Noname manuscript No.
(will be inserted by the editor)

Managing Bias and Unfairness in Data for Decision Support

A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems

Agathe Balayn · Christoph Lofi · Geert-Jan Houben

Received: date / Accepted: date

Abstract The increasing use of data-driven decision support systems in industry and governments is accompanied by the discovery of a plethora of bias and unfairness issues in the outputs of these systems. Multiple computer science communities, and especially machine learning, have started to tackle this problem, often developing algorithmic solutions to mitigate biases to obtain fairer outputs. However, one of the core underlying causes for unfairness is bias in training data which is not fully covered by such approaches. Especially, bias in data is not yet a central topic in data engineering and management research.

We survey research on bias and unfairness in several computer science domains, distinguishing between data management publications and other domains. This covers the creation of fairness metrics, fairness identification and mitigation methods, software engineering approaches and biases in crowdsourcing activities. We identify relevant research gaps and show which data management activities could be repurposed to handle biases and which ones might reinforce such biases. In the second part, we argue for a novel data-centered approach overcoming the limitations of current algorithmic-centered methods. This approach focuses on eliciting and enforcing fairness requirements and constraints on data that systems are trained, validated, and used on. We argue for the need to extend database management systems to handle such constraints and mitigation methods. We discuss the associated future research directions regarding algorithms, formalization, modelling, users, and systems.

Agathe Balayn E-mail: a.m.a.balayn@tudelft.nl · Christoph Lofi E-mail: c.lofi@tudelft.nl · Geert-Jan Houben E-mail: g.j.p.m.houben@tudelft.nl
Delft University of Technology, Delft, the Netherlands

Keywords Bias and Unfairness · Decision Support Systems · Data Curation · Bias Mitigation · Bias Constraints for DBMS

1 Introduction

Context. Data-driven decision-support systems [143] are applied to many scenarios to allow for faster and more informed decision-making. For example, such systems help to decide which candidate to hire for a job (as used by Amazon [87]), inform judges of the risk of an offender to re-offend (like the COMPAS system in the US [36]), decide on how to react when an accident is foreseen in a self-driving car [77], etc. However, these systems can suffer from various ethical issues: i) they are often accused to lack transparency, ii) their outputs are often not explainable, iii) they might infringe the privacy of multiple stakeholders, and iv) they are claimed to be unfair towards certain groups of the population.

Problem focus. We focus on the unfairness of such data-driven decision-support systems arising from uncontrolled biases. For instance, the Amazon screening system exhibited an unfair gender bias, while the COMPAS system was accused of being racist [36]. These issues recently came to prominence due to reports in public media and rulings such as the European Union General Data Protection Regulation (GDPR) [128], while also recently mentioned in the Seattle Report on Database Research [45].

Data-driven decision support systems have a data management component and a data analytic component, which typically utilizes machine learning models.

One of the main sources of the unfairness of such systems lies in biases within the data on which the decision models are trained [68]. The machine learning model of the COMPAS system might have been trained on a dataset imbalanced with respect to a protected attribute such as race, and hence the decision model trained on it makes more errors for the underrepresented minority class. The Amazon system might have been trained on a dataset of previous hiring decisions where men have a higher chance of receiving positive decisions, and thus the decision model also exhibits a skewed distribution towards men. These biases are often not detected unless a deployed system behaves unfairly towards a subgroup of the population.

Motivation. Works stemming from the machine learning and data mining communities have started to tackle unfairness from certain angles like evaluating the outputs of trained models [211]; and mitigating unfairness by post-processing the outputs of the system [72,42,35], or modifying the training process of the inference algorithms [98,181,25,62,168,43,91,141], or pre-processing the training data [114,207,57,70,71]. Nonetheless, most of these approaches do not focus on the root cause of unfair systems –uncontrolled biases in the training data– but on the data analytics aspects. Furthermore, it is pointed out that they are not easily accessible and applicable by practitioners to real-life cases [76,173].

We believe that more extensive works on bias should be undertaken by the data management community, and this paper highlights the research gaps towards that goal. Our focus is on data-driven systems that have a machine learning component for decision making, and the biases that arise from these systems. It allows us to scope our work to a subset of decision-support systems and to identify concrete gaps for these types of systems, while encompassing all data management research that discusses bias and unfairness since machine learning models do use data.

Approach. We survey data management and other computer science literature on fairness separately. For this, we highlight and discuss: 1) quantitative overview of the research, 2) research topics, 3) methods and their limitations. We continue with a gap analysis that outlines issues and possible solution spaces to tackle unfairness from a data-management perspective, arguing that bias and unfairness should be a central topic in data management. Additionally, we propose a novel approach addressing several of these gaps by introducing requirements-driven bias and fairness constraints into database management systems. In Figure 1, we summarize in details the steps that we take to achieve the contributions discussed below.

Contributions. With this survey, we aim to foster the interest of the data management community in unfairness in data-driven decision-support systems by presenting state-of-the-art literature in various fields. We also identify gaps in current data management research which, if addressed, should bring systems closer to a fair state. We discuss those gaps and provide directions for future data management work. This survey paper is both a research proposition and a call to this community to address such fairness challenges, as some of them are the result of uncontrolled data management activities, while others would be best addressed by adopting existing data management works.

In summary, we make the following contributions:

- We outline the state-of-the-art of computer science domains actively working on bias and unfairness (section 4, section 5).
- We systematically survey existing research on bias and unfairness issues related to data management (section 7)
- We identify bias and unfairness-related research gaps (section 8) in data management, and propose new research directions (section 9) and challenges (section 10).

2 Background: Bias, Unfairness and Decision Support Systems

In this section, we set the context of the literature survey by outlining the current formulation of the problem and showcasing industry practices on the topic.

2.1 Terminology

We assume that there is a function at the core of a data-driven decision-support system that applies labels (representing decisions) to data instances (representing cases), mimicking an intelligent (human) decision-making process. This function is typically a machine learning model (like a classifier, regression, or a ranking algorithm, etc.) trained on labeled data and exposed to unseen instances after deployment.

Decisions, whether made by people or systems, may show bias. A bias is observed if data instances belonging to certain classes show a systematically different label distribution compared to instances belonging to other classes. Classes group data instances that relate to the same conceptual types, which is typically expressed by sharing certain attribute values (e.g., data instances representing females, data instances associated with a negative or positive target label, etc.). Thus, a bias is a

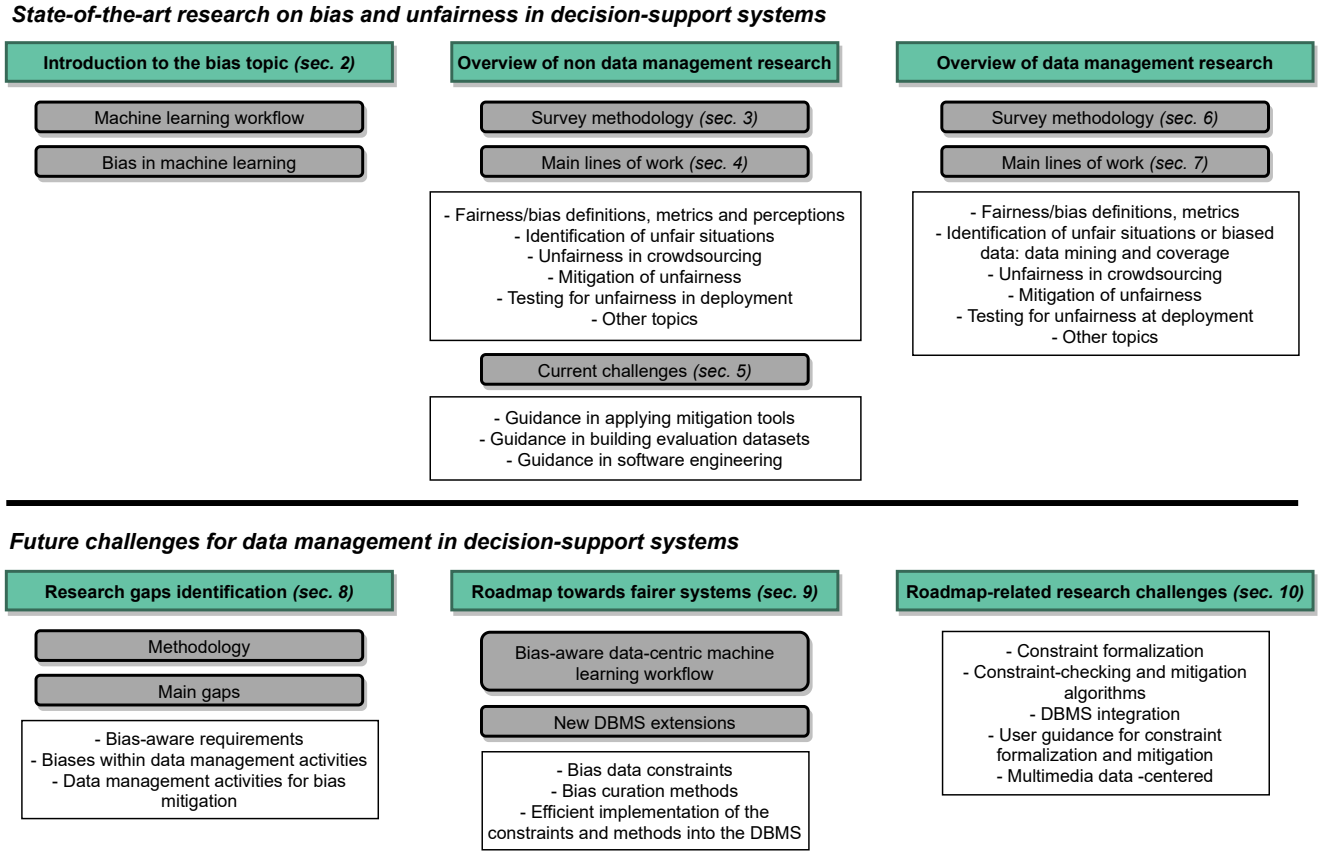

Fig. 1 Overview of the paper structure. After performing a survey of the state-of-the-art in various communities tackling issues of fairness and bias in some relation to machine learning, we identify research gaps and propose a set of research challenges for the data management communities.

statistical statement on class distributions, and it relies on the human judgment if a given bias is indeed problematic or not; some biases can be non-problematic or irrelevant, while others would require intermediate intervention. We will formalize this further in later sections.

We propose this general statistical definition of bias to address machine learning applications, while remaining in accordance with previous definitions: For instance, Olteanu et al. [129] define data biases as "a systematic distortion in the sampled data that compromises its representativeness.", where the distortion in our definition is referred to by mentioning the difference between classes. Many notions of unfairness in decision-support systems [21] (see the survey of definitions and metrics in section 4) are also based on some notions of inequality between groups (or classes).

Some biases can be *desired* and part of correct system functionality. For instance, in a system that predicts the likelihood of a criminal offender to re-offend, the class of individuals who actually re-offend should indeed be systematically attributed a higher probability to re-offend than the class of individuals who did not (as we see bias as a systemic difference in class-label distribution). While our example is simplified and allows to talk about "desired bias", we would like to warn the reader of the complexity of this idea in practice. Indeed, the data attributes and target labels that are used in practice in datasets are often distorted proxies for the actual notion to infer, or for the targeted use of the system, hence the observed bias might not be meaningful.

An *undesired bias* is a bias that is considered problematic, possibly *unfair* by the stakeholders of the system or other persons impacted by the system. Typically, this is observed when biases relate to *protected* attributes of sensitive nature. Defining protected attributes is often the result of an ongoing societal or ethical discussion, and protected attributes can emerge or change over time. For example, the COMPAS system [2] was accused of being racist towards Black defendants (i.e. "race" is a protected attribute in this case), as the rate at which it incorrectly inferred that certain Black defendants would recidivate was significantly higher than the same rate for White defendants. Stoyanovich et al. [178] explains through the informal

mirror metaphor that a bias (here what we term undesired bias) arises either when a system's outputs or the data reflects the world situation (with the idea that generally, data reflect the world), but this situation is not desirable; or when a system's outputs or data differ from the current world situation due to errors in the measurement process. In our example, the bias is an example of the second type, where the model's inferences are differing from the world situation (actual recidivism).

Lastly, an *unimportant bias* is detectable by statistical tools as a systematic difference in label distribution across classes, but societal discourse does not see these classes as sensitive and the resulting bias as problematic. For instance, the COMPAS system ignores age bias: even though there might be a systematic age bias in the system's decisions, this bias has been explicitly considered not important by the lawmakers (in this case, "age" was stated to be not a protected attribute by the system designers –however, over time, the view on this could change and "age" could also be seen as a protected attribute and thus making an observed age bias undesired). In general, bias is unimportant when it refers to classes defined by attributes that are meaningless for the context at hand, such as T-shirt color, or shoe shape.

These biases observed in the outputs of a system are usually traced back to the data that are fed to the underlying machine learning model at training and deployment time, or to the machine learning algorithm that is employed.

