# Cognitive and Emotional Response to Fairness in AI – A Systematic Review

Janine Baleis[1], Birte Keller[1], Christopher Starke[1], Frank Marcinkowski[1]

[1] Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1
40225 Düsseldorf

janine.baleis@hhu.de | birte.keller@hhu.de christopher.starke@hhu.de | frank.marcinkowski@hhu.de

**Abstract.** Artificial intelligence is increasingly used to make decisions that can have a significant impact on people's lives. These decisions can disadvantage certain groups of individuals. A central question that follows is the feasibility of justice in AI applications. Therefore, it should be considered which demands such applications have to meet and where the transfer of social order to algorithmic contexts still needs to be overhauled. Previous research efforts in the context of discrimination come from different disciplines and shed light on problems from specific perspectives on the basis of various definitions. An interdisciplinary approach to this topic is still lacking, which is why it is considered sensible to systematically summarise research findings across disciplines in order to find parallels and combine common fairness requirements. This endeavour is the aim of this paper. As a result of the systematic review, it can be stated that the individual perception of fairness in AI applications is strongly context-dependent. In addition, transparency, trust and individual moral concepts demonstrably have an influence on the individual perception of fairness in AI applications. Within the interdisciplinary scientific discourse, fairness is conceptualized by various definitions, which is why there is no consensus on a uniform definition of fairness in the scientific literature to date.

**Keywords:** Algorithmic Fairness, Perception, Justice, Discrimination

## 1 Introduction

Since the last decade, digital applications have been following us more and more at every turn – especially through the possession of smartphones, algorithms have been able to penetrate our everyday lives while indirectly influencing our decisions (Logg, 2017). AI applications – in the form of algorithms[1] – are increasingly involved in pro-

[1] In the further course of this paper, the term algorithm is used, since algorithms often form the basis for AI systems and the scientific literature cited here uses the term algorithms in connection with artificial intelligence and fairness.

cesses that go beyond everyday decision-making processes (Shank et al., 2019; Srivastava, Krause & Heidari, 2019; Žliobaitė, 2017) – in lending (Petrasic et al., 2017; Lessmann et al., 2015), in the legal system (Christin, Rosenblat & boyd, 2015; Monahan & Skeem, 2016) or in medicine (Deo, 2015). Due to the penetration of all areas of life by algorithms, the discussion of the effects of such applications on the population becomes more and more relevant. Algorithms are increasingly being put in a position to shape society's life decisively and thereby influencing it negatively (Srivastava, Krause & Heidari, 2019; Woodruff et al., 2018). Discrimination against certain population groups or persons can be a consequence (Altman, Wood & Vayena, 2018; Dodge et al., 2019; Friedler et al., 2019; Žliobaitė, 2017), because according to scientific findings algorithms are not purely objective (Woodruff et al., 2018). To counteract this, it seems essential to develop systems that take fairness into account in decision-making (Khademi et al., 2019).

In recent years, the topic of fairness in algorithms has established itself as an important field of research and is being researched by various scientific disciplines such as ethics, social sciences, law and computer science. It has not yet been clearly clarified what fairness means in this context. While the different disciplines cite different views of the concept of fairness, there is no universally valid definition of fairness (Saxena et al., 2019). While mathematical-fair algorithms were developed in mathematics or economics, they do not necessarily meet the fairness criteria of other disciplines such as the social sciences (Lee, Kim & Lizarondo, 2017). But it is not only in science that the understanding of fairness in algorithms differs – among those who are affected by algorithms, the individual perception of what is perceived as fair varies greatly.

In order to capture the scientific and social findings in the context of fair algorithms, it is considered sensible to systematically summarise previous research results across disciplines in order to find parallels and combine common fairness requirements. This endeavour is the aim of this paper.

With the help of a systematic narrative literature review, the question of how individuals perceive fairness in decision-making algorithms will be answered. To answer this question, both theoretical and empirical contributions are identified and synthesized through systematic research.

The following study is structured as follows: At the beginning, the suitability and presentation of the method of systematic review are presented. Subsequently, the research objective and methodological conception of the review are described, which explains the derivation of the research question, the search strategy and the selection of the research units. In the following, the analysis process for the review is presented before the systematic narrative review follows. The summary of the results of the review is followed in the discussion section by a critical reflection on the quality of one's own approach and a conclusion on potential further research efforts.

## 2 Suitability and presentation of the method

The examination of the nature of algorithms is a topic of interest that is in its infancy – also in science. This development can be seen not least in the fact that, for example, the

submissions for the Association for Computing Machinery's conference on fairness, accountability and transparency in algorithms, in short ACM FAT, increased by 80% within one year (ACM FAT, 2019). Accordingly, the body of literature on this topic is in the process of further condensing and becoming multidisciplinary (e.g. Favaretto, De Clercq & Elger, 2019).

In order to be able to record the findings of scientific research in this context, it is necessary to systematically process them – especially against the background that this topic is prominently researched in many scientific disciplines. This can be done by means of a systematic literature review. In this way, a step can be taken to bring together the still existing heterogeneity of identified claims and to be able to formulate more reliable statements with the help of the research results (Petticrew & Roberts, 2006, p. 21).

Systematic records of existing literature serve to summarise previous results according to predefined selection rules and, in contrast to classical literature reviews, can support the accumulation of knowledge through transparent selection and analysis processes (Cooper & Hedges, 1994; Petticrew & Roberts, 2006). The aim is also to critically evaluate research results and extract relevant information to answer the question (Higgins & Green, 2008). Such a synthesis should also identify weaknesses in the evidence and highlight the need for further research (Booth, Sutton & Papaioannou, 2016, p. 11).

The approach of systematically recording research results became known above all through the international organization Cochrane Collaboration, which has set itself the goal of producing reviews for the evaluation of medical therapies (Booth, Sutton & Papaioannou, 2016, p. 303; Green et al., 2008, p. 3; Petticrew & Roberts, 2006, p. 19f.). For the preparation of a Cochrane Review, seven steps must be observed in order to ensure a comprehensible and at the same time reproducible procedure for the preparation of a review (Higgins & Green, 2008). First, a *precise question needs to* be worked out . Based on this, *criteria will be* defined on the basis of which relevant studies will be included or excluded for the review. This will be followed by a *systematic literature search in at* least two subject databases and a manual search in journal issues and conference contributions – "grey literature"[2] and foreign-language contributions will also be considered. The *choice of literature* for the review should be based on the title, abstract and full text of an article. Both the selection of the literature and the *assessment of the quality of the contributions should be* carried out by at least two independent reviewers (Higgins & Green, 2008).

Not only in medical research have organisations with guidelines for the production of reviews excelled (Booth, Sutton & Papaioannou, 2016; Moher et al., 2009), but also in the social sciences (e.g. Gough, Oliver & Thomas, 2012; Petticrew & Roberts, 2006) and in computer science (e.g. Kitchenham & Charters, 2007) methods have been developed in this respect.

[2] This refers to literature that has not yet been published (Lefebvre, Manheimer & Glanville, 2008, p. 106) and includes conference papers and working papers in this paper.

Thus, in addition to the Cochrane system, the present paper also takes into account the approach of Petticrew and Roberts (2006), since the following review is set in a social science context.

Petticrew and Roberts (2006, p. 27) follow a similar approach as the Cochrane reviews – the focus is on the conception of systematic literature summaries in the social sciences – because these authors also recommend a 7-step guide for the preparation of a review. In a first step, the guidance comprises the definition of a *research question* or *hypothesis to be* answered by the review. In the following, *inclusion criteria for* studies are to be defined before a *comprehensive literature search* is carried out. Next, Petticrew and Roberts propose a *screening of the results* that classifies the contributions found according to relevance. A *critical evaluation* of the selected studies will then be undertaken before a *summary of the contents* and *dissemination of the results of* the review follows.

For this paper, various elements of the systematics presented were adopted and combined for the subsequent review. The methodological concept of the review is described in the following chapter.

## 3 Research objective and methodological conception of the review

The following literature review aims to identify relevant studies on the perception of fairness in relation to algorithms in order to (1) identify reasons and consequences of perceived unfairness/discrimination in algorithms, (2) uncover previous obstacles to fair algorithms, and (3) find potential solutions to this problem.

### 3.1 Derivation of the research question

The aim of this analysis is to give an overview of contributions to the perception of fairness in algorithms in different disciplines and to record in which research areas this connection was investigated and which methods were used.

In order to be able to develop a systematic review on this topic, it is essential to derive a research question (Booth, Sutton & Papaioannou, 2016, p. 13). The question should be formulated as precisely as possible, since it represents the benchmark of the review (Counsell, 1997, p. 381). In the beginning, three elements were used to develop a research question based on Ibrahim (2008) and presented by Booth, Sutton and Papaioannou (2016, p. 84ff.). Part of these elements is the subject or object of the question (who), the object of the question (what) and the influence on the subject/object (how).

The focus of the question should be the human perception of fairness in algorithms; the characteristics of the persons are irrelevant at first. All human beings can therefore be fundamentally regarded as subjects of the question. The object of the question is the perception of the persons, influenced by the fairness in the algorithm.

In a next step, the generated components of the research question were substantiated by the "**PICOC**" method (Petticrew & Roberts, 2006, p. 44).[3] The method offers a formal structure that makes it possible to break down the fundamental components of the question into their individual parts in more detail (Booth, Sutton & Papaioannou, 2016, p. 86). This approach can be used to identify potential search terms for subsequent searches in subject databases (Booth, Sutton & Papaioannou, 2016, p. 87). Along the authors' scheme it is necessary to clarify *population*, *intervention*, *comparison*, *outcome* as well as *context in* relation to the research project. For the analysis, the following result can be seen when transferring the method to the research project (see Table 1).

**Table 1.** Application of PICOC method according to Roberts & Petticrew (2006).

| | |
|---|---|
| population | All gender, all individuals |
| intervention | Fairness in Algorithms |
| comparison | none |
| outcome | Perception (of Fairness) |
| context | globally |

*The population* – the subjects of the question – is also defined as individuals whose age and gender do not have to be specified. The *intervention* represents fairness in and through algorithms, a *comparison* was not defined. The *outcome* should be the individual subjective fairness perception of the algorithmic decision-making process. The *context* – meant here as a country-specific reference – was not narrowed down for the following review. Thus, the following research question can be derived for the efforts of the present work:

*How do individuals perceive fairness in decision-making algorithms?*

Before the search for relevant literature within the framework of the derived research question can be carried out, an adequate search strategy has been developed, which is outlined in the following chapter.

