discrimination and unconscious bias. *ABA Journal of Labor & Employment Law*, *32*(2).

Schoeffer, J., Kuehl, N., & Valera, I. (2021). A Ranking Approach to Fair Classification. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 115–125.

Soleimani, M., Intezari, A., Taskin, N., & Pauleen, D. (2021). Cognitive biases in developing biased artificial intelligence recruitment system. *Proceedings of the Annual Hawaii International Conference on System Sciences* (HICSS), 5091–5099.

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *MIS Quarterly: Management Information Systems*, *45*(3), 1483–1499.

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

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, *37*, 205–224.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2).

Wolgast, S., Bäckström, M., & Björklund, F. (2017). Tools for fairness: Increased structure in the selection process reduces discrimination. *PLOS ONE*, *12*(12), e0189512.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A Fair Top-k Ranking Algorithm. *CIKM '17*, 1569–1578.

Zhang, N., & Zhang, N. (2018). Understanding the Roles of Challenge Security Demands , Psychological Resources in Information Security Policy Noncompliance. *Pacis 2018*.

Zhong, Z. (2018). A Tutorial on Fairness in Machine Learning. In *Towards Data Science*. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.