2.2 Current Practices

The typical development workflow for a data-driven decision-support system [142,143] follows a traditional software engineering approach. Typically, no explicit consideration about bias or unfairness is included in this process. Many modern real-life decision-support systems are based on machine learning algorithms that are trained and tuned on training data, and evaluated using test data [2,1,58]. These datasets typically contain (historic) examples of cases and their decisions (e.g., photos and their labels, job applications and their evaluations, etc.). Often, the training focus is on accuracy, and can be captured by metrics like precision, recall, or F1-scores [190]. After deploying a trained system, unfairness issues may arise during its operation [76]. It is not uncommon that this is the first time the system developers become aware that their system might treat certain cases unfairly in a systematic manner (by for example implicitly discriminating with respect to race, social class, or other protected attributes). Often, this is a direct result of optimizing the system for high accuracy scores with the given training and test data sets, while not including bias-related constraints into data collection and training.

As an example, consider a bank that uses data about their customers, their behaviors, and previous banking habits to build a system predicting if a customer would default on a loan. This system would then be used to recommend or even decide if a customer will be granted a loan. Unfairness could arise when certain categories of the population might have been discriminated against in the past (not always purposefully), and hence are also discriminated against by a system trained on historical decision data. If the training data have not been investigated for unfair biases and the trained machine learning model has not been evaluated for unfairness, such unfairness would only be discovered at deployment time, when certain customers would be treated unfairly.

After such unfair behavior is detected [76,190,173], the system designers often try to correct it by focusing on the class for which most complaints were raised: a new inference model with an architecture correcting unfairness could be trained, or an additional decision layer to correct the models' outputs for fairness (section 4) could be added. This requires in-depth knowledge and experience in the machine learning fairness literature, freedom on the choice of inference models, and computing resources to train additional models. The current process is costly and time-consuming. Also, unfair decisions of the deployed system incur allocation harms and potentially further damages like media outrages.

3 Data Analytics: Survey Approach

In this section, we explain how we proceeded to the survey of research on bias and unfairness outside data management, research that mainly focuses on the data analytics aspects of data-driven decision-support systems.

3.1 Methodology for the Selection of Papers

Our survey is based on a list of the different computer science domains that we consider to be working on topics related to the unfairness of decision-support systems, either because they use such systems, or because they have parts of such systems as an object of their research. This list is the following: machine learning, data mining, computer vision, natural language processing, recommender systems, computer-human interaction, human computation, software engineering, data management, and the interdisciplinary FAT (Fairness,

Accountability, Transparency) conferences (i.e. FAT* and AIES). For each of these domains, we retrieved papers of the main conferences (e.g. NeurIPS, KDD, CVPR, ACL, CHI, HCOMP) related to unfairness using two search engines (Google Scholar and DBLP). The approach to this was two-fold: 1) using unfairness-related keywords and the name of the domain, 2) using unfairness-related keywords and restricting the search to a list of the main venues of each domain. The list of keywords can be found in section 7. We reviewed the retrieved research papers from the different domains, compiled a list of major research topics currently addressed, and identified the main solutions proposed and their limitations. In this section, we do not cite all of the papers but only a selection of popular ones as there would be too many publications.

3.2 General Overview

The literature on bias within data-driven decision support systems spans a wide range of topics. The applications of these systems are diverse. These can be to support making decisions about individuals (e.g. deciding whether an offender's jail sentence should be extended based on its likelihood to recidivism, deciding whether to give a loan to someone based on their likelihood to reimburse it, etc.). In these cases, the systems are often trained on structured data about the individuals to make a decision on (e.g. data about the number of previous reimbursed loans, data about the number of crimes the offender previously committed, demographic data, etc.), but also sometimes on image or text data (e.g. deciding whether someone should get a treatment based on the description of their symptoms, deciding whether a scene is violent and police should be sent based on an image of the scene). It can also be to provide new knowledge for a later decision on someone or something, generally based on images (e.g. classifying whether someone is a doctor or a nurse based on their picture) or text (e.g. deciding whether a sentence is toxic).

In the next section, when it is not mentioned, we report works that mostly tackle applications using structured data, as research on unfairness for other types of data is more recent, and hence not all research outcomes are directly applicable to such data.

3.3 Main Research Directions

From our analysis of literature, we identified six main directions of research on unfairness and bias, which generally correspond to the perspective that different research communities have on the issue. While research starts with both the machine learning and data mining communities to define, formalize and measure unfairness, it then splits into two main directions –even though certain approaches are overlapping–: either identifying cases on unfairness in datasets, or developing ways to mitigate the unfairness when such datasets are used jointly with machine learning techniques for data analytics.

Stemming from the software engineering community and its recent interest in machine-learning-based systems, testing unfairness in the outputs of software is another developing direction. Finally, the human-computer interaction and the crowdsourcing communities started as well to develop an interest in the topic, respectively in understanding how humans perceive the unfairness of data-driven decision-support systems, and in investigating how humans might create certain of the biases that are found in the outputs of the systems.

As no other research community was identified with other research directions relevant to any case of data-driven decision-support systems, that is following these six directions that we organize our survey. In the last subsection, we mention other works that have not been widely adopted by computer science research yet.

4 Data Analytics: State of the Art

The goal of this section is to provide an overview of the current research topics and related state-of-the-art in the general computer science literature on bias and unfairness. We perform this survey through the lens of decision-support systems where bias and unfairness problems are currently most prevalent, i.e., where decisions suggested by the systems can be perceived as unfair or discriminating by certain stakeholders.

This section will serve as a foundation for our survey into bias in data management introduced in section 7, where we map the topics found in general computer science literature to the common data management workflow of most decision-support systems to identify research gaps.

4.1 Definitions and Metrics

Most works first propose *definitions* and *metrics* to quantify unfair situations, often based on definitions of discrimination in law[1].

[1] A survey and comparison of these definitions is in Zliobaite [211].

4.1.1 Overview

The mathematical definitions vary depending on the type of decision-support system: classification, ranking, regression, recommendation, etc.; but also based on underlying fairness notions like group fairness, individual fairness, or causal fairness [191].

Recently, new notions of fairness (e.g. multi-sided fairness [33]) involving more than one type of stakeholder and protected group were proposed for recommender systems: recommendations could be fair not only for the clients but also for the reviewers or providers of a service [104], or also for items presented in the system [86, 92, 171, 210].

New fairness notions could be identified from social sciences in order to make the systems more aligned with actual fairness values. Many of the proposed fairness definitions and metrics have multiple limitations [81]. For instance, group fairness does not account for unfairness within a given group and hence individual fairness was later proposed by Dwork et al. [52]. The fairness definitions are mostly based on equality notions of fairness but others might be more relevant for certain use-cases (e.g. affirmative actions [125], equity, need [61]). Besides, the identification of unfair situations through causality is also exploited by Madras et al. [117]. Indeed, most definitions rely on notions of correlations and not causation, whereas the ultimate goal of the systems and the metrics is to support making decisions ideally based on causal arguments.

4.1.2 Fairness Metrics

Here we give examples of the main mathematical definitions and metrics of fairness used for classification tasks.

All definitions and metrics assume the preliminary definition of a protected and a non-protected group of records (usually each record refers to a different individual) defined over the values of one or multiple sensitive attributes (also called protected attributes). For instance, in the aforementioned bank example, each record would represent a client of the bank with the attributes representing the information about this client. A sensitive attribute could be the gender, nationality, or age of the client. A protected group could be defined as all the clients whose age is between 15 and 25 years old, or as all the female clients whose age is in this interval. In the rest of this section, for the sake of clarity, we will take as a non-protected group the male clients, and as a protected group any other client. Most existing metrics only handle having one protected group and the rest of the records being aggregated into the non-protected group.

The definitions and metrics also require knowing the label the classifier predicted for each record (e.g. a positive prediction when a loan is granted and a negative prediction otherwise).

Most definitions rely on the comparison of statistical measures, and more specifically on checking equality of multiple probabilities, while the unfairness is quantified either by computing the difference or ratio of these probabilities. The definitions and metrics differ in the underlying values of fairness that they reflect, and on the exact measures and information required to compute them.

Group Fairness. *Group fairness based on predicted labels.* The first group of metrics only require knowledge of the predictions of a classifier for each record in a dataset and the membership of each record to the protected or non-protected group at stake. An example of such a metric is *statistical parity*[52]. Statistical parity is verified if the records in both the protected and unprotected groups have an equal probability to receive a positive outcome. An extension of such metric is the *conditional statistical parity* [42] which is verified when the above probabilities are equal, conditioned on another attribute.

In our bank example, the model would be considered fair according to this definition if the male applicants and the other applicants would have the same probability of being labeled as likely to repay the loan given all other attributes are equal.

Group fairness based on predicted labels and ground truth labels. The second group of metrics requires knowing both the classifier predictions and the ideal label that a record should be associated with. A classifier is fair according to these metrics when a measure of accuracy or error computed independently for the protected and the non-protected groups is equal across groups. This measure can be the true positive rate, the true negative rate, the false positive rate, the false negative rate, the sum of the true positive and false positive rates (named *equalized odds* [72]), the error rate, or the positive predicted value, the negative predictive value, the false discovery rate, the false omission rate, or ratios of errors (e.g. ratios of false negatives on false positives) [40]. All these metrics have different ethical implications outlined in Verma et al. [191].

In our example, a model would be fair based on these definitions if the selected measure of accuracy or error rate is the same for both male and female clients. For instance, for the true negative rate, the model would be fair when the probability for male clients labeled as

likely to default to actually default is equal to this probability for the non-protected group. For the definition based on recall, the model would be fair if the recall is the same for male and other clients, i.e. if the proportion of male clients being wrongly labeled as likely to default among male clients that would actually repay the loan is the same as for the clients of the protected group.

Group fairness based on prediction probabilities and ground-truth label. The third group of metrics requires knowing the prediction probabilities of the classifier and the ideal label. For instance, *calibration* [97] is verified when for any predicted probability score, the records in both groups have the same probability to receive a positive prediction. For our example, this would mean that for any given probability score between 0 and 1, the clients getting this score belonging to the protected and non-protected groups should all have the same likelihood of actually repaying the loan.

These conceptions of fairness all take the point of view of different stakeholders. While the recall-based definition satisfies what the bank clients would ask themselves –"what is my probability to be incorrectly rejected?"–, the true negative rate-based definition better fits the bank point of view –"of my clients that I decided to reject, how many would have actually repaid my loan?". The statistical parity metric could be considered to take the society viewpoint as supported by regulations in some countries –"is the set of people to whom a loan is granted demographically balanced?".

Individual Fairness. Another set of metrics, often named *individual fairness metrics* in opposition to the above metrics that compare measures computed on groups (referred to as *group fairness metrics*), relies on the idea that similar individuals should be treated similarly independently of their membership to one of the groups. The computation of such metrics requires the knowledge of each attribute that defines the similarity between records, and the knowledge of the classification outputs.

Fairness through unawareness [102] is associated to the idea that the sensitive attribute should not be used in the prediction process. In our example, this would simply mean that the gender of the clients is not used by the model, either during training or deployment.

Causal discrimination [59] is verified when the outputs of a classifier are the same for individuals who are represented with the same attribute values for all attributes except the sensitive attributes. Two bank clients asking for the same loan, having similar financial and employment situations, and simply differing on their gender should receive the same predictions from a model.

Finally, *fairness through awareness* [52] is verified when the distance between the output distributions of the different records is lower than the distance between these records. The different bank clients, all being more or less similar, should receive predictions that follow the same order of similarity, i.e. two clients being similar according to the metric employed should receive predictions that are under this high similarity measure, while two clients being farther apart can receive predictions that are not necessarily as similar as the two previous ones.

Generally, the underlying idea behind these notions of individual fairness is that group fairness notions do not allow to take into account unfairness that could arise within the groups, contrary to these new notions. Essentially, group fairness reflects averages over sets of individuals –if the averages across groups are similar, then the model is considered fair–, while individual fairness is interested in each of the individuals and how they are treated in comparison to all other individuals –while a group average might seem high, two individuals within the same group might receive disparate treatment, which in average look fair. In our example, an unfairness measure such as disparate impact could be low, meaning that both male and female clients are given similar percentages of loans, indicating that the model is fair. However, under this measure, two female clients having similar financial status could be treated differently, one receiving the loan and the other not, as in average the measure could still be close to the one for the male group. That is the type of issue that individual fairness metrics target.

"Combinations" of Metrics. Kearns et al. [93] showed that both group fairness and individual fairness metrics present important limitations in scenarios where multiple protected groups are defined over the intersection of multiple sensitive attributes, despite these scenarios being the most common ones in practice. Typically, the metrics might not account for unfairness issues in certain intersectional groups. In reaction to such limitations, they introduced a new set of metrics that rely on combining the underlying ideas of both group and individual fairness, and a new set of algorithms to optimize machine learning classifiers for them.

Causal Fairness. A last set of metrics relies on causal relations between records and predictions, and requires the establishment of a causal graph [95]. For instance, *counterfactual fairness* [102] is verified when the predictions do not depend on a descendent of the protected

attribute node in the graph. In our example, using such metrics would require providing a causal graph, where the protected attribute would be one of the nodes, and would entail verifying that the node representing the loan acceptance/rejection decision is not dependent on the protected attribute node.