### 3.2 Development of the search strategy

The definition of search terms to be used in the literature search is central to the development of a search strategy for a systematic review. The basis for this is initially provided by the terms of the previously derived question. The "**pearl-growing**" method was used to add further terms for the subsequent search. By identifying a relevant article – the *pearl* –further search terms or keywords can be identified on which the subsequent literature search is to be based. The identified article is then searched for terms relevant to the research project – for example, keywords used –which are extracted for further

[3] This method is an extension of the scheme by Guyatt et al (2008), which was developed as a tool for formulating clinical research questions.

research. Accordingly, *pearl* should already have close points of contact with the concept of interest (Booth, Sutton & Papaioannou, 2016, p. 115) and devote itself to the initial research question.

For the research of *pearl*, terms had to be derived from the research question in order to identify a contribution that addresses the connection between fairness and algorithms in relation to the perception of individuals. Thus, analyses are of interest that illuminate the context empirically.

For the present analysis, the article "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management" by Lee (2018) was identified as *pearl.* In it, Lee deals with the social perception and attitude towards algorithmic decision making processes in comparison to human decisions, with reference to the theory of fairness.

Due to the fact that the content of the article corresponds to the research efforts of the present review, the keywords already collected were extended by the "**PICOC**" method with those of the *pearl* (see Table 2).

**Table 2.** Collection of keywords.

| Keywords through "**PICOC"** | Keywords used in *pearl* |
|---|---|
| - fairness<br>- perception<br>- algorithmic decision-making | - Algorithmic management<br>- perception<br>- folk theory<br>- fairness<br>- trust<br>- emotion |

Following the extension of the list of search terms, the literature used by Lee (2018) was searched using the snowball system method to find possible contributions dealing with the topic of interest. **Snowballing** is used in literature research to find further sources based on an article, book or contribution that the author mentions in the bibliography. If suitable publications have been identified, they can be researched and, if necessary, used for one's own work (Booth, Sutton & Papaioannou, 2016, p. 121). In this way some contributions were identified for the present work which, like the *pearl*, are relevant for the subsequent review. As before, the contributions were searched for keywords and catchwords in order to extend the list of final keywords (see Table 3). The search terms identified now include terms related to algorithms (first component) on the one hand, and terms that touch on the theoretical concept of fairness as well as the objective of the study perception (second component) on the other.[4]

[4] In this approach, the search terms were deliberately not divided into three components in order not to approach the literature search too preconditionally from the very beginning.

**Table 3.** Presentation of keywords for literature searches.

| Component 1 | Component 2 |
|---|---|
| - big data | - trust |
| - digital data | - fair |
| - artificial intelligence | - just |
| - machine learning | - discriminat* |
| - algorithmic* | - perception |
| | - response |
| | - emotion |

Some of the search terms were changed in a last step by truncation in order to abbreviate them when searching in databases. The advantage of this is that, depending on the placement of the special character (*) after a certain syllable of the word, various final syllables are automatically added by the database in the search process (Booth, Sutton & Papaioannou, 2016, p. 116).[5]

In order to test the suitability of the search terms, they were entered into the *Web of Science* literature database for a first test. However, since the combination of the terms led to too many results (> 4,000), it seemed sensible to slim down the terms of the second component, since when reviewing the results of the first search, it was above all the terms *emotion* and *response in* combination with the terms of the first component that revealed irrelevant contributions. After these two terms were removed, *perception was* excluded because the number of results with the keyword *perception* in the title corrected to a very small number – an indication that the term is too specific. *Trust* was removed from the series of buzzwords that were supposed to represent the theory. This can be justified by the fact that, when reviewing the database results, contributions relating to trust did not provide a reference to fairness or justice and would therefore not serve to answer the research question.

The final list of search terms of the second component is therefore *fair**, *just** and *discrimina**. Both of the components should appear in combination in the title of the contributions for the following review in order to be taken into account for the analysis.

### 3.3 Selection of examination units

The selection of the contributions to be examined was made in several steps. In the following, the sub-steps of this systematic approach will be explained. To illustrate this, Figure 1 has been created to illustrate the individual steps. The basis for this presentation is the PRISMA flowchart by Moher et al. (2009), which has been adapted for this paper.

[5] In this way, for example, the term "fair" is searched for as well as the term "fair" or "fairness".

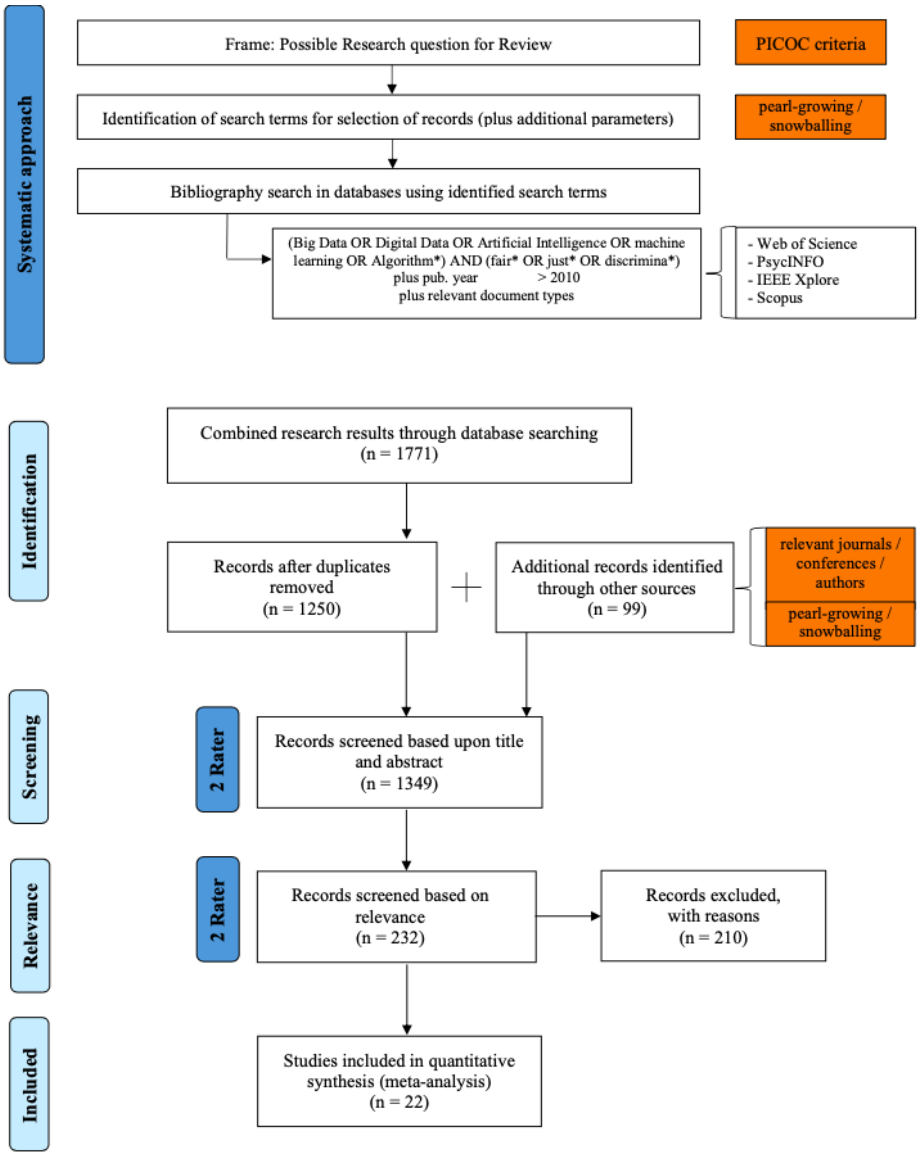


**Fig. 1.** Flowchart for the selection of the examination units according to Moher et al. (2009).

**Literature search in electronic databases.** For the selection of the research units, a search in various, partly interdisciplinary electronic literature databases – Web of Science, PsycINFO, IEEE Xplore and Scopus – was carried out using the search function, following the example of Petticrew and Roberts (2006) as well as Booth, Sutton and

Papaioannou (2016).[6] As fairness in algorithms is investigated in different research disciplines, as mentioned above, the selection of the subject databases was based on a rough thematic classification, which included the following disciplines (see Table 4).

**Table 4.** Overview of databases.

| database | scope[7] | fields of research |
|---|---|---|
| Web of science | about 12,000 journals<br>> 160,000 conference proceedings | Science, Social Science, Arts & Humanities |
| PsycINFO | About 2,500 journals | Psychology and Social Science |
| IEEE Xplore | > 190 journals | Computer Science, Electrical Engineering and Electronics |
| Scopus | > 23,000 journals | Science, Technology, Medicine, Social Science, Arts & Humanities |

With the help of the "**Boolean operators**" (Lefebvre, Mannheimer & Glanville, 2008, p. 132f.), the search terms of the two components were transformed into two combinatorial units:

> ("Big Data" *OR* "Digital Data" *OR* "Artificial Intelligence" *OR* "Machine Learning" *OR* "Algorithm*") **AND** (fair* *OR* just* *OR* discrimina*)

Similar terms for a concept are linked with an *OR* and in a further step the generated combinatorial units are linked with an *AND. In this way, it is* possible to combine a term of the first component with a term of the second component, and to apply the database to find results for all possible combinations (Booth, Sutton & Papaioannou, 2016, p. 118). After it was determined which terms the examination units to be considered must contain in the title, now the condition of the contributions is to be determined.

Although the majority of systematic reviews only include articles from peer review procedures and leave out other forms of contributions (Jungnickel, 2017, p. 45), the following review will include not only scientific articles from journals but also (proceeding) papers, as recent research efforts can be recorded in this way. Contributions from anthologies are also included in the analysis in order to derive possible further findings. Since the topic of fairness in algorithms is dealt with internationally and it can therefore be assumed that the scientific discourse is dealt with primarily in English, the present paper focuses exclusively on English-language publications. The methodological approach was not restricted – quantitative, qualitative and theoretical contributions were taken into account.

[6] The selection fell on databases with large stocks, which nevertheless cannot be assumed to be exhaustive. Therefore, some contributions that were not listed there may not be included in the analysis. To counteract this, additional manual research and snowballing were carried out.

[7] The information about the stock and the comprehensive scientific fields was taken from the homepages of the electronic databases.

As with Favaretto, De Clercq and Elger (2019), who conducted a systematic literature analysis on a related topic – discrimination in Big Data – publications have been included since 2010. This can be achieved by the renewed upswing of the AI topic. in the last ten years. In this phase, a new branch of AI research, Deep Learning, has emerged, which has aroused enormous interest in research in numerous disciplines (Adams Becker et al., 2017, p. 46; Kirste & Schürholz, 2019, p. 29).

**Identification.** According to the PRISMA flow chart, the step "Identification" follows, in which the contributions determined via the search function are documented, duplicates are removed and further contributions are added via other channels. The search terms and criteria in the various databases identified a total of 1,771 contributions as of the cut-off date (26 June 2019), of which 1,250 remained after a duplicate cleanup. In a further research step, 99 articles on the archives of relevant journals and specialist conferences, as well as on a search for authors and snowballing were found by hand. For the search in archives the identical search terms were used, as for the search in the literature databases. Snowballing has identified authors who have already published relevant articles on the topic itself or related topics.

**Screening and Relevance.** The following steps *screening* and *relevance* were carried out by two experts. For the *screening***,** the 1,349 remaining contributions were examined on the basis of their title and abstract. In concrete terms, the present articles were evaluated to determine whether they could provide useful information for the research interest. The selection criteria for the screening can be found in the chapter Codebook. In the course of this, deductive categories were formed. In case of contradictions between the reviewers, the contributions were transferred to the next coding process in order to extract possible further suitable studies from the cited literature of the contributions. In this way, 232 contributions were identified which were found to be relevant after an initial screening.

For the step *relevance*, the contributions determined were measured for relevance to answering the research question.[8] In order to obtain an answer to the derived research question, the experts, in consultation after a first reading, found such contributions which deal with the theoretical concepts mentioned above in connection with the perception of algorithms. The relevance of the abstract or article, the relevance to the derived research question of the review and the comprehensibility of the presentation of the research project and the method were assessed. These criteria were recorded numerically and added up for each contribution – if a criterion applies, the value one is assigned, so that a contribution can receive a maximum value of 4. This was converted into a color scheme after analog paper-and-pencil coding and then recorded in the codingsheet. As a result of this process, 22 contributions were selected for the systematic review and then excerpted extensively and divided into 4 content categories for further evaluation.

[8] If a contribution was not found as full text, it was excluded for further steps.

**Codebook.** A codebook has been prepared for the systematic analysis (see Annex A). This defines in separate chapters the *objective of the study*, the *period and units of investigation*, the *unit of analysis* as well as *formal criteria* and *selection criteria.* Furthermore, the *content categories for the* preparation of excerpts from the selected contributions and the *categorisation of the contributions to be included are* presented.

In the chapter *Selection Criteria*, access to the 1,771 recorded contributions was first removed. In addition, the conformity of the contributions was assessed – those that did not have an abstract and could not be found despite research were not taken into account for the rest of the process. Contributions whose title were in English but the abstract and the entire content were written in another language were not taken into account either. The *relevance of a contribution* in the preselection of contributions is given if at least one of the following criteria applies: comprehensibility of the abstract (rel_a), relevance to the derived research question of this review (rel_b), comprehensibility of the presentation of the research project (rel_c) and the methodology (rel_d). The relevance of the contributions from the preselection was coded in a next step with a colour scheme (green – at least 3 relevance criteria apply to the contribution and it is therefore relevant, yellow – at least 2 criteria can be transferred and the contribution is potentially relevant, turquoise – the contribution could rather become relevant in another context, as only one criterion is fulfilled, red – the contribution is not relevant and should not be included in the preselection).

The chapter *excerpts* contains the categories for the systematic recording of the relevant contributions of the final literature selection. The information was inserted into a ready-made Excel table as free texts. All information recorded in the excerpt was supported by indirect or direct quotations of the contribution to be excerpted, indicating the corresponding page number. In addition to the bibliographical information, the research discipline from which the contribution originated, the research question and the research project were also recorded. If mentioned, the used or recurred theory was given, as well as the method followed to answer the research question. In the case of empirical studies, both independent and dependent variables as well as the sample or selection of respondents should be recorded. Where mentioned, hypotheses and general assumptions were also documented. In addition, the results of the contributions were collected, which help to answer the research question of the review or, in addition, contribute relevant information for the facts of the case. Additional relevant information from the full text as well as literature references from the article were added. If a contribution contained appealing graphics, figures or tables that can be used sensibly from the point of view of the reviewer, these were also included as a screenshot in the excerpt document below under the item Appendix.

**Reliability.** For the selection of the contributions for the systematic review, all abstracts of the database results were examined with regard to relevance as a first step. The review of the abstracts was done by two raters. To ensure that this step corresponded to the selection standards of the research, a number of 72 abstracts were read by both raters in order to ascertain the reliability of the decision results and to statistically substantiate these results. Based on Higgins and Deeks (2008, p. 155), the consistency measure Kappa is calculated for the coherence of the coding results between the two raters. This

calculation provides for determining the percentage of matches in the total number of codings (p0) as well as the probability of random matches (pE) and normalizes this difference with the expected frequency of random non-compliance (Higgins & Deeks, 2008, p. 155; Hammann & Jördens, 2014, p. 177). A first test was performed based on a common coding process of 33 abstracts. Coding was done using the colors green (include – 2), yellow (potentially include – 2), turquoise (unclear – 1) and red (exclude – 0). The color green stands for a clear connection with the research question. Yellow, on the other hand, means that on the basis of the abstract it is not yet possible to conclusively determine whether an article is worth considering in the review. In this case, the entire contribution was used to make a final classification. For the calculation of Kappa, the color scheme was converted into numerical values. All in all, a Cohen's $\kappa = .55$ can be recorded for the first test. Thus, according to Higgins and Deeks (2008, p. 155), who refer to Orwin (1994), values up to .59 reflect "fair agreement".

In order to improve the quality of the agreement between the two raters, some examples were discussed after the first test in which the raters diverged. In the next step, 39 abstracts were read by both experts and subsequently coded. Table 5 shows the number of matches.

**Table 5.** Calculation of the kappa coefficient of agreement (second test).

| | | **Rater 2** | | | |
|---|---|---|---|---|---|
| | | excluded (0) | unclear (1) | included (2) | total |
| **Rater 1** | excluded (0) | 37 | 0 | 0 | 37 |
| | unclear (1) | 0 | 0 | 1 | 1 |
| | included (2) | 0 | 0 | 1 | 1 |
| | total | 37 | 0 | 2 | 39 |

With the second test a match of $\kappa = .74$ could be determined, considered to be a value of "good agreement" (Higgins & Deeks, 2008, p. 155). The percentage of agreement between the two experts is 97.44 percent – thus a very good reliability could be produced.

### 3.4 Analysis process

The analysis process of a systematic review can be performed in different ways: On the one hand through a meta-analysis, on the other hand it is possible to carry out a narrative literature analysis. For a meta-analysis, the empirical data from various studies are combined into a data set using statistical methods in order to generate further results and derive a global answer to the research question of the review from previous study results. A meta-analysis can be implemented if, for example, the design of the studies is similar or the same independent and dependent variables are used – the studies should therefore be as homogeneous as possible (Petticrew & Roberts, 2006, p. 164). A meta-

analysis is considered not suitable for this paper – the selected studies show too different methodological approaches and research questions. A systematic narrative review is therefore carried out.

According to Petticrew and Roberts (2006, p. 164), the implementation of a narrative systematic review is a common method in the social sciences. This method includes the summary of the study results and the description of the most important characteristics of this – e.g. the chosen method, the population studied as well as the details of the intervention. Furthermore, the methodological problems or bias of the studies are to be evaluated and possible effects on the study and review results are to be addressed (Petticrew & Roberts, 2006, p. 166). The authors propose a three-stage process for synthesis: (1) Division of the studies into logical categories (e.g. design, outcome, intervention, population, sample size, etc.), (2) narrative description of the individual study results, and (3) cross-study summary. The latter contains the number of studies analysed, the effects of the variables examined and the differences between the studies (Petticrew & Roberts, 2006, p. 166ff.).

The analysis process of the following systematic review was carried out in several steps. At the beginning, the 22 selected contributions were systematically extracted and then presented in a tabular overview comparing various characteristics. In order to prepare the narrative review, the contributions were categorised according to their suitability for answering the research question on the basis of the excerpts. This first step of the analysis process corresponds to the proposal of the first step by Petticrew and Roberts (2006, p. 166ff.).

For the systematic review, a distinction is first made between theoretical and empirical contributions before the relevant empirical study results are described. The results are described in groups. This step can be assigned to the second stage of the procedure after Petticrew and Roberts (2006). This is followed by a cross-study summary of the results.

**Excerpts.** The selected 22 contributions were systematically extracted parallel to the reading. Against the background of the heterogeneity in methodology and the research design of the contributions, a summary of the individual studies on excerpts appeared to be useful as preparation for the systematic analysis. For this purpose, a grid was created (see Appendix B) in which the respective information was inserted as free texts.

The excerpt contains information about the research discipline from which the contribution originates, the research question and the research project. Furthermore, the theoretical concept to which the contribution refers was recorded. The methodological approach, the study design – including information on the sample and the population studied – and hypotheses were also recorded. In order to be able to apply the contributions to the derived research question, the results of the articles and their theoretical connection were documented. Additional information, such as critical statements by the authors or the like, was recorded separately.[9]

[9] A detailed overview can be found in the Codebook (Annex A).

**Categorisation of contributions.** The 22 selected contributions were subdivided into content groups (see Annex A). This subdivision was intended on the one hand to determine the suitability of the contributions for answering the research question, and on the other hand to pre-structure the review.

The first group contains relevant contributions which, in addition to dealing with theory, also deal empirically with the recording of perceptions on algorithms and are thus useful for answering the research question. For example, the contribution of Lee (2018) was divided into this group along with 13 others.

The second group includes contributions that deal empirically but more technically with fairness or discrimination in algorithms. Although these contributions can be used to determine which technical conditions are advisable for fair algorithms, the gain in knowledge for answering the research question is smaller.

The third group includes such contributions that deal with theory and examine the operationalization of fairness-aware machine learning algorithms (3.1) and the operationalization of fairness in machine learning and social science literature (3.2).

## 4 Systematic Review

Particularly in the last two years, the topic of fairness in algorithms has increasingly been dealt with in various research disciplines. In order to gain a comprehensive overview of the contributions of the various disciplines that deal with the perception of individuals to fairness in algorithms, the systematic processing of these is to be classified as meaningful. The systematic review is also intended to contribute to insights into which concepts of fairness should be applied in algorithms in order to maximize the justice of decision from the point of view of individuals. In this way the practice of effective methods can be pointed out and at the same time the need for further research can be pointed out to the scientific community.

In order to be able to use the information from the excerpts for a higher-level analysis, it has been stored in an overview table. (see table 6). In addition to the research design and methodology, the discipline from which the contribution originated, the field of application, the definition of fairness used and the interest in knowledge were also recorded.