4.1.3 Conflicting Perceptions of Fairness

While there exists all these mathematical fairness definitions and metrics, they tend to be conflicting and it is impossible to comply with all of them simultaneously, as shown by Chouldechova et al. [40]. Consequently, few papers [195,108,107,20,65] study how the fairness of data-driven decision-support systems is perceived in order to choose the most relevant definitions taking into account stakeholders' preferences and mathematical trade-offs. Srivastava et al. [174] show that one simple definition of fairness (demographic parity) solely matches the expectations of users of hypothetical systems. Conversely, Lee et al. [108,107] and Grappiolo et al. [65] show that different stakeholders might value different and possibly multiple notions of fairness (e.g. efficient, egalitarian, or equalitarian allocations).

Biases of the end-users of the systems are also investigated since their decisions informed by the predictions impact the (un)fairness of the systems. For example, Zhang et al., Solomon et al. and Peng et al. [209,170, 139] study how cognitive biases of the systems' users influence how they use the outputs of the systems to make the final decision. Peng et al. [139] show in the context of candidate hiring that the final human decision might be gender-biased by the proportion of male/female candidates exhibited by the algorithm.

4.2 Identification of Bias and Unfairness

4.2.1 Data mining research

Many data mining papers, dating from 2008 to 2016, deal with discovering and measuring discrimination within datasets, the results being potentially useful for "debugging" the datasets for later training machine learning models. They investigate scenarios of direct and indirect discrimination, further complicated by additional privacy concerns [152] and cases where the protected attributes are unavailable.

Methods. At first, methods relied on learning rules based on the dataset features potentially used for making the decisions, and on identifying features leading to discrimination [138,153]. Later, situation testing was used to account for justified differences in decisions concerning individuals from different protected groups [114]. "Unlike hypothesis testing, where a statistical analysis is adopted to confirm a predetermined hypothesis of discrimination, the aim of discrimination discovery is to unveil contexts of possible discrimination." [151]. Certain papers combine data mining methods with additional statistical testing in order to verify the potential discrimination situations discovered [155].

Example. In our bank example, rules would be mined from the available dataset with the target label as consequent and other dataset attributes as antecedent.

A rule would be potentially discriminatory with direct discrimination if the antecedent contains one or more protected attributes. Actual direct discrimination would then be verified by setting a threshold α, and comparing it to the difference of rule confidence, for rules with and without the protected attributes –if the difference exceeds α, that would mean that the protected attributes have a strong effect on the rule and hence there is direct discrimination.

Let's use the following highly simplified rules for the sake of giving an example: (*permanent job, low amount loan* $\rightarrow$ *medium risk not to repay*, confidence 0.1) and (*permanent job, low amount loan, woman* $\rightarrow$ *medium risk not to repay*, confidence 0.6). If the difference between the two confidences (here $\alpha = 6$) is deemed important with regard to discrimination, then the second rule would be deemed directly discriminating: for instance if $\alpha = 3$, then it is not discriminatory, while with $\alpha = 7$, it is.

As for indirect discrimination, it manifests in certain cases when a rule is not potentially discriminatory as its antecedents do not contain a protected attribute. If background knowledge is available about the context of the data, and protected attributes are shown to be connected to the antecedents within this knowledge, then the rule might be indirectly discriminating.

An example of such would be if a rule such as *permanent job, low amount loan, district1234* $\rightarrow$ *medium risk not to repay* was found with high confidence, and from prior human knowledge, we would also know that the rule *district1234* $\rightarrow$ *Black community* holds with high confidence. Then, proposed algorithms could estimate the confidence of the rule *permanent job, low amount loan, district1234, Black community* $\rightarrow$ *medium risk not to repay*, and identify it as discriminatory.

4.2.2 Research on multimedia applications

Natural language processing. Natural language processing (NLP) [183] focuses on social, undesired biases usu-

ally related to gender or race. For example, text completion models are shown to perform better on text from majority languages such as Standard-American English than on text from socially-restricted dialects such as African-American English. These works usually identify undesired biases from their knowledge around the context of the application, and propose methods to quantify these biases, often through the use of semi-synthetic datasets.

Computer vision. On the contrary, in computer vision, most papers tackle systematic dataset biases that are not necessarily related to human values but to properties of the world, such as image extrinsic properties like illumination [119,197] or image quality [169], or intrinsic properties like the background when classifying the sentiment of a picture [135] or the actions represented in images [110], or properties of the object to detect such as face orientation [100], or object scale in scene recognition [75].

Some works however investigate the diversity of the samples with regard to their cultural provenance for object detection tasks [167] or to protected attributes (e.g. gender bias in text for image captioning [73]). For instance, facial recognition models were shown to be trained on datasets which do not necessarily reflect the diversity of the populations on which the models are applied to, leading to an imbalance of accuracy for the different populations [175,32]. It is shown that these bias issues impact the performance and generalization of the trained models to new samples and datasets [94, 186].

4.3 Mitigation of Bias and Unfairness

4.3.1 Works dealing with tabular data

Mitigation methods decrease the unwanted biases in the outputs of the decision-support systems, consequently decreasing unfairness. When the input consists of tabular data, these methods can be divided into three categories that focus on different parts of the systems [21]: *dataset pre-processing*, *in-algorithm treatment*, and *post-processing of the outputs*. While the literature does not provide guidance in the selection of the method to apply, it seems to primarily depend on the notion of fairness to optimize for, and on the actual context of the application. For instance, certain developers might only have access to the machine learning models and then would apply in-algorithm methods, while data engineers might have the opportunity to transform the data before any kind of learning, which supports an earlier tackling of biases.

Mitigation through dataset pre-processing. For pre-processing, Luong et al. [114] propose a method that is inspired from situation testing, an experimental legal procedure to identify discrimination, in order to identify and later modify discriminative data labels. Zhang et al. [207] bring the ideas to use causal graphs to identify significant cases of unfairness, and to remove unfairness in the data through constrained optimization in order to maintain both utility and fairness of the dataset. Feldman et al. [57] propose data repairing methods. Hajian et al. [70,71] target simultaneously fairness and privacy preservation in datasets through an optimization algorithm.

Mitigation through in-algorithm treatment. Algorithmic modifications of the training process mostly focus on adjusting the loss function of machine learning models through the addition of regularization terms to include the selected notions of fairness, for classification [43, 91,141], for ranking [62,168], for matching tasks [98, 181], but also recently in the context of recommender systems [25].

Mitigation through output post-processing. Post-processing relies on the idea that model's predictions can be made fair by defining specific thresholds that transform the continuous outputs of the inference model into binary labels [72,42]. Specific methods vary in order to adapt to the specific group fairness metrics to optimize for, and sometimes to provide the option to defer the decision to the human operator [35].

4.3.2 Works dealing with multimedia data

In multimedia data research, we mainly identify two types of methods for mitigating biases. These are either pertaining to dataset pre-processing, or to in-algorithm treatment. These works are generally more recent and less numerous than for tabular data.

In computer vision, in order to make the outputs of the systems less biased, datasets are often modified to increase the diversity of present objects and extrinsic properties (e.g. collection or transformation of data samples, creation of synthetic datasets [100]). However, the goal of these efforts is typically to improve model performance, not necessarily fair treatment of certain classes. This is for example addressed by Amini et al. and Quadrianto et al. [9,144] who introduce fair feature representations that hide protected attributes. Directly controlling fairness in computer vision datasets is not a major topic yet [202,51].

Natural language processing [183] typically modifies the training dataset (semi-manual data augmentation

or annotation of samples with protected attributes), the embeddings of the samples as these have been shown to integrate unwanted biases from the large corpora of text on which they are trained, or the inference models. A more detailed account of these methods is given in [183].

4.4 Testing for Bias and Unfairness

4.4.1 Tabular data

Few works focus on evaluating the fairness of machine learning-based data-driven decision-support systems at deployment time, i.e. when ground truth for the new data samples is not known.

Galhotra et al., Angell et al., Udeshi et al. and Aggarwal et al. [59,10,187,6] propose test-suites to evaluate the fairness of software that relies on machine learning models, focusing on individual unfairness and developing methodologies for auto-generation of test inputs. For instance, the Aequitas framework [156] first proceeds to a random sampling of the input space to generate test cases, then the samples that are identified as discriminatory are used to further generate more test cases, by adding perturbations to these samples. In this case, it is not needed to know the ground truth, only the comparison between the model's inferences for the similar generated samples is important. Certain methodologies can identify more or fewer discrimination cases.

In contrast, Albarghouthi et al. [7] adopt a programming language perspective: they propose a way to formally verify whether certain decision-making programs satisfy a given fairness criterion (group or individual fairness) through encoding fairness definitions into probabilistic properties.

4.4.2 Multimedia data

For multimedia data, the same metrics are used as for tabular data. The difference however lays in that the required information to compute the metrics, such as the protected attributes, are often not readily available, and often impossible to extract easily solely from looking at the data samples (for instance, it is questionable whether race or gender can be annotated simply by looking at the picture of someone without knowing the person). Additional context or expertise might be required, such as in the cases of annotating the dialects employed in text samples or the race of the person who wrote the samples.

In computer vision, a few manually created benchmarks such as Gender Shades of Buolamwini et al. [32] are used to test specific applications like face detection.

In natural language processing, Sun et al. [183] explain that biases are quantified either by measuring associations between terms related to protected attributes, or by computing the prediction error of the data-driven decision-support system for the different subgroups represented by the protected attributes. This often requires generating data samples where the protected attribute is controlled to perform a systematic evaluation, especially because a large set of protected attributes can be considered in these spaces.

4.5 Bias in Crowdsourcing

Crowdsourcing is an essential component of many machine learning data-driven decision-support system workflows. It allows to collect data samples, or to label these samples so as to create ground truth labels to train the machine learning models on. From our analysis of existing works, we identify two meanings and research directions around bias in crowdsourcing. Closer to our topic, bias here refers to the way labels are attributed to data samples by annotators who project their own biases in the annotations. Another meaning however refers more to unfairness, and the pay inequality of various annotators among each other or compared to the minimum pay in their respective countries.

4.5.1 Biased annotations

Collecting "unbiased" data samples (biases from the data collection) and data labels (biases from the crowd workers like the biases in descriptions of people's pictures [133,134,132]) or identifying biases in datasets using crowdsourcing have been investigated with the purpose of later training machine learning models with such data. For example, Supheakmungkol et al. and Hu et al. [79,78] respectively propose a platform to obtain representative data samples and labels for various machine learning tasks (e.g. translation, computer vision, etc.) and a workflow to discover biases in image datasets.

For labels, methods to mitigate crowd worker biases are proposed: leveraging psychology and social computing theory [185] for political social media content; resolving disagreement in mined resources such as data from social media [67], or review ratings of items [111]; disambiguating biases from the task design [90,54]; and allocating crowd workers based on their demographics [19].

The effects of these biases on the outputs of machine learning models (e.g. unfairness as exclusion of opinions [17]) have not been studied extensively.

4.5.2 Unfair crowdsourcing tasks

Another research direction is the investigation of unfairness towards crowd workers. For example, Boyarskaya et al. [118] propose a scheme to pay workers fairly as a function of their work accuracy and the crowdsourcing task goals (maximum cost, minimum overall accuracy). A crowdsourcing plug-in [19] to allocate crowd workers based on their demographics and the related minimum wage is also investigated.

4.6 Other Focuses

Analysing the publications we retrieved from our systematic survey, we identify a few other emergent research directions, that have been developed to less extent until now, but that we believe are relevant to our topic, since they indirectly inform on issues around bias and unfairness either in the general development of the systems or in the data that could be used for these systems.

4.6.1 "Fair" software engineering

Other lines of work within computer science research are also interested in fairness. We specifically highlight works on designing methods to develop fairer software [192,109], coping with software designer biases [96,163, 34,83,193,149], fair processes to design software [61, 146,27]. For instance, German et al. [61] see code reviewing as a decision process where codes from different categories of population might be more or less often accepted, Rahman et al. and Bird et al. [146,27] point out that bug-fix datasets are biased due to historical decisions of the engineers producing data samples. Other papers such as [137,82,166,64,24,18,26,189] reflect on how projects (data science process, creation of fairness definitions) are conducted and how unfairness is seen and might arise in general from the problem formulation perspective.

Inspired by these works, in section 9, we also propose expanding the software engineering process of data-driven applications with additional fairness requirements.

4.6.2 Application-focused adaptation of the works on bias and unfairness

Certain works focus on bias and unfairness identification and mitigation methods for specific applications such as text analysis –e.g. Diaz et al. [47] address age bias in sentiment analysis–, social media news and existing polarization biases [48], fairness in self-driving vehicles [77], text processing [106]), web information systems and biases arising from them [164,46,127,112, 136,150,123,145,172].

Certain of these works are especially important for the goal of developing fair decision-support systems since they raise awareness of potentially biased sources of data, that are later used to train the machine learning models. For example, Das et al. and Quattrone et al. [46,145] show that user-generated content on Web platforms is biased towards certain demographics of the population due to the varied proportions of activity these demographics have (e.g. OpenStreetMap contributions are mostly from male users). We foresee this will have an impact on decision systems trained on datasets crawled from these platforms since the samples would be biased.

4.6.3 Human-computer interaction research

Certain researchers from the human-computer interaction community work on identifying the needs of data and machine learning practitioners in relation to new unfairness issues that arise from the application of data-driven decision support systems in real-life scenarios both for public and private sectors [76,190].

Besides, the Fairness, Accountability, Transparency (FAT*) community is also interested in problems related to social sciences, like the impact of publicly pointing out biases in company software [147], or the influence of decision-making systems on populations [126]. These works outline new research challenges for which technical processes and tools could be further developed.

5 Data Analytics: Limitations

In this section, we highlight the main limitations of current works on bias and unfairness, as they are argued by different research communities.

5.1 Limitations within each Research Direction

The topics of the previous subsections each bear certain limitations and research challenges.

Methods for identifying, testing, and mitigating biases do not allow for the development of fully fair and accurate systems and do not enable understanding where the unwanted biases come from in the systems for each of the different unfairness metrics. Besides, these methods are only adapted to increase fairness scores as measured by current metrics, but a system fair according

to one metric might not be fair for humans, as existing fairness definitions do not align fully with human perceptions of unfairness.

Also, due to the impossibility theorems between multiple metrics, there is currently no solution to build systems that are considered fair with regard to multiple metrics, whereas the combination of multiple metrics might be closer to the human notions of fairness. Methods do not all handle well intersectionality –when fairness is defined over the combination of multiple protected attributes–, whereas this is a closer notion of fairness than formalizations over single protected attributes.

Finally, existing methods almost all assume the prior knowledge of the protected attributes but this assumption might not hold in practice.

As for crowdsourcing works, not all biases coming from crowd workers are known from researchers or dataset developers until now, and hence they are not all dealt with when creating datasets.

5.2 Limitations in the Choice of Directions

Besides the above challenges tied in with the current approach of the issue that centers around machine learning algorithms, more general limitations are highlighted by certain works.

Mainly, the human-computer interaction community [76] suggests conducting more research to bridge the gap between existing machine learning methods and their applicability by industry practitioners. Works with professionals have been conducted to understand industry needs to deal with unfairness and bias and compared to existing research, showing that both bias mitigation and evaluation methods might not be adapted to real uses. Also, the software engineering community suggests taking a step back on the development of the systems to consider fairness in all development and deployment steps.

We discuss these gaps in more details below.

5.2.1 Algorithms and Tools for Data Bias Mitigation

Holstein et al. [76] point out that certain practitioners have more control on the data collection and curation steps than on the machine learning algorithm development, but that existing methods primarily focus on mitigation in the algorithm. Thus, we later advocate focusing on the data aspect of biases and unfairness.

Also, frameworks to help the selection of appropriate unfairness mitigation methods accounting for tradeoffs with other performance measures are needed.

5.2.2 Support for Evaluation

Practitioners also lack tools to facilitate the building of representative evaluation datasets and to identify and apply adapted metrics.

Most metrics are adapted for cases of allocative harms, that can arise when the goal of a system is to allocate resources to multiple stakeholders. They are however not often adapted for representational harms that arise from the classification of individuals in different categories, or from the association of individuals to (stereotyped) characteristics. This would be especially relevant in natural language processing (e.g. word embeddings denoting females are more closely associated to a number of job categories like maids and janitors contrary to the male embeddings) and in computer vision (e.g. images representing Black persons are more often classified as containing violence than images representing White persons). Also, most metrics assume knowledge of individual-level features whereas for privacy reasons this knowledge is often absent.

Besides, many unknown unknowns such as identifying before implementation or deployment the populations that could suffer from unfairness remain. Most research assumes the knowledge of the protected categories of population, generally coming from legislations, but there might be additional alarming context-dependent unfairness cases.

5.2.3 Guidance in Software Engineering

Many research opportunities are foreseen in the software engineering process in order to build ethics-aligned software. Roadmaps to develop ethical software are proposed [16,30], where the needs for methods to build ethical software, to evaluate the compatibility of the software with human values, and to help stakeholders formulate their values are highlighted. In this direction, Hussain et al. [80] and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [99] respectively argue for a collaborative framework to create software design patterns including social values (such values would be unwanted biases and different types of unfairness in our case) and for standards on algorithmic biases in order to provide a development framework that could support the creation of value-aligned algorithmic software. We believe this is also highly relevant for the data management community as, for instance, the data schemas developed in discussion with stakeholders need to be aligned with the values to integrate into the decision-support systems.

6 Data Management: Survey Approach

In this section, we first explain our survey methodology for bias and fairness research specifically in data management, and establish a quantitative research overview. This will serve as a starting point to identify research gaps in the next sections. Especially, in the previous sections, we established the general state-of-the-art in computer science research, and in the next sections, we compare it to data management works. Particularly, we investigate the extent to which data management research has differentiated until now from other research, with the intuition that more data management-specific activities should be investigated in the future. Besides, we map the data management research to the workflow of decision-support systems to identify important research gaps.

6.1 Survey Methodology

We surveyed a selection of data management venues for articles dealing with unfairness. This was conducted between August 2019 and December 2020, using two search engines (Google Scholar and DBLP). We retrieved papers using the keywords "bias", "fair", "disparate", "discrimination", "responsible", "diversity" and "coverage" combined with OR clauses, appended with constraints on the publication venues, covering the full publication history of the venues. The keywords were chosen to encompass as diverse publications as possible, as we noted that "fairness' is not the only term used for describing related works, but also notions of "discrimination", "bias", "diversity", or more general notions of ethics and responsible computing are employed.

In particular, we included publications from the ACM TODS, VLDB and TKDE journals, CIDR, ICDT, ICDE, SSDBM, EDBT, SIGMOD/PODS proceedings and the Data Engineering Bulletin [2]. With snowball sampling, we also selected the data management papers cited by the initially retrieved papers.

We filtered out the ones not actually addressing fairness topics of systems where some kind of decision is made, which relates to human individuals. Excluded papers mostly concern the fair allocation of computing resources or tasks between components of a computing system.

In our analysis, we distinguish the type of articles, e.g., full papers, tutorials, panels, keynotes, etc, but do not exclude any of them because we noticed that few full papers have been published, while many discussions on the topic happen either orally or in shorter papers.

[2] The Data Engineering Bulletin has a full special issue on fairness. [31]

6.2 Quantitative Overview

From the quantitative analysis of data management papers concerning unfairness and bias, we first of all notice that only 34 papers focus on the problems of biases in data-driven decision-support systems (*DDSS*), of which only 17 full papers; other than those, we see that mainly demos (5), tutorials (3), review papers (3) or vision papers (2) are presented, next to short papers (2), workshop paper (1), panel discussion (1), keynote (1). Most of these works have been published in the last 2 years.

This number is rather low compared to other research domains in computer science like machine learning, human-computer interaction, or data mining where unfairness is a common topic since 2010 and where there are more than a few hundred papers. While this observation is hardly surprising as most issues related to unfairness stem from the application of automated, often machine learning-based, data analysis techniques to human-related data, we argue that there should also be algorithm-agnostic bias considerations on the data management side.

6.3 Main Research Directions

All of the papers that we retrieved from data management venues, searching for a wide range of publications related to unfairness, fall into one of the topics also addressed by research outside of data management introduced in section 4. However, two topics identified in section 4 are not covered at all in data management (perceptions of fairness and testing of data-driven decision-support systems).

Yet, it is also important to note that several works are interested in questions of fair rankings, set selections, and data coverage, that are not discussed specifically in other disciplines. These questions are of importance for machine learning workflows where the pre-retrieval of "unbiased" datasets from databases could be necessary. These works can also be used independently of any machine learning model, simply as data analytics tools that provide decisions on data samples, such as for the tasks of ranking or selecting a limited number of candidates for job hiring.

The application areas are diverse; most of the times, the proposed methods are of a general nature, but sometimes specific to selected use-cases such as fair web page ranking [39], fair OLAP queries [158], fairness and

trust in multi-agent systems [194], or fair urban mobility [198].

7 Data Management: State of the Art

Here, we discuss current related research topics worked on in the data management community, map them to the topics discussed in the previous sections, and outline the main existing approaches.

7.1 Definitions

Three papers propose formal definitions of fairness, expanding on existing machine learning and data mining literature. Yang et al. [203] propose measures of fairness in ranking tasks, whereas Salimi et al. [162] propose a fairness definition for classification tasks to overcome limitations of previous definitions solely based on correlations or causality. Farnadi et al. [56,55] introduce fairness definitions, a first-order logic language to specify them, and mitigation methods. They argue that fairness is a concept depending on relations between the individuals within a dataset.

7.2 Identification

We identify multiple works that relate to the identification of undesired biases in datasets. These works seem to divide into three main categories depending on the approach they follow, and to the problem conditions that they define for themselves. While the first category of works is close to the data mining topics discussed in prior sections, the other two –coverage and unbiased query results– are specific to the data management community.

7.2.1 Data mining approaches

Similarly to other data mining works, some papers aim at identifying biases seen as discrimination within datasets. The context ranges from datasets of potentially discriminative historical decisions [208,69], with methods potentially encoded into the database system [154], to datasets of ranking scenarios [53,63] where unfair treatment towards specific groups might arise (these groups are not predefined), and to text datasets [205] where the semantics of certain user-generated comments might be discriminatory.

7.2.2 Coverage

Another topic related to the identification of biases within datasets more specific to data management literature is the notion of *data coverage*. Coverage relates to the idea that data samples in a dataset should sufficiently cover the diversity of items in a universe of discourse [15]. Without adequate coverage, applications using such datasets might be prone to discriminative mistakes. For example, certain computer vision models of Google performing image classification and object detection have been reported to have mistakenly labeled a Black woman as "gorilla", likely because the original training dataset did not cover enough images of Black women.

Dataset coverage characterization and mitigation methods Asudeh et al. [15] first proposed a formalisation of the coverage problem. They also present and evaluate methods both to efficiently evaluate the coverage of a dataset with respect to thresholds set by a practitioner for each dataset attribute, and to identify the type of data samples that are preferable to collect to solve the coverage issue accounting for the cost of data collection. These methods are based on the idea that representing a dataset as a pattern graph allows pruning a large amount of insufficiently covered data patterns represented as pattern relationships. Their link to coverage can then be exploited efficiently, instead of linearly traversing the whole dataset to identify uncovered patterns and to reason about their relationships.

Moskovitch et al. [124] take a different approach, aiming at efficiently estimating the number of items fitting different patterns in a dataset. This is based on pattern profiling and caching their statistics under resource constraints. Estimation functions estimate the count of any selected pattern with trade-offs between accuracy and efficiency based on those cached statistics. Lin et al. [113] argue that one of the main limitations of many previous works is the assumption that the considered dataset is constituted only of a single table. Applying existing methods to a realistic multi-table setup is shown prohibitively expensive. Instead, the authors propose a new parallel index scheme and approximate query processing to explore dataset coverage efficiently.

Coverage-informed database queries The previous approaches aimed at identifying coverage issues in a dataset that was "found" in a general fashion (as opposed to collected for a specific application in mind). Other methods focus on a setup with data present in a data warehouse, and propose to retrieve a subset of the data in

such a way that the data verify a specific application-oriented coverage objective. In this context, Accinelli et al. [5] propose a method to rewrite queries whose results would violate a specific coverage constraint into a similar query whose results now fulfill the constraint. In a similar fashion, Salimi et al. propose a way to identify biased results of OLAP queries, and rewrite similar queries to obtain unbiased results [158, 157].

Dataset nutritional labels. Some works promote the idea of creating *nutritional labels* for datasets, similar to the machine learning community which proposes to make datasheets to report on the creation of datasets [60] or to describe machine learning models [120]. In machine learning, these datasheets are intended for accountability, easier auditing of models, or for understanding of the limitations of models or datasets with respect to generalization abilities to extended tasks. Nutritional (data) labels in data management take a lower-level and more in-depth look at the datasets, and allow practitioners to interactively explore dataset distributions to identify diversity and coverage issues within the datasets themselves.

Particularly, Sun et al. [182] develop MithraLabel, which aims at providing flexible nutritional labels for a dataset to practitioners, showing the distributions of each selected attribute, functional dependencies between attributes, and the maximal uncovered patterns. When a dataset is added to the system, a set of dataset labels that summarize information about the dataset are shown, such as how representative of minorities the data is, how correlated the different attributes are (especially with respect to the protected attributes, the number of errors (e.g. missing values), etc. In addition to showing such data, its back-end optimizes for the trade-off between the amount of information given (through the widget), and the space the widgets use, by "learning" how preferable each widget is for different tasks based on logs of practitioners' use. Additionally, MithraCoverage [88] allows interaction with aforementioned coverage methods, e.g. to filter out the invalid patterns, but also to fix the parameters of the method such as the coverage threshold, or the attributes the practitioner wants to investigate particularly.