**Table 6.** Overview of included articles.

| **Author (Year) Country** | **design** | **discipline** | **„field of application“** | **Definition of fairness / Concept of fairness** | **Aim** |
|---|---|---|---|---|---|
| Lee (2018) USA | empirical | multidisciplinary | employment | “Fairness is defined as treating everyone equally or equitably based on people’s performance or needs“ (p. 4) | “We posit that how people perceive algorithmic and human decision-makers may influence their perceptions of the managerial decisions that are made.“ (p. 3) |
| Shin & Park (2019), UAE & USA | empirical | Computer Science | Algorithmic Services | “Fairness in algorithm contexts means that algorithmic decisions should not create discriminatory or unjust consequences“ (p. 278) | “this study aims to conceptualize FAT in relation to the increasing use of algorithms and clarifies the roles of such problems in the user acceptance of algorithm services.” |
| Woodruff et al. (2018) USA | empirical | Computer Science | social | Algorithmic discrimination | “explore ethical and pragmatic aspects of public perception of algorithmic fairness” (p. 1) |
| Binns et al. (2018) UK | empirical | Computer Science | Various | Informational, procedural & distributive justice | “We undertake three experimental studies examining people’s perceptions of justice in algorithmic decision-making under different scenarios and explanation styles.” (p. 1) |
| Vallejos et al (2017) UK | empirical | multidisciplinary | Digital citizenship | None named | „explores the policy recommendations made by young people regarding algorithm fairness“ (p. 247) |
| Koene et al. (2017) UK | empirical | Computer Science | Online Services | “a context-dependent evaluation of the algorithm processes and/or outcomes against socio-cultural values. Typical examples might include evaluating: the disparity between best and worst outcomes; the sum-total of outcomes; worst-case scenarios; everyone is treated/processed equally without prejudice or advantage due to taskirrelevant factors” (p. 1) | “we present two pilot studies aimed at getting a better understanding of the conceptualisation of algorithmic fairness by users.” (p. 1) |

| Author (Year) Country | design | discipline | „field of application“ | Definition of fairness / Concept of fairness | Aim |
|---|---|---|---|---|---|
| Grgić-Hlača et al. (2018a) Germany, UK & USA | empirical | Computer Science | law | none named | “we propose to understand why people perceive certain features as fair or unfair to be used in algorithms.” (p. 1) |
| Grgić-Hlača et al. (2018b) Germany & UK | empirical | Computer Science | law | procedural & distributive justice | “we propose notions of procedural (rather than distributive) fairness, based on which input features are used in the decision process and how including or excluding the features would affect outcomes.” (p. 1) |
| Lee, Kim & Lizarondo (2017) USA | empirical | Computer Science | algorithmic services | equality & equity | “we take a human-centered approach in order to identify considerations for building fair and motivating algorithmic services.” (S. 1) |
| Lee & Baykal (2017) USA | empirical | Computer Science | social | equality & equity | “We investigated people’s perceptions of mathematically-proven fair division algorithms making social division decisions.” (p. 1035) |
| Saxena et al. (2018) USA | empirical | multidisciplinary | salery | multiple (Computer Science) | “our goal is to understand how people perceive the fairness definitions proposed in the recent computer science literature” (p. 1) |
| Araujo et al. (2018) The Netherlands | empirical | social science | none | none named | “an overview of public knowledge, perceptions, hopes and concerns about the adoption of AI and ADM across different societal sectors in the Netherlands.” (p. 3) |
| Dodge et al. (2019) USA | empirical | Computer Science | law | discrimination | “we conducted an empirical study with four types of programmatically generated explanations to understand how they impact people’s fairness judgments of ML systems” (p. 275) |

| Author (Year) Country | design | discipline | „field of application“ | Definition of fairness / Concept of fairness | Aim |
|---|---|---|---|---|---|
| Srivastava, Heidari & Krause (2018) USA & Switzerland | empirical | Computer Science | law, medicine | multiple (Computer Science) | “We take a descriptive approach and set out to identify the notion of fairness that best captures lay people’s perception of fairness. We run adaptive experiments designed to pinpoint the most compatible notion of fairness with each participant’s choices through a small number of tests.” (p. 1) |
| Khademi et al. (2019) USA | empirical | Computer Science | salery, law | “We introduce two definitions of group fairness grounded in causality: fair on average causal effect (FACE), and fair on average causal effect on the treated (FACT).” (p. 2907) | “Our analyses of two real-world data sets, the Adult income data set from the UCI repository (with gender as the protected attribute), and the NYC Stop and Frisk data set (with race as the protected attribute), show that the evidence of discrimination obtained by FACE and FACT, or lack thereof, is often in agreement with the findings from other studies.” (p. 2907) |
| Altman, Wood & Vayena (2018) USA | empirical | Computer Science | none | “When analyzing fairness, one should measure all of the aspects of life that are widely recognized within social science and health fields as fundamental for well-being. Specifically, the literature identifies five key measures for a life course analysis: wealth, lifespan, health, subjective life-satisfaction, and the ability to make substantial choices about one’s life (sometimes referred to as “capability”).” (p. 6) | “We further demonstrate how counterfactual frameworks for causal inference developed in statistics and computer science can be used as the basis for defining and estimating the foreseeable effects of algorithmic decisions.” (p. 2) |

| Author (Year) Country | design | discipline | „field of application“ | Definition of fairness / Concept of fairness | Aim |
|---|---|---|---|---|---|
| Speicher et al. (2018) Germany, Switzerland & UK | empir-ical | Computer Science | social | group & individual fairness, inequality | “Given two unfair algorithms, how should we determine which of the two is more unfair? Our core idea is to use existing inequality indices from economics to measure how unequally the outcomes of an algorithm benefit different individuals or groups in a population.” (p. 2239) |
| Lepri et al. (2017) Italy & USA | theo-retical | philoso-phy | social | “lack of discrimination or bias in the decisions” (S. 5) | “Instead, focusing on discrimination and lack of transparency, we provide the readers with a review of recent attempts at making algorithmic decision-making more fair and accountable, highlighting the merits and the limitations of these approaches.” (p. 4) |
| Veale & Binns (2017) UK | theo-retical | multidis-ciplinary | none | discrimination | “This paper focuses on how fairness and discrimination in machine learning systems can be mitigated within practical institutional constraints.” (p. 2) |
| Binns (2018) UK | theo-retical | multidis-ciplinary | none | “Questions of discrimination, egalitarianism and justice” (p. 1) | “This paper draws on existing work in moral and political philosophy in order to elucidate emerging debates about fair machine learning.” (p. 1) |
| Gajane & Pechenizkiy (2018) Austria & The Netherlands | theo-retical | none | none | multiple (Computer Science) | “The aim of this article is to survey how fairness is *formalized* in the machine learning literature and present these formalizations with their corresponding notions from the social sciences literature.” (p. 1) |

Among the 22 contributions included for the review are 18 empirical and 4 theoretical contributions. The theoretical contributions deal with the operationalization of fairness in algorithms and in the scientific literature. Among the empirical contributions are empirical-technical contributions to the nature of fair algorithms (n=4) and empirical contributions to the perception of fairness in algorithms (n=14). Among the latter are three qualitative studies, seven quantitative and four with mixed-method design. None of the contributions were published before 2017. A large part of the analyses (n=8) comes from the US, five from Great Britain, one from the Netherlands and another from Finland. A further seven contributions come from an international team of authors. In terms of scientific disciplines, more than half of the papers (n=14) were published in journals or at computer science conferences, five others multidisciplinary and only two from the social sciences or philosophy. An included contribution has not yet been published

The concrete research project as well as the results of each contribution will be described in detail in the following chapters, divided according to the research design. Furthermore, the articles are grouped according to results and evaluated in order to answer the research question.

### 4.1 Theoretical contributions

In addition to empirical contributions, theoretical contributions were also used to obtain an answer to the research question of the review. Since the following theoretical articles do not deal with the perception of fairness in algorithms, but nevertheless provide insights into the use of fairness theory in algorithms, it was considered reasonable to include these contributions in the review. In addition, it seems sensible to consider the findings from theory for further empirical research.

Among the contributions that deal theoretically with the operationalization of fairness-aware machine learning algorithms are the investigations of Lepri et al. (2018) and Veale and Binns (2017).

Lepri et al. (2018) give in their article an overview of applicable technical solutions to promote fairness, accountability and transparency in algorithmic decision-making processes. From the literature presented, the authors conclude that the definition of fairness in an algorithm must be used individually and problem-centered. This task is also a challenge for many scientific disciplines, as the topic is already widely researched. Lepri et al. conclude with the appeal that multidisciplinary research teams are needed to ensure fairness and transparency in algorithms.

Veale and Binns (2017) outline three approaches for fairness-aware machine learning in organisations. In the authors' view, a lack of fairness is due to the fact that organisations are not allowed to collect sensitive data for legal or institutional reasons. A fairness enhancement approach involves the integration of trusted third parties who, unlike the organisation, could selectively store data to detect discrimination and implement fairness restrictions while respecting privacy. Collaborative online platforms will enable organisations to gather contextual and experiential knowledge from researchers, other practitioners and stakeholders in order to increase fairness. The third approach

exploratively deals with the formulation of fairness hypotheses and interpretable models to improve the fairness of a system.

It can be deduced from the two contributions that questions of fairness in algorithms should be examined contextually and from different scientific perspectives in order to ensure these and other factors such as transparency. It is also pointed out that attributes which may give rise to discrimination have not yet been legally regulated (Veale & Binns, 2017, p. 2).

Among the contributions that deal theoretically with the operationalization of fairness in machine learning (ML) and social science literature are the studies by Binns (2018) and Gajane and Pechenizkiy (2018) from Computer Science.

In his contribution Binns (2018) combines the fairness definitions from the basic literature of political philosophy with those used in ML literature and practice. The aim is to combine the findings of both disciplines and to identify in the philosophical literature those aspects which are helpful for future ML research on algorithmic fairness or which have not yet been considered. Binns points out that discrimination, egalitarianism and justice concepts are of particular interest when it comes to the mathematical definition of fairness for an algorithm. In detail, the author points out that fairness in ML literature, spoken in the language of political philosophy, stands for a multitude of normative egalitarian considerations.

Gajane and Pechenizkiy (2018) examine in their contribution similar to Binns (2018) how fairness is formalized in ML literature for predictions and contrast these findings with the ideas of distributive justice in social science literature. The authors point out that two concepts of fairness have not yet been discussed in ML literature, but have already been studied intensively in the social sciences: *Equality of resources* and *Equality of capability of functioning.* According to the authors, fair predictions by algorithms cannot be made without considering these social questions, although it is difficult to implement these attributes in algorithms.

It can be deduced from the two contributions that the concepts of fairness in the social sciences are related to those in computer science and are in part congruent. However, it is necessary for computer science and social sciences to think together in order to program fair algorithms. Above all, social issues, such as access to resources, must be integrated into algorithmic decision-making processes.

### 4.2 Empirical contributions

In addition to the different disciplines from which the contributions originate, they also differ with regard to the examination of the theory of fairness, the field of application of the investigation and the chosen methodology. Before these aspects are dealt with, the contributions that deal technically with fairness in algorithms should be distinguished from those that deal empirically with the perception of fairness in algorithms. The latter are of particular relevance for answering the research question, which is why these findings are described in detail and only brief references should be made to the more technical studies.