7.2.3 Unbiased query results

Most previously presented works focus on retrieving a fair or diverse set of data tuples from a single dataset. Orr et al. [131] adopt a different setup and problem. They assume that existing databases are biased in a sense that they might not accurately reflect the world distributions of samples, and that practitioners can have additional access to aggregate datasets which contain information that might reflect the real distributions. From this new framing of the bias problem, they propose Themis, a framework that takes as input the original dataset, the aggregate dataset, and a practitioner's query, and outputs results that are automatically debiased by learning a population's probabilistic model and reweighting samples accordingly. This is the first work in the area of open-world databases that aims at debiasing query results in that sense of bias.

7.3 Mitigation

Mitigation methods focus on modifying datasets, e.g. for classification tasks [162, 184, 103], or ranking tasks [14, 103, 66]. Most methods are seen as data repair methods where the tuples or labels are modified, and would merit being unified with other data cleaning methods as their application might influence unfairness [184].

We identify three main trends in mitigation methods, that focus either on data or feature representations. Data works consist in transforming data for classification tasks by relying on causality notions, or in tackling the problem of retrieving fair, possibly ranked, data subsets. feature representation works aim at learning data representations for which the outputs of classification tasks are fair. We further explain these three trends below.

7.3.1 Dataset de-biasing through causality

Salimi et al. [159] focus on causal algorithmic fairness –a recent topic emerging in several research domains. They outline research directions and present how data management methods such as query rewriting, dataset repairing, data provenance, and weak supervision algorithmic models to fairly label data could be applied to mitigate dataset biases with a causal sense. While causal fairness is argued to better reflect human notions of fairness for instance by accounting for disparities due to relevant attributes, it is currently hard to use this formalisation of fairness for measurement and mitigation because they require knowing the causal graph of the dataset –which is typically not available.

In [160], causality analysis is adopted to rely on the availability of the causal graph to mitigate biases within the datasets, accounting for admissible biases. The authors note that the causal fairness constraints that ask for the absence of edges in the graph between certain nodes are equivalent to independence conditions between attributes, and that ensuring fairness could be seen as ensuring such independence. Hence, they propose to rely on existing works on integrity constraints,

e.g. multivalues dependencies, which are closely related to this idea, and frame dataset fairness mitigation as a database repair problem for these dependencies. The algorithm they develop, Capuchin, inserts new samples in the database to ensure the independence between protected attributes and target labels for any direct paths, except for the ones with attributes that a practitioner would define as admissible.

While most works assume all data to be in a single table, Salimi et al. [161] also adapt previous works around causality to the context of relational databases since the prior formalisation can not directly apply there. They propose a declarative language –CaRL: Causal Relational Language– that allows them to represent their relational data into a causal paradigm and specify the potential causal dependencies between attributes. With this, they also propose a method to answer causal queries formulated within the language that practitioners would pose.

7.3.2 Diversity in sets and rankings

Some works investigate algorithms to retrieve fair data such as group-fair and diverse set selection [179] or ranking [12, 203], group fair recommendations in the health domain [180], or to fairly allocate public resources [23]. Such notions of fairness are primarily associated with the notion of *diversity* in the data management community [50], the idea that "different kinds of objects are represented in the output of an algorithmic process". In certain cases, the problem extends to identifying several sets of diverse items where the items across sets are different (termed aggregate diversity), such as for recommender systems where the recommended items should be diverse across users not to recommend always the same items as certain item publishers would otherwise not be appearing in the systems.

In their survey [50], the authors explain the different formalisations of diversity through metrics, and the different algorithms existing to return diverse sets. They note that diversity usually comes hand in hand with the notion of utility. For instance, in the context of hiring, the candidates to select should both be "useful" to the hiring entity and diverse for example to avoid structural bias.

Variations of the problems of rankings and set selection are explored. The difference between *diversity* and certain notions of *fairness* is discussed in [50], and is based on that fairness in certain cases means that the algorithmic system represents objects or individuals in proportions equal to the input data, and these proportions might not necessarily be reflecting diversity in the objects or individuals. Yang et al. [199] further highlight the difference by identifying the trade-off that can arise between utility, diversity and fairness in certain contexts such as hiring. Selecting a set of candidates to hire that maximizes utility constrained over diversity might not lead to selecting the best candidates for each protected group or intersectional protected group, which could be considered unfair within each group. In response to that, they propose new in-group fairness constraints to integrate to the set selection problem, and formulate the optimization task into integer linear programs to solve it.

Asudeh et al. [13] take a different view on the problem and focus on the task of ranking items by assigning weights to each attribute characterizing the items, allowing to compute a score for each item. They highlight the stability issue that this formalization might encounter –similar weight assignments might lead to different rankings–, and propose a measure of ranking stability as well as algorithms for providing stable rankings by leveraging the geometric properties of weights and rankings. Interestingly, Yang et al. [201] propose to integrate the causality approach in order to identify intersectionally fair rankings. Kuhlman et Rundensteiner [101] tackle a variation of the ranking problem, considering that multiple stakeholders provide individual rankings, and these rankings must be combined while conserving group fairness notions for the items. They propose a formalisation for this problem and algorithms with guarantees to identify optimal fair ranks.

In contrast, Chen et al. [38] focus on spatial allocation tasks, where resource items placed in a space should be allocated fairly to different individuals. They propose a formalisation of fairness in this context, and objective functions that integrate both fairness and "convenience" concerns (e.g. minimum traveling distance) for the individuals, as well as algorithms to achieve such task.

7.3.3 Representations

In contrast, some works are interested in feature representations and their connection to output fairness. Particularly, Lahoti et al. [103] propose a method to ensure individual fairness by learning fair representations of the data. Yahav et al. [196] are interested in biases found within text features (specifically tf-idf) resulting from biases in text mining datasets due to the context around the datasets samples. These methods do not touch upon the raw samples but their feature representation.

7.4 Crowdsourcing

Unfairness in crowdsourcing is also investigated, similarly as in the other domains studied in the previous sections. Works either look at unfairness towards the crowd workers, such as Borromeo et al. [28] who propose a list of axioms to guide the creation of fair and transparent crowdsourcing processes –task assignment, task completion, and worker compensation–; or look at resolving unwanted biases in labeled data. It is argued that such biases in labels can stem from personal preferences or differing expertise of crowd workers [206], from labeling "trends" [74,122], or from the subjectivity of the object to review in evaluation systems [105].

7.5 Data science workflow

Different from works in the other domains, a few recent works are interested in developing tools at the intersection of data management and machine learning. For instance, Schelter et al. [165] note that the existing tools developed for fairness do not support practitioners (and researchers) fully in developing the whole data science workflow responsibly. Instead, they simply let them apply various fairness metrics and bias mitigation methods without being aware of their interaction with other parts of the workflow such as data cleaning, separation of the datasets into independent training and test sets, etc. They build FairPrep, a framework on top of the existing IBM toolkit AIF360, in order to fill this gap: practitioners input data and their desired preprocessing methods, as well as choose a machine learning algorithm, and the framework automatically processes this information, trains the model and outputs its complete evaluation based on both performance and fairness measures. This allows avoiding errors in building the workflow, such as for instance leaking data information from the training to the test set when handling data errors such as missing values, when engineering features or tuning a model's hyperparameters, etc. Besides, experiments with their framework show the lack of consideration of existing fairness works from the machine learning community for critical data engineering activities such as data cleaning.

With the same idea that the data pipelines might unintentionally inject biases, Yang et al. [200] developed a tool that automatically extracts a directed acyclic graph representation of the data pipelines and data flows from the code of the pipelines, and provides information on the way each vertex impacts the distribution of samples based on protected attributes and target labels. By generating a report with the graph and this information, a practitioner can investigate potential bias issues of its pipelines.

8 Data Management: Research Gaps

In this section, we identify research gaps between data management research on bias and unfairness (section 7), bias and unfairness research in other fields of computer science (section 4), and typical development practices of data-driven decision-support systems. These gaps are summarised in Figure 2. This is the basis for developing a new approach to the issue in the next section.

8.1 Methodology

Approach. To identify these gaps, we first outline a list of all activities performed over the full lifecycle of a data-driven decision-support system, from development to deployment. This list provides us with the basis to reflect on potential research gaps, as it encompasses the necessary set of activities to develop the systems, and these activities are by design both the sources of bias and unfairness and the opportunities to solve these issues. These activities can be associated with one or multiple general unfairness-agnostic research areas, usually stemming from machine learning and data management. For instance, the construction step of a decision-support system consists of building both a data management and a data analytics set-up. Data management activities at this step map to multiple research areas within data management such as data integration or data curation.

Then, we map the research activities that we identified in the previous sections onto the aforementioned mapping. This allows investigating the extent to which the different unfairness-agnostic research and non-research activities are covered by bias and unfairness-related research. In cases where an activity is not covered, it might be because it does not interact with unfairness at all, or because it has not been studied yet. In any case, we analyze it because it could still be useful to resolve certain unfairness issues. Such analysis brings us to identify three main gaps, either related to data management activities for addressing unfairness, or to data management activities that create unfairness, or more generally to whole stages in the lifecycle that have not been thoroughly investigated.

Lifecycle of a data-driven decision-support system *(in orange).* The development process of a data-driven decision-support system is divided into five main

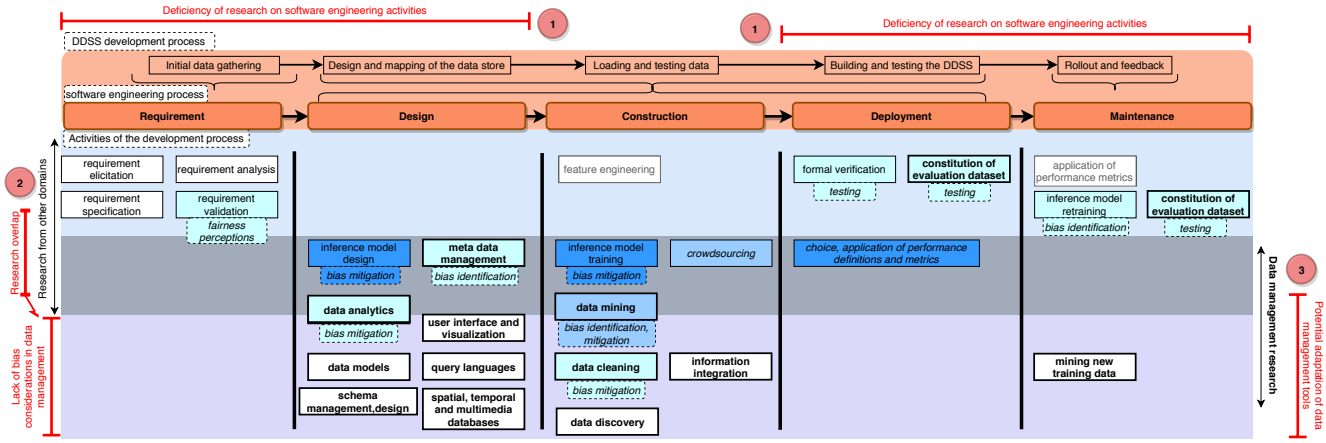

Fig. 2 Activities relevant to bias and their amount of bias-related research (white: no research; light to dark blue from few to plenty of research). Data management activities are bold.

stages as described in [142]: 1) the initial data gathering, 2) the design and mapping of the data store, 3) the loading and testing of the data, 4) the building and testing of the system, and 5) its rollout and inclusion of feedbacks from its users. These stages are easily mapped to the typical software engineering process [29]: 1) requirements engineering, 2) system design, 3) system construction, 4) system testing, and 5) maintenance of the system after deployment. While the description of the lifecyle of the decision-support systems focuses on the distinction between data and other aspects of the system, the software engineering description mostly focuses on the general stages of development.

Activities performed during the lifecycle *(activities placed in boxes, we differentiate between data-related activities in bold, and other ones).* We identify the specific activities performed in each stage of the lifecycle. To do so, software engineering literature [29] indicates the activities which are general to any kind of software. These activities span the requirement engineering stage (requirement elicitation, analysis, specification and validation), the design stage (system and user-interface design), and both the testing and maintenance phase (these last two stages are not detailed for simplicity and because they might not be applied thoroughly yet for the specific case of data-driven decision-support systems).

Data management literature presents activities or topics that are specific to the data aspects of the lifecycle. These are extracted from the common list of research topics in data management venues [3]. For the design stage, we identified data models, query languages, schema management and design, meta-data management, user interface and visualization, data analytics, and specific issues on spatial, temporal, and multimedia databases. For the construction phase, we found data mining, data cleaning, information integration, data discovery, and crowdsourcing.

Additional activities that are specific to machine learning [8] are found in the design stage (inference model design), and in the construction stage where we identify data collection (shown as data mining because of the overlap with data management literature), data labeling (shown as crowdsourcing for the same reason), feature engineering, and inference model training. In the testing stage, only model testing is added. For the maintenance stage, model monitoring and model update are identified. These last two stages are further subdivided. Testing is composed of the choice and application of performance definitions and metrics, the constitution of evaluation datasets (these two are for experimental testing), and the formal verification. For the maintenance phase, we found mining new training data, inference model retraining, application of performance metrics, and constitution of new evaluation datasets, since the context of application of a system might shift or expand, and hence new data must be collected, and the machine learning model must be retrained to account for this shift.

Mapping to current research on bias and unfairness *(colors of boxes).* We map current research on unfairness (from light to dark blue, representing the quantity of current literature on that topic) outlined in sections 4, 7 to these activities (the topics identified in the previous sections are in italic for easy identification). This enables to identify where research is focusing and where it is lacking.

In the following, we explain the findings of this analysis, grouped by their main topics.

[3] List from `https://vldb2020.org/research-track.html`.

8.2 Bias-Aware Requirements

A first observation is that some stages of the development process are more researched than others. Specifically, the design and implementation of inference models are the most covered topics [84], along with metrics or definitions for fairness. There is also a shorter line of work on data mining, mostly focusing on structured data and text data.

In contrast, works on requirement engineering and subsequent database design (elicitation, translation to specifications), system testing, and maintenance (continuous testing with respect to the identified requirements) are much fewer. These limitations are also partly highlighted within the Human-Computer Interaction (HCI) and the Software Engineering communities, as explained in section 5. Yet, many researched methods mostly focus on bias mitigation in the algorithmic part. Hence, developing tools to model, design, and construct better datasets should be a priority.

8.3 Biases in Data Management Activities

A second observation is that for many traditional data management activities which might introduce unwanted biases, there is little to no research investigating their impact on biases at the output of the system. This covers for example data cleaning, data discovery, or data integration [11]. On that note, Stoyanovich et al. [178] encourage the exploration of the possibilities to mitigate biases early in the data life cycle of the decision support systems.

Abiteboul et Stoyanovich [4] further outline that several principles from regulations about responsible data-driven systems, possibly outside the scope of bias and fairness such as the right to "data portability", would require investigation and adaptation of the data management community. For instance, ensuring "the right to be forgotten" for an individual would mean investigating how this right translates in every layer of a database, while accounting for possible dependencies with the data tuples representing this individual and other connected individuals.

Furthermore, we could not identify any significant effort on bias and unfairness considerations in data modeling, schema design, and data provenance topics, even though these activities define the information on which the inference model and decisions are based.

8.4 DBMS Activities for Bias Mitigation

A third observation is that part of the encountered research efforts in data management mirrors the works in other domains on bias and unfairness for data-driven decision-support systems (section 4) with similar approaches and limitations. Especially, there is also a focus on definitions, metrics, and mitigation at the algorithm level. However, further re-purposing or adapting some of the approaches developed in other data management works could serve to identify or mitigate certain biases already in the datasets. This holds especially for data cleaning methods like error detection and data repairing, data analytics and efforts in data modeling, and also research on multimedia data.

Only a small part of current data management research makes use of such methods. The idea of mitigating unwanted biases through data repair methods is similar to those proposed in data mining, but tends to be more general and agnostic with respect to the employed analytic methods as presented by Salimi et al. [159]. Two vision papers are of note on the topic. The first one proposes to unify data pre-processing and inference systems arguing that fairness, accountability, and transparency could be seen as database system issues before applying machine learning and outlining how a platform for data analytics could help solve these issues [177]. On the other hand, Stoyanovich et al. [176] claim that methods to automatically attribute labels to datasets and machine learning models (meta-data) to prevent their misuse are needed to prevent the creation of additional biases.

Asudeh et Jagadish in a tutorial [11] suggest that works around data profiling and provenance could be adapted to fulfill the need of practitioners for tools to explore biases in data. Besides, Abiteboul et Stoyanovich. [4] discuss how various regulations such as the GDPR in Europe advocate for responsible development and use of data and data-driven decision support systems, and make the case there that the data management community could support progress on principles like transparency by adapting existing works for instance on data profiling to better expose the data statistics for a richer interpretation of the systems' outputs.

Orr et al. [130] proposed an in-DBMS method for practitioners to query a database and retrieve results which are automatically cleared from dataset sampling biases introduced during the data collection step. This work is the closest to the approach we advocate in the next section since it aims at helping practitioners to mitigate biases within the database, although it is not made for the purpose of further training a machine learning model.

9 Roadmap for Future Research

In the previous sections, we identified both limitations and gaps stemming from the current approach to tackle unfairness of data-driven decision-support systems, i.e. approaches focused on the machine learning algorithms themselves, and general research gaps stemming from existing data management activities. The main limitations are the difficult application of existing algorithmic methods by practitioners, and the fact that such methods do not allow to build fully fair systems.

In this section, we reflect on a way forward to overcome these limitations. Particularly, the limitations hint at a possible research shift in order to solve existing unfairness issues: not only should we develop algorithms robust to unfairness but also data methods to mitigate unfairness, and practical tools to support and ensure the use of such methods by practitioners. In the next section, we discuss the challenges arising from this way forward.

9.1 Eliciting and Enforcing Fairness Requirements

We advocate focusing on eliciting and enforcing bias and fairness requirements already early in the system design workflow. This allows to clarify the goals of a system in relation to fairness, and then brings the possibility to guide practitioners along the system development cycle to create a system that verifies these goals. Thus, the fairness requirements serve as a foundation of a bias-aware data-engineering pipeline. Here, we outline how such bias and fairness requirements can be applied conceptually and how they integrate into existing database management system architectures.

9.1.1 Proposed workflow

We propose a new workflow for practitioners building data-driven decision-support systems, encouraging fairness-by-design.

Ideally, before designing and building a system, a practitioner would define a list of requirements, including fairness requirements.

These requirements would then be translated into constraints on both the data used for training the system and inputted at deployment time. These constraints would impose statistical conditions with regard to defined protected attributes that would ensure that a dataset could be considered fair for the requirements at hand. At training time, this would increase the likelihood that the outputs of a model trained on such dataset are fair (note: an "unbiased" training dataset does not guarantee an unbiased resulting system since new unwanted biases might arise from the machine learning algorithm used or small unwanted biases in the data might be reinforced by the machine learning model, but helps); while at deployment time, it would monitor whether the predictions made for new data points are fair. Constraints at training and deployment time might differ depending on the initial fairness requirements, the associated characteristics that a training data should bear, and the appropriate slack for such training data characteristics needed to ensure reasonable fairness measures.

Continuous checks of bias constraints on the system's outputs are needed, analogously to continuous testing in software deployment, since the fairness of the system might vary in case a distributional shift happens between the training data and deployment data.

In cases where the data would not follow such constraints, either data curation methods could be employed to remedy such issue at training time, or this would be an indication that it is mathematically impossible to verify simultaneously the multiple fairness requirements and other requirements, and hence the system should not be developed or the requirements should be reviewed. At deployment time, the constraints not being verified would indicate the necessity to defer the decision to a human agent, or the necessity to retrain the model on updated data.

9.1.2 Addressed limitations

This new approach considers the quality of the data as a core issue, contrary to the approach outlined in section 2 coming from most research on algorithms and metrics for the outputs of a machine learning model or for the outputs of other types of inference models. Our intuition is that it would overcome multiple challenges that are typical concerns of different research communities, besides unfairness, and that interact with unfairness considerations: cost, time, robustness and practicality for the machine learning and software engineering communities, societal impact and trust for the human-computer interaction community. They are the following challenges:

Fairness. The main source of biases is data, hence investing research to understand, detect, and control bias in data allows to build less biased datasets with regard to specific fairness requirements and consequently to train fairer systems.

Robustness. Modifying optimization functions of machine learning algorithms or post-processing decisions can have unforeseen effects in cases where the application context and data would change. In contrast, we argue that enforcing inspection of data biases in

the early stages of development and during deployment would result in more robust systems since potential issues would be identified earlier.

Practicality. Practitioners might understand issues and methods in the data stages of the development of a data-driven decision-support system better than those related to the inference model. For example, obtaining extra training data to balance a dataset might be easier than adjusting machine learning algorithms; hence, data-focused tools could be more applicable than current methods. Considering that transfer learning is becoming a common practice (i.e. using pre-trained general models and then fine-tuning them for a specific application), the availability of "unbiased" data for the fine-tuning phase is crucial.

Cost and Time. By ensuring that training data has no bias issues, the resulting trained models will likely behave in a more desirable fashion, thus fewer costly training and retraining cycles are needed to achieve the desired system behavior. Ultimately, the process would be more effective and less costly.

Societal Impact. Establishing requirements would encourage considering societal impact already in the initial stages of development. Past cases which did not explicitly state and enforce their fairness requirements showed the potential negative impact of building these systems without accounting for potential issues: Microsoft's chatbot Tay became racist after its deployment because it was constantly retrained on data fed to it by layman users and had to be shutdown [3], while the automatic CV screening tool of Amazon was shown to be discriminating against women after release [87]. Many of these issues could have been foreseen and mitigated if undesired bias identification and fairness were central design goals of these systems.

Trust and Informed Decision-Making. Finally, by explicitly communicating bias and fairness design goals and validating systems respectively, trust can be facilitated between the system and stakeholders or users who will have a better understanding of its behavior. This can also support building an accurate abstract model of the capabilities of a system. This will lead to better decisions, as the performance of a human decision-maker is dependent on his/her mental models of the problem and of the system and on tools at hand [140].

9.2 Required DBMS Extensions

By shifting the focus from the algorithms to the data, we foresee the need for two new core extensions to database management systems, that would support the application of the proposed workflow.

9.2.1 Bias Data Constraints.

Fairness requirements identified in the requirements elicitation phase need to be formalized such that they can guide the system's development. Furthermore, they need to be validated or verified across the system's lifecycle. New *bias data constraints*, expanding on existing data constraints, could be used to encode and enforce data-related bias requirements.

9.2.2 Bias Curation Methods.

Data curation methods addressing bias by transforming, adding, or removing data instances would be needed in cases where the constraints are violated. While also algorithmic mitigation techniques (see section 4) can be used, we argue that data curation is often more effective or practical [76]. If the constraints are violated, the system designers would be warned to take action or prevented to train the models.

9.2.3 Embedding into the DBMS

In order to support and enforce the use of bias constraints and curation methods, existing database management systems should be extended to integrate them, an idea also suggested in [11]. This will be important as checking bias constraints can be very data-intensive. By embedding this into the database management system, we can take advantage of existing components like indexes or system catalog information, allowing for more efficient implementation. The creation and integration of these components bring a multitude of data management research challenges that we highlight in the next section.

10 Open Research Challenges

Here we highlight the specific research challenges which need to be addressed for realizing the bias and unfairness-mitigating extensions proposed in the previous section.

10.1 Formalization and Modelling Challenges

10.1.1 Bias-Aware Schema Design

While selecting fairness notions for a specific use-case is not an easy task, defining the exact attributes and their allowed values to base the constraints on and the subsequent design of the database schema is also complex. Formally understanding how the granularity and ranges of the values in the database schema influence

performance of the system and measurement of its bias remains to be investigated. For example, let's assume that the loan attribution model should not discriminate against young black men, and that the dataset contains gender and race as categorical attributes and age as an integer. After choosing a fairness definition, deciding how to transform age into a categorical attribute can have direct bias consequences. Defining protected classes (male, black, [10-23]) or (male, black, [10-25]) as protected attributes would both surface and measure different biases. Different mappings of age to its protected class "young" can create different system behaviors: the granularity of the categories chosen would influence both the performance and fairness of the trained inference model. This gets even more complex when the bias constraints are defined over several attributes to transform. Similarly, this transformation might have an impact on the similarity measures used in the constraints for individual fairness since tuples similarity depends on their attributes.

10.1.2 Predicting the Feasibility of a Data-Driven Decision-Support System

At the start of the workflow, determining whether bias constraints can be verified along with other requirements (e.g. accuracy performance, cost, amount of data) and other data constraints before designing and implementing a system would enable to save a great amount of time and computing power, while it would also allow to possibly refine requirements and resources allocated for a system. For instance, in case a practitioner has a specific amount of loan data and wants to build a data-driven decision-support system to automate the decision of giving out a loan, knowing before building the system and training a model that it will not be able to reach a minimum required accuracy and fairness would save efforts. Until now, few theoretical works [97,40] have been proposed that investigate such feasibility of requirements. Existing results focus on the diverse fairness notions that can contradict each other.

Using impossibility results for fairness notions [40], certain impossible scenarios can already be determined analytically. Predicting a measure of each requirement, potentially via simulation through the training of simple inference models could also give empirical indications of the feasibility.