**Empirical-technical contributions to fairness in algorithms.** The four contributions, which deal with the technical integration of fairness in algorithms, all come from Computer Science and investigate different measures of fairness and discrimination in order to increase fairness in algorithmic decisions.

Khademi et al. (2019) investigate the discrimination potential of algorithms by considering causal models. The aim is to determine whether a group of individuals who share a sensitive attribute (e.g. gender) has been discriminated against by causality in an algorithmic decision-making system. The results show that fairness for the groups does not necessarily go hand in hand with fairness for the individual. Altman, Wood and Vayena (2018) investigate the effects of algorithmic decision making on the lives of individuals using counterfactual analysis. The results show that decisions are classified as unfair if certain individuals are confronted with foreseeable damages or if they have to bear higher costs than others. Based on the results, the authors also provide technical recommendations for achieving fairness in algorithms. Žliobaitė (2017) evaluates various measures of discrimination from Computer Science and evaluates their suitability for measuring discrimination in algorithms. In addition, the author highlights measurements from other disciplines that are not yet applicable, but are potentially suitable for the context, and discusses the need to extend the legal situation to algorithmic decisions in the context of non-discrimination. In this context, the author pleads for the use of the expertise of computer scientists for future legislation. Speicher et al. (2018) deal with measurements for algorithmic unfairness using inequality indices from economics. The results show that the unfairness of an individual-level algorithm can be divided into a between-group and a within-group component. According to the authors, the unfairness of an algorithm increases when the between-group component – i.e. the average benefit of the group – is minimized.

**Empirical contributions to perception.** In order to be able to shed more light on the findings of the 14 empirical papers on the perception of fairness in algorithms, the methodology of these algorithms will be examined in more detail beforehand. On the basis of the methodology, the results of the contributions can subsequently be compiled in groups.

Three contributions were carried out purely qualitatively – one by means of qualitative interviews (Lee, Kim & Lizarondo, 2017), the others by means of deliberative procedures such as group discussions (Vallejos et al., 2017; Woodruff et al., 2018). Seven contributions were implemented using quantitative methods, including four online experiments (Dodge et al., 2019; Lee, 2018; Saxena et al., 2019; Srivastava, Heidari & Krause, 2019), two online surveys by Grgić-Hlača et al. (2018a, 2018b) and a standardized survey using CAWI by Araujo et al. (2018). A combination of quantitative and qualitative elements can be found in four further works: Two studies show qualitative interviews in combination with offline and online surveys (Binns et al., 2019; Shin & Park, 2019), two others conduct experiments together with a stakeholder group discussion (Koene et al., 2017) or a qualitative laboratory study (Lee & Baykal, 2017).

*Contributions (qualitative).* The three selected contributions with qualitative methodology deal with the recording of individual fairness perceptions of particular target

groups. Lee, Kim & Lizarondo (2017) explore how service algorithms need to be programmed to make decisions that are not only efficient, but also fair and motivating. The authors investigate this question in the context of a food sharing project in which an algorithm will be used in future to decide to whom donations should be allocated. Stakeholder interviews will be used to determine the human considerations behind an allocation decision and how an algorithm can implement these requirements fairly. Vallejos et al. (2017) use deliberative methods to examine the political recommendations of young people on the nature of fairness in algorithms. The discussion should be stimulated by inputs on the influence of algorithms on Internet users and should trigger opinion-forming processes on the topic. Woodruff et al. (2018) investigate the perception of fairness in algorithms of marginalized population groups in the USA through group discussions and subsequent qualitative interviews, with the aim of examining the ethical and pragmatic aspects of public fairness perceptions.

*Contributions (quantitative)*. On the one hand, the contributions with experimental designs examine the question of which fairness definitions can be applied to the perception of individuals for the assessment of justice in an algorithm (Saxena et al., 2019; Srivastava, Heidari & Krause, 2019). Saxena et al. (2019) are interested in finding out which of the three fairness definitions presented appears fairest to respondents in the context of decisions on lending. Srivastava, Heidari and Krause (2019) proceed similarly; the aim of the experiments is to determine in which social context (e.g. creditworthiness) which mathematical concept of fairness is regarded as ethically more desirable.

On the other hand, other contributions examine the question of how different decision-makers (Lee, 2018) and explanatory approaches (Dodge et al., 2019) affect the individual's perception of fairness. Lee (2018) focuses on the analysis of the perception and evaluation of algorithmic versus human decision makers using the example of management decisions. Dodge et al. (2019) investigate how various programme-driven explanations affect the individual fairness assessments of ML systems.

Grgić-Hlača et al. (2018a) investigate why individuals consider the use of certain information (features) for algorithmic decision-making to be (un-)fair. In a further article Grgić-Hlača et al. (2018b), the evaluation of the use of certain information (features) in a decision-making process and the question as to which of these features can be used to fairly design a process are dealt with. Araujo et al. (2018) deal in their study on public perception and attitudes towards artificial intelligence and automated decision-making in the Netherlands, among other things, with the objectivity of algorithms.

The contributions using mixed methods designs examine the influence of fairness, accountability and transparency on individual attitudes towards algorithms (Shin & Park, 2019), and the influence of reasoning approaches for algorithmic decisions on the fairness evaluation of algorithmic decisions (Binns et al., 2018). On the other hand, the conceptualization of fair algorithms from the user's perspective (Koene et al., 2017) and the differences in perception in decision-making through algorithms and group discussions (Lee & Baykal, 2017) are examined.

*Results empirical contributions.* The results of the empirical contributions can be divided into four groups: On the one hand, general assessments on the topic of fairness in AI are given, on the other hand guidelines for the implementation of fairness in algorithms are derived from the results. Furthermore, the application of certain fairness definitions in algorithms will be discussed and influences for the perception of (in)fairness in algorithms will be investigated and uncovered.[10]

**Table 7.** Division of literature into content groups.

| | |
|---|---|
| General results on the topic | Koene et al. (2017), Lee (2018), Lee & Baykal (2017), Araujo et al. (2018) |
| Application of a specific definition of fairness in the algorithm | Saxena et al. (2019), Srivastava, Heidari & Krause (2019) |
| Guidelines for Fairness in KI | Lee, Kim & Lizarondo (2017), Vallejos et al. (2017), Woodruff et al. (2018), Shin & Park (2019) |
| Influence on fairness perception | Grgić-Hlača et al. (2018a, b), Saxena et al. (2019), Dodge et al. (2019), Shin & Park (2019), Binns et al. (2018), Lee & Baykal (2017) |

Three contributions derive more general results for fairness in algorithms. In dealing with public perception, knowledge, hopes and concerns about the integration of artificial intelligence and automated decision systems in various social sectors in the Netherlands, Araujo et al. (2018) in relation to fairness deduce that more than one third of respondents agree that artificial intelligence and automated decision systems lead to a more objective treatment of people. These technical applications can also lead to fairer decisions. Like Araujo et al. (2018), Lee and Baykal (2017) pointed out that the algorithm was considered fair from the respondents' point of view due to the perceived objectivity and equal treatment of all participants at the beginning. However, in a study with comparative elements, the authors found that one-third of respondents felt that algorithmic decision-making was less fair to group decisions because they did not consider multiple concepts of fairness. From the respondents' point of view, more room for compromise could be allowed in group decisions in order to increase the fairness of a decision. In a further analysis of fairness in algorithmic decision-making towards human decision-makers, Lee (2018) shows that respondents evaluate both human and algorithmic decisions equally fairly in relation to mechanical tasks; likewise, management decisions are equally just and trustingly independent of the decision-maker. However, when it comes to decisions that involve human qualities (e.g. empathy) - professional recruitment and evaluation tasks – human decisions are seen as more trustworthy and fairer. Also, algorithmic decisions that involve human traits evoke more negative emotions. Koene et al. (2017) deal generally with the operationalization of fairness for

[10] Three contributions can be assigned to several groups due to the application of multiple methods and can therefore provide different insights for the review.

the purpose of implementation in algorithms and derive from the results of the group discussion and a survey that fairness seems to be very context-dependent in the eyes of the interviewees and depends on the information of the users and come to the conclusion that there is no "(...) unique, globally approved, definition of fairness" (p. 2).

Two contributions agree with the findings of Koene et al. (2017) on the definition of fairness, but argue along the empirical lines that the application of a particular definition of fairness in algorithms should be implemented in the future. Saxena et al (2019) argue that individuals prefer calibrated fairness ("ratio") to equal and performance-based fairness in the context of a fictitious lending scenario. Srivastava, Heidari and Krause (2019) point out that the simplest mathematical definition of fairness – demographic parity – in two different application scenarios (criminal risk prediction and skin cancer prediction) is closest to the interviewees' idea of fairness.

Three contributions derive guidelines and requirements for the achievement of fairness in algorithms from the empirical results. Lee, Kim and Lizarondo (2017) argue in this context on the basis of interview results that algorithms should consider different notions of fairness, taking into account the population as well as the context, in order to work fairly. Vallejos et al. (2017) show that the deliberative procedures used indicate that young people want to know more about algorithms; more transparency and more control over the way algorithms use their personal data is required to make them fairer. The participants in the group discussion are also of the opinion that there should be a global approach to the regulation of fairness and ethics guidelines in algorithms. Woodruff et al (2018) also found that respondents see a connection between unfairness in algorithms and national dialogue on ethnic inequality and economic inequality. Based on these results, the authors conclude that fairness should be integrated as a value in the design and development of an algorithm. Furthermore, preliminary studies are to be set up for the development of fair algorithms that integrate test persons with different perspectives and ethnic backgrounds and thus establish dialogue with different social groups.

Seven contributions deal with the influence of different factors on the perception of fairness in an algorithm. Analogies can be found between the results of Dodge et al. (2019) and Binns et al. (2018). Both studies examine the effect of explanatory approaches on the perception of the fairness of an algorithmic decision or the fairness assessment of ML systems.

In order to examine the influence of different explanatory approaches on the perception of fairness of an algorithmic decision, Binns et al. (2018) have selected five fictional contexts and related scenarios that lead to a negative decision for an individual (e.g. lending). The respondents were told which information about the individual was available and which of four explanations was used to decide the algorithm. The assessment of the fairness of this decision should be based on five statements, each assigned to a dimension of justice. The results show that some of the respondents found the processing of personal information about an individual in an algorithmic decision-making process unfair, others found the system to be statistically fair. Overall, it can be seen that explanations can have either an effect or no effect on the fairness assessment of a decision – it depends on the combination of scenarios and explanatory approaches.

Dodge et al. (2019) also show that certain approaches to explanation ("cased-based explanation") are perceived as less fair, while others can strengthen confidence in the fairness of the algorithm ("global explanation", "local explanation"). Furthermore, the individual preferences of respondents on the fairness of an algorithm can influence their reaction to the different explanatory approaches – the authors explain this with a lack of agreement on the significance of moral concepts.