10.1.3 Formalizing the Tensions between Privacy and Fairness

Conflicting orthogonal efforts are put into preserving the privacy of individuals [115,89,49] in the training and test data. This might include aggregating tuples, decreasing the granularity of certain attributes (like the ones used for diversity constraints or the protected attributes, e.g., by collapsing a specific age to an age range that is different from the age ranges chosen for categorizing age in bias constraints), or completely dropping attributes from the view. Common protected attributes for fairness are often also considered sensitive for privacy. Hence, despite good intentions, not having these relevant attributes, classes or tuples creates obstacles to check the bias constraints, whereas biases on these private-sensitive attributes might still exist due to remaining other attributes correlated with the protected ones. Thus, more work on understanding the interactions between privacy and unfairness [85], and on accurately inferring the missing attributes from the available data is needed [44,37]. This would be part of the checking process of the bias constraints.

10.2 Algorithmic Challenges

There exist few bias curation methods from the data mining and machine learning communities, however, they are still limited in scope (e.g. the intersectionality of multiple protected attributes is not usually handled by current methods). More research is needed to establish approximation algorithms that would guarantee bias constraint satisfaction on the training data. These algorithms could transform existing data (like data resampling, data label modification, or variants of database repairing methods [159]) possibly with inspiration from existing data cleaning methods, synthesize new ones, or guide the collection of additional records.

Additionally, nearly all data-driven decision support systems rely on elaborate data engineering pipelines for preparing, transforming, integrating, cleaning, and finally ingesting training data, test data, and live data. Bias curation needs to be integrated within such data engineering pipelines. Also, existing steps of data engineering pipelines might have unforeseen and insufficiently understood consequences and effects on data bias. For instance, cleaning a dataset from its outliers might remove data from the protected minority class and hence a bias curation method would not have access to such data anymore, missing-value imputation methods might skew the dataset towards the protected or non-protected group and hence might add unwanted biases, so new methods would be needed to allow for the application of the bias curation methods, etc.

Only the interaction between bias and data cleaning has received preliminary attention [184,165]. Hence, future work needs to investigate the impact of the previous activities on data biases, and the interaction with

the bias curation methods. This would lead either to providing guidelines on the workflow to follow, or to the creation of new algorithms that would integrate curation and integration or cleaning simultaneously.

10.3 System-Oriented Challenges

Adapting existing mechanisms in database management systems for supporting the bias constraints exhibits multiple challenges. The bias constraints would bear some similarities with existing database constraints, but also differences that would make their implementation and use not straightforward. We develop here the comparison with traditional constraints and highlight foreseen challenges.

10.3.1 Constraint Expression

Translating fairness metrics into SQL constraint language, possibly by additionally using user-defined functions, is the first step and challenge to allow the support of bias constraints. The way to encode these constraints would need to be as flexible as possible to accommodate most definitions of fairness and possibly new ones.

Certain constraints would be specified on protected attributes, other attributes of the data, and possibly on the decision attributes (actual decisions and/or predictions). The exact test of the constraint could cover statistical tests for undesired biases such as unwanted correlations between protected and other attributes or checking for potential "wrong" decision labels (e.g. [153]). For instance, in case fairness towards groups is important, the acceptable data distributions for each protected class can be specified. In many cases, these would be egalitarian distributions [191], but also non-egalitarians constraints could be relevant. For example, an AI-assisted hiring tool might want to positively discriminate against female applicants to address issues with employee diversity.

Inspiration from existing ways to encode data cleaning rules could be taken to express the bias constraints. For instance, denial constraints which are declarative specifications of rules a dataset should respect [41], could be investigated, especially for individual fairness which relies on the similarity between tuples.

10.3.2 Constraint Checking Mechanism

A new set of challenges in order to implement bias constraints efficiently using current database technologies is the result. The use of triggers could be investigated as a tool to check for the constraints.

Because the constraint functions are expensive to compute, an envisioned research direction is to investigate how to incrementally compute the statistics that make the constraints over multiple batches of data, in order to avoid the whole re-computation at each check. Possibly existing system catalog statistics used for query optimization could allow to speed up such computation while reducing the resource consumption.

Bias constraints could be checked when a sufficiently larger number of records has been added or modified. Several policies for monitoring them would be useful: checking for constraint violations after initially populating the database, checking for violations when training data is retrieved for training an inference model, or when adding a large number of training tuples during system maintenance phases, and finally checking for violations when a significant number of new decisions are suggested by the system before accepting them.

Certain fairness metrics require the computation of error rates between ground truth decisions and predicted decisions. Hence, some constraints would also require having predictions made by a machine learning model available –training data in the database by itself would be insufficient.

10.3.3 Frameworks and Tools

Once a bias constraint is violated, tools need to be available to facilitate the use of data bias curation methods. Such methods should be integrated into existing data debugging tools such as Dagger [116]. To our knowledge, only two data-focused systems have been implemented toward this goal – by Ruggieri et al. [154] for discrimination discovery, by Ramadan et al. [148] to uncover unwanted data dependencies through data flow analysis –, and a few frameworks (e.g., AIF360 [22], Aequitas [156]) with fairness metrics and algorithmic mitigation methods not integrated into the entire system lifecycle. However, such frameworks do not allow handling complex cases where protected groups would not be binary and defined on single attributes.

Besides, bias *meta-data* for the data views could be generated to help the communication [76] about potential dataset biases [204,176,60,121].

10.4 Guidance for DBMS Users

As a major practical challenge, we identify the need for guiding a practitioner through the process of specifying fairness requirements and bias constraints. Certain applications might rely on country-specific regulations, while others might not have well-established policies. As there are a plethora of different fairness

definitions, choosing the correct metric and setting the correct parameters is far from trivial due to the abstraction gap between application (fairness as an abstract norm) and constraint model (fairness as a mathematical object). Therefore, we envision a guidance component that could come in form of wizards, or an IDE that can provide suggestions based on data profiling of potential biases and on existing regulations.

A human-in-the-loop approach could highlight these biases, and then from feedback provided by the practitioners about the biases, it could uncover the undesired ones and automatically infer related fairness requirements, bias constraints and their prioritization. User studies could also be conducted to understand the actual difficulties and questions that practitioners would like to address.

Similarly, practitioners could be helped by having guidance frameworks and interfaces for deciding on bias curation methods to apply, that would visualise their impacts on different categories of population and on the other important factors in the requirements (e.g. cost, time, accuracy, etc.).

10.5 Multimedia Data-based Challenges

Applications using multimedia data such as images, texts or videos have typically the same aforementioned challenges, but additional difficulties arise.

For instance, for checking bias constraints, it is difficult to extract protected attributes or other semantically interpretable features from an image or text. Hence, it is both difficult to generate necessary meta-data to apply the constraints, and to generate new representative test cases to check for the constraints. This task is currently performed manually for images and semi-automatically for text which hampers scalability and real-world applicability.

A similar issue arises when curating data for bias. Structured data algorithms would not be easily applicable since no interpretable attributes would be available to reason on. One direction to investigate could be to transform multimedia data into structured representations on which to apply the aforementioned algorithms. Possibly, crowd workers could be asked to annotate protected attributes, to produce or collect new related samples following certain templates (such as in [188]), or new automatic methods like GANs (Generative Adversarial Networks) could be used conditioned on meaningful attributes, in order to generate data with specific meta-data.

11 Conclusion

In this survey, we provided an overview of the state-of-the-art computer science works that address unfairness issues of data-driven decision support systems. While we showed that these works focus primarily on developing definitions and metrics for unfairness, and algorithmic approaches to mitigate this unfairness in the underlying machine learning models, we also observed that there are still only few works emanating from the data management community that exploit existing data management research to approach unfairness. This leads us to highlight research gaps that future data management research could fill, such as investigating how data management activities like data integration, data discovery, data cleaning might create or reinforce data biases that would result in algorithmic unfairness.

We then took a step back from the current machine learning-centered approaches (which are typically hard to apply in real-world scenarios). We argued for a new data-centered approach that would mitigate these higher-level challenges. Eliciting data requirements and enforcing them through the extension of database management systems with bias constraints and bias curation methods would reduce the spread of unfairness in the outputs and possibly ensure better monitoring of potential biases both before and after deployment of the systems. Furthermore, by making such constraints explicit already in early development phases, many common pitfalls and issues could be avoided by simply having a higher degree of awareness and planning during development.

Realizing such approaches, however, presents novel data management research challenges. New algorithmic solutions, formalisations, and modelling informed by theory and also system- and user-oriented research need to be considered to allow for building database management systems that ensure fairness in the outputs of later trained machine learning models and the systems using such models.

Acknowledgements

This work has been partially supported by the *HyperEdge Sensing* project, funded by Cognizant.

References

1. : HireVue.com . `https://www.hirevue.com/` (Jan. 2020)
2. Practitioners guide to compas. Tech. rep., Northpointe (2012)
3. James Vincent : Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a

day. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (Feb. 2020)
4. Abiteboul, S., Stoyanovich, J.: Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. Journal of Data and Information Quality (JDIQ) **11**(3), 1–9 (2019)
5. Accinelli, C., Minisi, S., Catania, B.: Coverage-based rewriting for data preparation. In: EDBT/ICDT Workshops (2020)
6. Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D.: Black box fairness testing of machine learning models. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2019, pp. 625–635. ACM, New York, NY, USA (2019). DOI 10.1145/3338906.3338937. URL http://doi.acm.org/10.1145/3338906.3338937
7. Albarghouthi, A., D'Antoni, L., Drews, S., Nori, A.V.: Fairsquare: probabilistic verification of program fairness. Proceedings of the ACM on Programming Languages **1**(OOPSLA), 80 (2017)
8. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T.: Software engineering for machine learning: A case study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 291–300. IEEE (2019)
9. Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure (2019)
10. Angell, R., Johnson, B., Brun, Y., Meliou, A.: Themis: Automatically testing software for discrimination. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2018, pp. 871–875. ACM, New York, NY, USA (2018). DOI 10.1145/3236024.3264590. URL http://doi.acm.org/10.1145/3236024.3264590
11. Asudeh, A., Jagadish, H.: Fairly evaluating and scoring items in a data set. Proceedings of the VLDB Endowment **13**(12), 3445–3448 (2020)
12. Asudeh, A., Jagadish, H., Stoyanovich, J.: Towards responsible data-driven decision making in score-based systems. Data Engineering p. 76 (2019)
13. Asudeh, A., Jagadish, H.V., Miklau, G., Stoyanovich, J.: On obtaining stable rankings. Proc. VLDB Endow. **12**(3), 237–250 (2018). DOI 10.14778/3291264.3291269. URL https://doi-org.tudelft.idm.oclc.org/10.14778/3291264.3291269
14. Asudeh, A., Jagadish, H.V., Stoyanovich, J., Das, G.: Designing fair ranking schemes. In: Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, pp. 1259–1276. ACM, New York, NY, USA (2019). DOI 10.1145/3299869.3300079. URL http://doi.acm.org/10.1145/3299869.3300079
15. Asudeh, A., Jin, Z., Jagadish, H.: Assessing and remedying coverage for a given dataset. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp. 554–565. IEEE (2019)
16. Aydemir, F.B., Dalpiaz, F.: A roadmap for ethics-aware software engineering. In: Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018, pp. 15–21 (2018). DOI 10.1145/3194770.3194778. URL https://doi.org/10.1145/3194770.3194778
17. Balayn, A., Mavridis, P., Bozzon, A., Timmermans, B., Szlávik, Z.: Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations (short paper). In: Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018., pp. 67–71 (2018). URL http://ceur-ws.org/Vol-2276/paper7.pdf
18. Barabas, C., Virza, M., Dinakar, K., Ito, J., Zittrain, J.: Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In: S.A. Friedler, C. Wilson (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *Proceedings of Machine Learning Research*, vol. 81, pp. 62–76. PMLR, New York, NY, USA (2018). URL http://proceedings.mlr.press/v81/barabas18a.html
19. Barbosa, N.a.M., Chen, M.: Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pp. 543:1–543:12. ACM, New York, NY, USA (2019). DOI 10.1145/3290605.3300773. URL http://doi.acm.org/10.1145/3290605.3300773
20. Barlas, P., Kleanthous, S., Kyriakou, K., Otterbacher, J.: What makes an image tagger fair? In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '19, pp. 95–103. ACM, New York, NY, USA (2019). DOI 10.1145/3320435.3320442. URL http://doi.acm.org/10.1145/3320435.3320442
21. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. NIPS Tutorial (2017)
22. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)
23. Benabbou, N., Chakraborty, M., Zick, Y.: Fairness and diversity in public resource allocation problems. Data Engineering p. 64 (2019)
24. Benthall, S., Haynes, B.D.: Racial categories in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 289–298. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287575. URL http://doi.acm.org/10.1145/3287560.3287575
25. Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pp. 2212–2220 (2019). DOI 10.1145/3292500.3330745. URL https://doi.org/10.1145/3292500.3330745
26. Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: S.A. Friedler, C. Wilson (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *Proceedings of Machine Learning Research*, vol. 81, pp. 149–159. PMLR, New York, NY, USA (2018). URL http://proceedings.mlr.press/v81/binns18a.html
27. Bird, C., Bachmann, A., Aune, E., Duffy, J., Bernstein, A., Filkov, V., Devanbu, P.: Fair and balanced?: Bias