In their study, Shin and Park (2019) measured the influence of fairness, accountability and transparency (FAT) on individual satisfaction with algorithms with special consideration of trust – used here as a moderating variable. The results show that the perception of FAT of an algorithm by the user can significantly influence the cognition and acceptance of the user and that the perceived FAT plays an important role in the user satisfaction of algorithms. This study also found that the interaction effect between trust and the characteristics of algorithms (FAT) affects satisfaction with the algorithm.

The study by Lee and Baykal (2017) shows not only the results mentioned above, but also insights into the influence of interpersonal competence and programming skills on the fairness assessment of human and algorithmic decision-making. The interviewees, who have a high level of interpersonal competence, feel that the discussion-based decisions are much fairer than the algorithmic decisions. Programming skills of the individuals – as proxy for knowledge in the field of algorithms – on the other hand had no effect on the perception of fairness of the discussion-based decision. Saxena et al. (2019) also contribute additional insights by investigating the influence of sensitive information on the fairness evaluation of an algorithmic decision. The results show that information about the ethnic group membership of an individual only has an influence on which definition of fairness is perceived to be fairer by respondents in individual cases.

Grgić-Hlača et al. (2018a) examine in their analysis the influence of the use of certain features on the evaluation of the fairness of the algorithmic decision-making process using binary logistic regression. The COMPAS tool, which is used in the USA to estimate the risk of a defendant recidivism in court, was used as a case study for the fairness assessment of various features. The results support different views of the respondents as to which features are considered fair for use in the algorithm – for example, from the point of view of a majority of respondents it is fair if the 'criminal record' feature is used for the algorithmic decision. Nevertheless, the authors find large discrepancies between respondents' perception of which features are fairer than others for the decision-making process.

In a further study by Grgić-Hlača et al. (2018b) a similar question is examined with the aim of developing measures for procedural justice that take into account the characteristics used in the decision-making process and evaluate the moral judgements of human beings about the use of these characteristics. The investigation again took place in the context of criminal recidivism, using the example of the COMPAS tool and additionally the prediction of illegal possession of weapons. The analysis examined to what extent the perceived fairness of a characteristic is influenced by additional knowledge about increasing the accuracy of the prediction. Additionally, the extent to which knowledge of the increase in differences in decision outcomes influences the perceived fairness of a feature was measured. Overall, the assessments of the various

characteristics varied widely. However, it could be noted that respondents classified those features as fairer that improved the accuracy of prediction and those features as more unfair that led to discrimination against certain feature holders. Thus, these results could be used to quantify procedural justice. In a next step, the fairness of feature combinations was investigated and it was found that high procedural equity leads to high distributive equity. This could be measured by the fact that the demand for a high level of process equity limited the range of functions to features that many respondents considered fair.

### 4.3 Synthesis

In the following, the results of the contributions will be brought together in a cross-study synthesis. Recommendations for empirical practice can be derived from the theoretical contributions: The operationalisation of fairness in algorithms should be chosen depending on the context and taking into account different scientific perspectives and fairness definitions in order to produce not only transparency but also non-discriminatory decisions (Binns, 2018; Gajane & Pechenizkiy, 2018; Lepri et al., 2018; Veale & Binns, 2017).

The results of the empirical technical debate also take up the fact that multidisciplinary approaches to definitions of fairness and discrimination should be developed and that various research disciplines should be taken into account in this context in the future (Žliobaitė, 2017). Furthermore, the results show that group and individual fairness do not necessarily have to go hand in hand and that the technical definition must take into account the persons for whom fairness criteria are to be applied (Khademi et al., 2019; Altman, Wood & Vayena, 2018).

From the 18 empirical contributions on the perception of fairness in algorithms, various insights can be gained. The general arguments about fairness in AI and algorithms show that individuals often perceive decision-making through artificially intelligent applications as fairer due to the objectivity and potential equal treatment they attest (Araujo et al., 2018; Lee & Baykal, 2017). However, it is also necessary to consider the context dependence (Koene et al., 2017), because in the case of mechanical tasks the algorithms are just as familiar as human decision makers; in the case of decisions requiring human properties the algorithmic decision is less familiar in comparison, however (Lee, 2018). These findings go hand in hand with the findings derived from the theoretical contributions.

The results of the contributions to the application of certain definitions of fairness in algorithms also show that individuals prefer certain forms of fairness depending on the context – here the selected scenarios for empirical verification of the assumptions – (Saxena et al., 2019; Srivastava, Heidari & Krause, 2019) and therefore the perceived fairness of people varies depending on the application scenario.

From the results of the studies that have derived guidelines or demands for the achievement of fairness in algorithms, the argument of context dependence for individual fairness perceptions can also be read – which is already demanded in order to be able to guarantee fairness (Lee, Kim & Lizarondo, 2017). Transparency in the procedures and use of personal data in algorithms is also addressed (Vallejos et al., 2017;

Woodruff et al., 2018). Individuals seem to demand global approaches to the regulation of fairness measures and ethical guidelines for algorithms (Vallejos et al., 2017). Accordingly, the implementation of these in the design of algorithms must be taken into account (Woodruff et al., 2018).

The findings of the contributions, which deal with the influence of various factors on the individual's perception of fairness, also underline the need to consider the context of fair decision-making through algorithms (Binns et al., 2018; Grgić-Hlača et al., 2018a). Different individual moral concepts also lead to different fairness perceptions of algorithms (Dodge et al., 2019; Grgić-Hlača et al., 2018b). The accuracy of the prediction and confidence in the decision of an algorithm also influences the perception of fairness (Grgić-Hlača et al., 2018b, Shin & Park, 2019) and again points to the desire for objectivity in algorithms. There are repeated results that indicate that ensuring transparency in algorithms contributes to satisfaction – here expressed by fairness – with an algorithmic decision (Shin & Park, 2019).

## 5 Summary of the results of the review

Overall, it can be said that the theoretical and empirical literature provides valuable answers for the investigation of fairness in algorithms. The investigations come from different research disciplines; a large part of the included studies come from Computer Science. The fields of application of the studies vary – some were applied to social or legal scenarios, others to multiple or no explicit contexts. So far, no extended field has emerged.

For the analysis of the individual perception of fairness in algorithms, qualitative surveys and deliberative methods were used in addition to quantitative methods such as experiments or surveys. It should be emphasised that the qualitative studies have led to findings which the quantitative studies have not yet shed light on – these have dealt with specific subgroups and their perception of fairness (Woodruff et al., 2018; Vallejos et al., 2017).

To answer the research question 'How do individuals perceive fairness in the decision-making of algorithms', the following points can be highlighted along the lines of the results of the empirical studies:

(1) The perception of the fairness of an algorithm is significantly influenced by its context (Binns et al., 2018; Grgić-Hlača et al., 2018a; Koene et al., 2017; Lee, Kim & Lizarondo, 2017; Veale & Binns, 2017). Therefore, it is not possible to identify a particular definition of fairness that is applicable to each survey or that is fairest from the respondents' point of view (Koene et al., 2017, p. 2).

(2) Individuals often attest algorithms to objectivity and expect equal treatment in decision-making, which is why they consider them fairer to human decision-makers (Araujo et al., 2018; Lee & Baykal, 2017).

(3) Transparency, trust and individual moral concepts play a role in the individual perception of fairness in algorithms (Dodge et al., 2019; Grgić-Hlača et al., 2018b; Shin & Park 2019; Vallejos et al., 2017; Woodruff et al., 2018).

(4) Interpersonal competence of respondents, for example, tends to promote the preference of human decision-makers. Individuals also find it fairer and trust algorithms more when important decisions requiring human abilities (e.g. professional recruitment and evaluation tasks) have been made by human decision-makers (Lee, 2018).

## 6 Discussion of the methodological approach

A systematic review should be carried out along a transparent, structured, reliable and replicable approach based on specific quality criteria (see Chapter 2).

The reproducibility of the procedure is given from the point of view of the authors of the review, since the procedure was described in detail and the collected findings were carried out intersubjectively and comprehensibly on the basis of excerpts and tables. In addition, on the one hand the reliability was achieved by the systematic approach, and on the other hand the relevance assessment of the selected contributions was proven by Kappa ($\kappa = 0.74$), which was found to be statistically good.

The objectivity of the procedure can be assessed as partially limited, since the derived search terms and the further procedure of the literature search via pearl-growing and snowballing were determined by subjective discretion and knowledge of the authors. Nevertheless, relevant literature has been used to keep the selection mechanisms as objective as possible. The formation of inductive categories for the screening and the final decision on the inclusion of the contributions also cannot be recognised as completely objective, but the categories were discussed in detail by the experts and assessed as practicable.

It should also be mentioned that the selection of other criteria (e.g. an extension of the search terms to the abstract and not exclusively to the title, etc.) for inclusion would have led to different database results. The authors of the review therefore do not make any claim to completeness.

The evaluation of the contributions in terms of answering the research question was result-oriented, which led to some contributions being described in more detail in the Review than others. The reason for this is the gain in knowledge that was heard through such contributions. Nevertheless, beyond the design of the contributions, the aim was to identify similarities and to use these arguments for the synthesis.

## 7 Conclusion and outlook

In summary, it can be stated that various definitions of fairness in algorithms have already been scientifically illuminated and that the results on individual fairness perception vary considerably – depending on the chosen context of the study. Other contexts

or fields of application, such as the perceived fairness in higher education admission procedures, could provide new insights, since in this context other forms of resource allocation are at stake (Saxena et al., 2019, p. 6). With regard to further research on this topic, some contributions have already pointed out that there is a lack of interdisciplinary knowledge and that in the future scientists from different disciplines should develop, evaluate and validate common theoretical concepts and alternative measurement models for different tasks (Lepri et al., 2018, p. 618). In addition to empirical research, various stakeholders should strive to establish common standards for fairness in algorithms and, if necessary, integrate them into institutional or legal frameworks in the future in order to guarantee these (Lepri et al., 2018, Vallejos et al., 2017; Žliobaitė, 2017). In this way, the discrimination or marginalisation of individuals or groups in algorithmically controlled decision-making processes can be minimised or avoided altogether in the future.

With regard to the method of systematic narrative review to answer the chosen research question, it can be concluded that the extension to a Living Review is envisaged in order to be able to include the constantly growing field of empirical results on the topic.

## References


1. ACM FAT via Twitter (2019, August, 23). "The full paper deadline has now passed, and FAT*2020 has received *291 papers* across all tracks: a whole 80% more than were reviewed last year!". Available at: https://twitter.com/fatconference/status/1164894203840749568, zuletzt abgerufen am 13.09.2019.
2. Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthana-rayanan, V. (2017). NMC Horizon Report; 2017 Higher Education. Retrieved from http://cdn.nmc.org/media/2017-nmc-horizon-report-he-EN.pdf
3. Altman, M., Wood, A., & Vayena, E. (2018). A Harm-Reduction Framework for Algorithmic Fairness. *IEEE Security & Privacy*, *16* (3), 34–45. https://doi.org/10.1109/MSP.2018.2701149
4. Araujo, T., de Vreese, C., Helberger, N., Kruikemeier, S., van Weert, J., Bol, N., Oberski, D., Pechenizkiy, M., Schaab, G., Taylor, L. (2018). *Automated Decision-Making Fairness in an AI-driven World: Public Perceptions, Hopes and Concerns*. 1–20. Retrieved from https://www.ivir.nl/publicaties/download/Automated_Decision_Making_Fairness.pdf
5. Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Journal of Machine Learning Research*, *81*, 1–11. Retrieved from http://arxiv.org/abs/1712.03586
6. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3173574.3173951
7. Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review*. Los Angeles: SAGE Publications Ltd.

8. Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and Predictive Algorithms. *Data & Civil Rights: A New Era of Policing and Justice*, 1–11. https://doi.org/10.27.2015
9. Cooper, H., & Hedges, L. V. (1994). Research Synthesis as a Scientific Enterprise. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis. A Practical Guide* (pp. 3–14). New York: Russell Sage Foundation.
10. Counsell, C. (1997). Formulating Questions and Locating Primary Studies for Inclusion in Systematic Reviews. *Annals of Internal Medicine*, *127* (5), 380–387. https://doi.org/10.7326/0003-4819-127-5-199709010-00008
11. Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593
12. Dodge, J., Liao, V. Q., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *International Conference on Intelligent User Interfaces, Proceedings IUI*, *Part F1476*, 275–285. https://doi.org/10.1145/3301275.3302310
13. Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: Perils, Promises and Solutions. A Systematic Review. *Journal of Big Data*, *6* (1). https://doi.org/10.1186/s40537-019-0177-4
14. Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589
15. Gajane, P., & Pechenizkiy, M. (2018). *On Formalizing Fairness in Prediction with Machine Learning*. Retrieved from http://arxiv.org/abs/1710.03184.
16. Gough, D., Oliver, S., & James, T. (2012). *An Introduction to Systematic Reviews* (2nd ed.). Los Angeles: SAGE.
17. Green, S., Higgins, J. P., Alderson, P., Clarke, M., Mulrow, C. D., & Oxman, A. D. (2008). Introduction. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 3–9). Chichester, West Sussex: John Wiley & Sons Ltd.
18. Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018a). Human Perceptions of Fairness in Algorithmic Decision Making. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 903–912. https://doi.org/10.1145/3178876.3186138
19. Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018b). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 51–60. New Orleans, Louisiana, USA.
20. Guyatt, G., Rennie, D., Meade, M. O., & Cook, D. J. (2008). *Users' Guides to the Medical Literature: Essentials of Evidence-based Clinical Practice* (2nd ed.). New York: McGraw Hill Professional. https://doi.org/10.1036/0071590382
21. Hammann, M., & Jördens, J. (2014). Code open tasks. In D. Krüger, I. Parchmann, & H. Schecker (Eds.), *Methods in scientific didactic research (pp*. 169-178). https://doi.org/10.1007/978-3-642-37827-0_14

22. Higgins, J. P., & Deeks, J. J. (2008). Selecting studies and collecting data. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 151–185). Chichester, West Sussex: John Wiley & Sons Ltd.
23. Higgins, J. P., & Green, S. (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, West Sussex: John Wiley & Sons Ltd.
24. Ibrahim, R. (2008). Setting up a research question for determining the research methodology. *ALAM CIPTA, International Journal on Sustainable Tropical Design Research & Practice*, *3* (1), 99–102.
25. Young nickel, K. (2017). *Interdisciplinary opinion leader research*. Wiesbaden: Springer Trade Media. https://doi.org/10.1007/978-3-658-17786-7
26. Khademi, A., Foley, D., Lee, S., & Honavar, V. (2019). Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2907–2914. https://doi.org/10.1145/3308558.3313559
27. Kirste, M., & Schürholz, M. (2019). Einleitung: Entwicklungswege zur KI. In V. Wittpahl (Ed.), Künstliche Intelligenz. https://doi.org/10.1007/978-3-662-58042-4_1
28. Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in SE. *Keele University and Durham University Joint Report*, 1–44. https://doi.org/10.1145/1134285.1134500
29. Koene, A., Perez, E., Ceppi, S., Rovatsos, M., Webb, H., Patel, M., Jirotka, M., Lane, G. (2017). Algorithmic Fairness in Online Information Mediating Systems. *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17*, 391–392. https://doi.org/10.1145/3091478.3098864
30. Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5* (1), 205395171875668. https://doi.org/10.1177/2053951718756684
31. Lee, M. K., & Baykal, S. (2017). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1035–1048. https://doi.org/10.1145/2998181.2998230
32. Lee, M. K., Kim, J. T., & Lizarondo, L. (2017). A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, *2017-May*, 3365–3376. https://doi.org/10.1145/3025453.3025884
33. Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for Studies. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 95–150). Chichester, West Sussex: John Wiley & Sons Ltd.
34. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*, *31* (4), 611–627. https://doi.org/10.1007/s13347-017-0279-x
35. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European*

*Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

36. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000097
37. Monahan, J., & Skeem, J. L. (2016). Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology*, *12*(1), 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945
38. Orwin, R. G. (1994). Evaluating Coding Decisions. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 139–162). New York: Russell Sage Foundation.
39. Petrasic, K., Saul, B., Greig, J., & Bornfreund, M. (2017). Algorithms and bias: What lenders need to know. *White & Case*.
40. Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences*. Oxford: Blackwell Publishing. https://doi.org/10.1002/9780470754887
41. Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., & Liu, Y. (2019). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *ACM Conference on AI, Ethics, and Society (AIES)*. Retrieved from http://arxiv.org/abs/1811.03654
42. Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019
43. Speicher, T., Heidari, H., Grgić-Hlača, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices.*Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18, August 19–23, 2018, London, United Kingdom*, 2239–2248. https://doi.org/10.1145/3219819.3220046
44. Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*, 2459–2468. https://doi.org/10.1145/3292500.3330664
45. Vallejos, E. P., Koene, A., Portillo, V., Dowthwaite, L., & Cano, M. (2017). Young People's Policy Recommendations on Algorithm Fairness. *WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference*, 247–251. https://doi.org/10.1145/3091478.3091512
46. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4* (2), 1–17. https://doi.org/10.1177/2053951717743530
47. Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. https://doi.org/10.1145/3173574.3174230

48. Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31* (4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

## Appendix

### Appendix A – Codebook

**Inhaltsverzeichnis**

**1. Untersuchungsziel**

Ziel der Untersuchung ist es, einen systematischen Literaturreview anzufertigen, um einen Überblick über Beiträge zum Thema Fairness in Algorithmen geben zu können. Dabei sollen bisherige Forschungsbestrebungen aus unterschiedlichen Disziplinen untersucht werden.

**2. Untersuchungszeitraum**

Der Untersuchungszeitraum – hier gemeint als der Zeitraum, in welchem die erfassten Beiträge veröffentlicht wurden – erstreckt sich vom 01. Januar 2010 bis zum 26. Juni 2019.

**3. Untersuchungseinheit**

Untersuchungseinheiten der Analyse sind englischsprachige Beiträge, die Fairness im Zusammenhang mit Algorithmen thematisieren. Dabei ist zunächst nicht von Relevanz, aus welcher Disziplin die Beiträge stammen, soll es doch primär darum gehen, Forschungsbestrebungen und –ergebnisse zu diesem Thema zu identifizieren und systematisch zu erfassen. Berücksichtigt wurden (Proceeding) Paper, Artikel aus wissenschaftlichen Journals sowie Sammelbandbeiträge, die zum oben genannten Thema zwischen 2010 und dem Stichtag der Analyse – dem 26. Juni 2019 – erarbeitet wurden und in den folgenden elektronischen Literaturdatenbanken geführt werden:

- Web of Science
- PsycINFO
- IEEE Xplore
- Scopus

Um relevante Beiträge für die Analyse zu ermitteln, wurde eine Titelsuche in den genannten Datenbanken mithilfe Suchbegriffe durchgeführt, die zuvor über verschiedener Methoden abgeleitet wurden:

- Big Data
- Digital Data
- Artificial Intelligence
- Machine Learning
- Algorithm*
- fair*
- just*
- discrimina*

Mithilfe der Booleschen Sprache wurden die Begriffe in zwei kombinatorische Einheiten überführt, mit Operatoren versehen und in der nachfolgenden Schreibweise – fallspezifisch an die Suchmaske der Datenbank angepasst – zur Literaturrecherche verwendet:

(„Big Data“ *OR* „Digital Data“ *OR* „Artificial Intelligence“ *OR* „machine learning“ *OR* “Algorithm*”) **AND** (fair* *OR* just* O*R* discrimina*)

Zusätzlich wird eine händische Suche nach Beiträgen in relevanten Journals, Konferenzen und von Autoren angeschlossen und ein Duplikatcheck zur Reduktion doppelter Beiträge vorgenommen. Die Gesamtheit beläuft sich somit auf 1.349 Beiträge.

**4. Analyseeinheit**

Die Analyseeinheit umfasst neben den bibliographischen Angaben der Beiträge ebenso den Abstract, sowie den gesamten Beitrag.

**5. Erhebung**

Für die nachfolgende Codierung wird eine Vollerhebung der 1.349 Beiträge durchgeführt. Diese werden in einem ersten Schritt systematisch erfasst, indem die nachfolgenden formalen Kategorien für alle Beiträge codiert werden. Im Anschluss werden diese unter Berücksichtigung der Auswahlkriterien systematisch reduziert, sodass die Codierung der inhaltlichen Kategorien lediglich für die Beiträge vorgenommen wird, die zuvor als “relevant” codiert wurden.

**6. Formale Kategorien**

Die folgenden formalen Kategorien werden zunächst für alle Beiträge in einer Excel-Tabelle erfasst:

- Fortlaufende Nummer (Nr.)
- Database
- Authors
- Year
- Title
- Source Title
- Volume
- Issue
- Beginning Page
- Ending Page
- DOI
- Link
- Abstract
- Conference Title, Year
- Total Citations

## 7. Auswahlkriterien

Im Rahmen der Reduktion der Beiträge in Hinblick auf die Beantwortung der Forschungsfrage, werden neben dem Zugriff auf den Beitrag und der Konformität auch die Relevanz des Beitrags als Auswahlkriterien berücksichtigt.

### 7.1 Zugriff Beitrag

Ausgeschlossen werden zunächst Beiträge, auf die kein Zugriff möglich ist.

- Zugriff nicht möglich (0) → aussortiert
- Zugriff möglich (1) → in Liste behalten

### 7.2 Konformität Beitrag

In einem nächsten Schritt werden solche Beiträge nicht berücksichtigt, die keinen Abstract aufweisen und die trotz Recherche nicht gefunden werden können. Auch werden solche Beiträge nicht berücksichtigt, deren Titel zwar auf englisch geschrieben ist, der Abstract aber auf einer anderen Sprache verfasst ist.

- keine Konformität (0) → aussortiert
- Konformität (1) → in Liste behalten

### 7.3 Relevanz Beitrag

Um die Relevanz eines Beitrags für die anschließende Analyse bewerten zu können, werden die Abstracts aller Beiträge gelesen. Die Durchsicht der Abstracts wird von zwei Gutachtern übernommen. Damit dieser Schritt der Auswahl Forschungsstandards entspricht, wird eine Anzahl von Abstracts bestimmt, die zum Zwecke der Reliabilität von beiden Gutachtern gelesen werden. So sollen die Codier-Ergebnisse statistisch abgesichert sowie die Qualität der Auswertung zu erhöht werden. Die gebildeten Relevanzkriterien, die zuerst auf den Abstract angewendet werden, lauten:

- Verständlichkeit des Abstracts (rel_a)
  - trifft nicht zu (0)
  - trifft zu (1)
- Bezug zur abgeleiteten Forschungsfrage dieses Reviews (rel_b)
  - trifft nicht zu (0)
  - trifft zu (1)
- Verständlichkeit der Darstellung des Forschungsvorhabens (rel_c)
  - trifft nicht zu (0)
  - trifft zu (1)

- Verständlichkeit der Darstellung der Methodik (rel_d)
  - trifft nicht zu (0)
  - trifft zu (1)

Die Kriterien werden händisch in eine Printversion der Excel-Tabelle eingefügt und codiert. In einem nächsten Schritt werden die Ergebnisse dieses Vorgangs in ein Farbschema überführt, nachdem die Ergebnisse der beiden Gutachter zusammengefügt werden.

| **Relevance** | **Code (color)** |
|---|---|
| relevant (min. 3 points) | |
| potentially relevant (min. 2 points) | |
| relevant in other context (min. 1 point) | |
| irrelevant (no point) | |

Ein Beitrag, der aus Sicht der Gutachter drei der oben genannten Relevanzkriterien erfüllt, wird als *important* eingestuft. *Potentially relevant* sind solche Beiträge, die zwei oben aufgeführten Punkte auf sich vereinen können. Beiträge mit der türkisenen Farbgebung sind *relevant in other context* – beispielsweise solche, in denen sich ein Bezug zur abgeleiteten Forschungsfrage ausmachen ließ, für die konkrete Beantwortung aber nicht dienlich sind. *Irrelevante* Beiträge werden in rot markiert und können keine der Kriterien erfüllen. Zu allen für relevant bemessenen Beiträgen werden die Volltexte recherchiert und gespeichert.

## 8. Inhaltliche Kategorien – Exzerpte

Die nach den zuvor beschriebenen Auswahlkriterien ausgewählten Beiträge, die mindestens 3 Punkte der Relevanzkriterien erfüllen (Code Color: green), werden parallel zur Lektüre anhand der nachfolgenden Rubriken systematisch exzerpiert. Die Informationen werden dabei in natürlicher Sprache für jeden Beitrag in einem separaten Dokument festgeschrieben. Alle Informationen, die im Exzerpt erfasst werden, werden mit indirekten oder direkten Zitaten des zu exzerpierenden Beitrags belegt und dabei die dazugehörige Seitenzahl angegeben.

### 8.1 Bibliographie

In der Kopfzeile der Word-Tabelle, die als Vorlage erstellt wurde, werden die bibliographischen Angabe des Beitrages erfasst.

### 8.2 Forschungsdisziplin

Eingetragen wird hier, aus welcher Disziplin der Beitrag stammt. Sofern diese Information nicht im Abstract oder im Volltext gegeben wird, werden die Autoren recherchiert und deren Forschungsbereich festgehalten.

### 8.3 Forschungsfrage

Die Frage, welcher ein Beitrag nachgeht, soll erfasst werden. Vor allem über die Forschungsfrage lässt sich beurteilen, ob ein Beitrag für die Analyse dienlich ist. Sofern die untersuchte Fragestellung nicht explizit genannt wird, kann alternativ auch eine implizite Frage erfasst werden.

### 8.4 Forschungsvorhaben

Codiert wird entweder ein Teil der Kurzzusammenfassung aus dem Abstract des Beitrages oder es wird in wenigen eigenen Worten erfasst, was das Vorhaben des Beitrages ist. Für die weitere Zuordnung in der Analysestruktur stellt diese Information eine gute Grundlage dar.

### 8.5 Theorie

Da der Fokus des Reviews die Auseinandersetzung mit Fairness als theoretischem Gegenstand ist, soll hier vor allem aufgeführt werden, welche Definition von Fairness oder Gerechtigkeit oder Diskriminierung verwendet wird. Auch sollen Querverweise auf andere zitierte Beiträge oder Theorien festgehalten werden.

### 8.6 Methode

Hier wird notiert, ob qualitative/quantitativ/gemischt oder nur theoretisch gearbeitet wird. Sofern es sich um empirische Beiträge hält, soll die verwandte Methode ausführlich erfasst werden.

### 8.7 Abhängige Variable (Effekte)

Sofern es sich um eine empirische Untersuchung handelt und ein AV-UV-Design angewendet wurde, soll dies erfasst werden. Bei theoretischen Arbeiten als auch bei den Arbeiten, die zwar empirische Designs aber keine dieser Art verwendet haben, wird das Feld leer gelassen.

### 8.8 Unabhängige Variable

Sofern es sich um eine empirische Untersuchung handelt und ein AV-UV-Design angewendet wurde, soll dies erfasst werden. Bei theoretischen Arbeiten als auch bei den Arbeiten, die zwar empirische Designs aber keine dieser Art verwendet haben, wird das Feld leer gelassen.

### 8.9 Sample/Befragte

Sofern es sich um eine empirische Untersuchung handelt, sollen im Text angesprochene Informationen über das Sample oder die Befragten erfasst werden. Bei theoretischen Arbeiten oder Arbeiten, die darüber keinen Aufschluss geben, wird das Feld leer gelassen.

### 8.10 Hypothesen

Erfasst werden hypothetische Überlegungen, die sich aus dem Forschungsvorhaben des Beitrags ableiten lassen. Oftmals werden in Beiträgen allgemeine Vermutungen geäußert – diese können ebenfalls erfasst werden, sollten aber als diese gekennzeichnet werden (z.B. durch den Zusatz „allgemeine Hypothese/Vermutung“)

### 8.11 Ergebnisse

Um die Forschungsfrage des Reviews zu beantworten, bedarf es der Erfassung der Ergebnisse der relevanten Beiträge. Sofern möglich, sollen die Ergebnisse mit Rückbezug auf die Hypothesen oder das Forschungsvorhaben festgehalten werden. Sofern es sich anbieten, die Ergebnisse bereits in eigenen Worten zusammenzufassen, sollte dies getan werden.

### 8.12 Erfassung zusätzlicher Informationen

Sollte es Informationen im Text geben, die noch nicht in einer der oben genannten Rubriken erfasst wurde, ist dieses Feld dafür vorgesehen. Auch können hier sinnstiftende direkte Zitate eingefügt werden. Gibt es kritische Äußerungen eines für den Review relevanten Sachverhaltes gegenüber, sollte dies hier erfasst werden.

### 8.13 Literaturverweise aus dem Text

Unter der Exzerpt-Tabelle sollen Literaturverweise notiert werden, die im Exzerpt aufgenommen wurden und im Initalbeitrag verwendet wurden. Dies dient auch dazu, auf diese Literatur zu einem späteren Zeitpunkt gegebenenfalls zurückgreifen zu können.

### 8.14 Anhang

Sollte ein Beitrag relevante Grafiken oder Tabellen aufweisen, können diese per Screenshot in die Freifläche unter den bibliographischen Angaben gesetzt werden.

### 9. Kategorisierung der einzuschließenden Beiträge

Nachdem die Exzerpte nach der oben beschriebenen Vorlage erstellt wurden, sollen die Beiträge auf ihre Eignung in Hinblick auf die Beantwortung der Forschungsfrage des Reviews kategorisiert werden. Für die Entscheidung darüber wurde die folgende Kategorisierung erarbeitet:

#### 9.1 Kategorie 1

In der ersten und damit relevantesten Kategorie sollen Beiträge aufgeführt werden, die empirische die Wahrnehmung von fairness, justice oder discrimination in Algorithmen untersuchen. Dabei muss der Titel des Beitrags nicht exakt diesen Wortlaut enthalten, sollte jedoch inhaltlich diese Thematik empirisch bearbeiten. Die Ergebnisse der Beiträge in Kategorie eins sollen eine Antwort auf die abgeleitete Forschungsfrage des Reviews geben können.

#### 9.2 Kategorie 2

In der zweiten Kategorie sollen Beiträge aufgeführt werden, die sich empirische und technisch fairness, justice oder discrimination in Algorithmen auseinandersetzen. Technisch wurde hier deswegen als Begrifflichkeit hinzugefügt, da die finale Auswahl der Artikel einige Beiträge aus der Computer Science umfasst, deren Methodik komplex ist und die Ergebnisse dieser Beiträge nicht wie die aus Kategorie 1 direkten Bezug zur Forschungsfrage des Reviews herstellen können.

#### 9.3 Kategorie 3.1

In der Kategorie 3.1 werden Beiträge eingeordnet, die sich theoretisch mit der Operationalisierung von fairness-aware machine learning Algorithmen beschäftigen und damit thematisieren, wie ein Algorithmus zu sein hat, der Fairness-Kriterien berücksichtigt.

#### 9.4 Kategorie 3.2

In der Kategorie 3.2 befinden sich Beiträge, die sich ebenfalls wie in 3.1 theoretisch sind und sich mit der Operationalisierung von Fairness in machine learning unter einem sozialwissenschaftlichen Blickwinkel auseinandersetzen.

**Appendix B – Template Excerpt**

| **Text** | | | |
|---|---|---|---|
| Bezug | | | Literaturverweis |
| Forschungsdisziplin | | | |
| Forschungsfrage | | | |
| Theorie (Dimension von Fairness) | | | |
| Methode | | | |
| AV | | | |
| UV | | | |
| Sample / Befragte | | | |
| Hypothesen | | | |
| Ergebnisse | | | |
| | | | |
| Notiz | S. | Inhalt | Literaturverweis |
| | | | |
| | | | |

Literaturverweise aus dem Text

Anhang
- z.B. relevante Tabellen/Graphiken etc.