in bug-fix datasets. In: Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering, ESEC/FSE '09, pp. 121–130. ACM, New York, NY, USA (2009). DOI 10.1145/1595696.1595716. URL http://doi.acm.org/10.1145/1595696.1595716
28. Borromeo, R.M., Laurent, T., Toyama, M., Amer-Yahia, S.: Fairness and transparency in crowdsourcing. In: Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017., pp. 466–469 (2017). DOI 10.5441/002/edbt.2017.46. URL https://doi.org/10.5441/002/edbt.2017.46
29. Bourque, P., Fairley, R.E., et al.: Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0. IEEE Computer Society Press (2014)
30. Brun, Y., Meliou, A.: Software fairness. In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2018, pp. 754–759. ACM, New York, NY, USA (2018). DOI 10.1145/3236024.3264838. URL http://doi.acm.org/10.1145/3236024.3264838
31. Bulletin of the Technical Committee on Data Engineering, IEEE Computer Society: Special Issue on Fairness, Diversity, and Transparency in Data Systems, Vol. 42 No. 3 (2019). Available at: http://sites.computer.org/debull/A19sept/A19SEPT-CD.pdf (Feb. 2020)
32. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp. 77–91 (2018)
33. Burke, R.: Multisided fairness for recommendation. arXiv preprint arXiv:1707.00093 (2017)
34. Calikli, G., Bener, A., Arslan, B.: An analysis of the effects of company culture, education and experience on confirmation bias levels of software developers and testers. In: Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering - Volume 2, ICSE '10, pp. 187–190. ACM, New York, NY, USA (2010). DOI 10.1145/1810295.1810326. URL http://doi.acm.org/10.1145/1810295.1810326
35. Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., Smith, A.: From soft classifiers to hard decisions: How fair can we be? In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 309–318. ACM (2019)
36. Carter, A.: Cathy o'neil (2016) weapons of math destruction: How big data increases inequality and threatens democracy, new york, st. martin's press and virginia eubanks (2018) automating inequality: How high-tech tools profile, police, and punish the poor, new york, broadway books (2018)
37. Chen, J., Kallus, N., Mao, X., Svacha, G., Udell, M.: Fairness under unawareness: Assessing disparity when protected class is unobserved. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 339–348. ACM (2019)
38. Chen, Z., Cheng, P., Chen, L., Lin, X., Shahabi, C.: Fair task assignment in spatial crowdsourcing. Proc. VLDB Endow. **13**(12), 2479–2492 (2020). DOI 10.14778/3407790.3407839. URL https://doi-org.tudelft.idm.oclc.org/10.14778/3407790.3407839
39. Cho, J., Roy, S., Adams, R.: Page quality: In search of an unbiased web ranking. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005, pp. 551–562 (2005). DOI 10.1145/1066157.1066220. URL https://doi.org/10.1145/1066157.1066220
40. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
41. Chu, X., Ilyas, I.F., Papotti, P.: Holistic data cleaning: Putting violations into context. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 458–469. IEEE (2013)
42. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 797–806. ACM (2017)
43. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, pp. 797–806 (2017). DOI 10.1145/3097983.3098095. URL https://doi.org/10.1145/3097983.3098095
44. Coston, A., Ramamurthy, K.N., Wei, D., Varshney, K.R., Speakman, S., Mustahsan, Z., Chakraborty, S.: Fair transfer learning with missing protected attributes. In: Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, USA (2019)
45. Daniel Abadi, Anastasia Ailamaki, David Andersen, Peter Bailis, Magdalena Balazinska, Philip Bernstein, Peter Boncz, Surajit Chaudhuri, Alvin Cheung, AnHai Doan, Luna Dong, Michael J. Franklin, Juliana Freire, Alon Halevy, Joseph M. Hellerstein, Stratos Idreos, Donald Kossmann, Tim Kraska, Sailesh Krishnamurthy, Volker Markl, Sergey Melnik, Tova Milo, C. Mohan, Thomas Neumann, Beng Chin Ooi, Fatma Ozcan, Jignesh Patel, Andrew Pavlo, Raluca Popa, Raghu Ramakrishnan, Christopher Ré, Michael Stonebraker and Dan Suciu: The Seattle Report on Database Research (2020). Available at: https://sigmodrecord.org/2020/02/12/the-seattle-report-on-database-research/ (Feb. 2020)
46. Das, M., Hecht, B., Gergle, D.: The gendered geography of contributions to openstreetmap: Complexities in self-focus bias. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pp. 563:1–563:14. ACM, New York, NY, USA (2019). DOI 10.1145/3290605.3300793. URL http://doi.acm.org/10.1145/3290605.3300793
47. Diaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D.: Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp. 412:1–412:14. ACM, New York, NY, USA (2018). DOI 10.1145/3173574.3173986. URL http://doi.acm.org/10.1145/3173574.3173986
48. Dingler, T., Choudhury, A., Kostakos, V.: Biased bots: Conversational agents to overcome polarization. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, UbiComp '18, pp. 1664–1668. ACM, New York, NY, USA (2018). DOI 10.1145/3267305.3274189. URL http://doi.acm.org/10.1145/3267305.3274189
49. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the twenty-second

ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 202–210 (2003)
50. Drosou, M., Jagadish, H., Pitoura, E., Stoyanovich, J.: Diversity in big data: A review. Big data **5**(2), 73–84 (2017)
51. Dulhanty, C., Wong, A.: Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. arXiv preprint arXiv:1905.01347 (2019)
52. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226. ACM (2012)
53. Elbassuoni, S., Amer-Yahia, S., Atie, C.E., Ghizzawi, A., Oualha, B.: Exploring fairness of ranking in online job marketplaces. In: Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019, pp. 646–649 (2019). DOI 10.5441/002/edbt.2019.77. URL `https://doi.org/10.5441/002/edbt.2019.77`
54. Faltings, B., Jurca, R., Pu, P., Tran, B.D.: Incentives to counter bias in human computation. In: Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA (2014). URL `http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8945`
55. Farnadi, G., Babaki, B., Getoor, L.: Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, p. 108–114. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3278721.3278733. URL `https://doi-org.tudelft.idm.oclc.org/10.1145/3278721.3278733`
56. Farnadi, G., Babaki, B., Getoor, L.: A declarative approach to fairness in relational domains. Data Engineering p. 36 (2019)
57. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, pp. 259–268 (2015). DOI 10.1145/2783258.2783311. URL `https://doi.org/10.1145/2783258.2783311`
58. Ferryman, K., Pitcan, M.: Fairness in precision medicine. Data & Society (2018)
59. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: Testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, pp. 498–510. ACM, New York, NY, USA (2017). DOI 10.1145/3106237.3106277. URL `http://doi.acm.org/10.1145/3106237.3106277`
60. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets
61. German, D.M., Robles, G., Poo-Caamaño, G., Yang, X., Iida, H., Inoue, K.: "was my contribution fairly reviewed?": A framework to study the perception of fairness in modern code reviews. In: Proceedings of the 40th International Conference on Software Engineering, ICSE '18, pp. 523–534. ACM, New York, NY, USA (2018). DOI 10.1145/3180155.3180217. URL `http://doi.acm.org/10.1145/3180155.3180217`
62. Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pp. 2221–2231 (2019). DOI 10.1145/3292500.3330691. URL `https://doi.org/10.1145/3292500.3330691`
63. Ghizzawi, A., Marinescu, J., Elbassuoni, S., Amer-Yahia, S., Bisson, G.: Fairank: An interactive system to explore fairness of ranking in online job marketplaces. In: Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019, pp. 582–585 (2019). DOI 10.5441/002/edbt.2019.61. URL `https://doi.org/10.5441/002/edbt.2019.61`
64. Glymour, B., Herington, J.: Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 269–278. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287573. URL `http://doi.acm.org/10.1145/3287560.3287573`
65. Grappiolo, C., Martínez, H.P., Yannakakis, G.N.: Validating generic metrics of fairness in game-based resource allocation scenarios with crowdsourced annotations. In: Transactions on Computational Intelligence XIII, pp. 176–200. Springer (2014)
66. Guan, Y., Asudeh, A., Mayuram, P., Jagadish, H.V., Stoyanovich, J., Miklau, G., Das, G.: Mithraranking: A system for responsible ranking design. In: Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, pp. 1913–1916. ACM, New York, NY, USA (2019). DOI 10.1145/3299869.3320244. URL `http://doi.acm.org/10.1145/3299869.3320244`
67. Guerra, P.H.C., Veloso, A., Jr., W.M., Almeida, V.A.F.: From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011, pp. 150–158 (2011). DOI 10.1145/2020408.2020438. URL `https://doi.org/10.1145/2020408.2020438`
68. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 2125–2126 (2016)
69. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering **25**(7), 1445–1459 (2012)
70. Hajian, S., Domingo-Ferrer, J., Farràs, O.: Generalization-based privacy preservation and discrimination prevention in data publishing and mining. Data Min. Knowl. Discov. **28**(5-6), 1158–1188 (2014). DOI 10.1007/s10618-014-0346-1. URL `https://doi.org/10.1007/s10618-014-0346-1`
71. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination- and privacy-aware patterns. Data Min. Knowl. Discov. **29**(6), 1733–1782 (2015). DOI 10.1007/s10618-014-0393-7. URL `https://doi.org/10.1007/s10618-014-0393-7`
72. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp. 3315–3323 (2016)
73. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias

in captioning models. In: European Conference on Computer Vision, pp. 793–811. Springer (2018)
74. Hernández-González, J., Inza, I., Lozano, J.A.: A note on the behavior of majority voting in multi-class domains with biased annotators. IEEE Transactions on Knowledge and Data Engineering **31**(1), 195–200 (2018)
75. Herranz, L., Jiang, S., Li, X.: Scene recognition with cnns: objects, scales and dataset bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 571–579 (2016)
76. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., Wallach, H.: Improving fairness in machine learning systems: What do industry practitioners need? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pp. 600:1–600:16. ACM, New York, NY, USA (2019). DOI 10.1145/3290605.3300830. URL http://doi.acm.org/10.1145/3290605.3300830
77. Holstein, T., Dodig-Crnkovic, G.: Avoiding the intrinsic unfairness of the trolley problem. In: Proceedings of the International Workshop on Software Fairness, FairWare '18, pp. 32–37. ACM, New York, NY, USA (2018). DOI 10.1145/3194770.3194772. URL http://doi.acm.org/10.1145/3194770.3194772
78. Hu, X., Wang, H., Dube, S., Vegesana, A., Yu, K., Lu, Y.H., Yin, M.: Discovering biases in image datasets with the crowd
79. Hu, X., Wang, H., Dube, S., Vegesana, A., Yu, K., Lu, Y.H., Yin, M.: Discovering biases in image datasets with the crowd. In: Proceedings of HCOMP 2019 (2019)
80. Hussain, W., Mougouei, D., Whittle, J.: Integrating social values into software design patterns. In: Proceedings of the International Workshop on Software Fairness, FairWare '18, pp. 8–14. ACM, New York, NY, USA (2018). DOI 10.1145/3194770.3194777. URL http://doi.acm.org/10.1145/3194770.3194777
81. Hutchinson, B., Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 49–58. ACM (2019)
82. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 49–58. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287600. URL http://doi.acm.org/10.1145/3287560.3287600
83. Imtiaz, N., Middleton, J., Chakraborty, J., Robson, N., Bai, G., Murphy-Hill, E.R.: Investigating the effects of gender bias on github. In: Proceedings of the 41st International Conference on Software Engineering, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019, pp. 700–711 (2019). DOI 10.1109/ICSE.2019.00079. URL https://doi.org/10.1109/ICSE.2019.00079
84. Jagadish, H., Bonchi, F., Eliassi-Rad, T., Getoor, L., Gummadi, K., Stoyanovich, J.: The responsibility challenge for data. In: Proceedings of the 2019 International Conference on Management of Data, pp. 412–414 (2019)
85. Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., Ullman, J.: Differentially private fair learning. In: International Conference on Machine Learning, pp. 3000–3008 (2019)
86. Jannach, D., Kamehkhosh, I., Bonnin, G.: Biases in automated music playlist generation: A comparison of next-track recommending techniques. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP '16, pp. 281–285. ACM, New York, NY, USA (2016). DOI 10.1145/2930238.2930283. URL http://doi.acm.org/10.1145/2930238.2930283
87. Jeffrey Dastin: Amazon scraps secret AI recruiting tool that showed bias against women . https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (Jan. 2020)
88. Jin, Z., Xu, M., Sun, C., Asudeh, A., Jagadish, H.V.: Mithracoverage: A system for investigating population bias for intersectional fairness. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, SIGMOD '20, p. 2721–2724. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3318464.3384689. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3318464.3384689
89. Johnson, N., Near, J.P., Song, D.: Towards practical differential privacy for sql queries. Proceedings of the VLDB Endowment **11**(5), 526–539 (2018)
90. Kamar, E., Kapoor, A., Horvitz, E.: Identifying and accounting for task-dependent bias in crowdsourcing. In: Third AAAI Conference on Human Computation and Crowdsourcing (2015)
91. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Model-based and actual independence for fairness-aware classification. Data Min. Knowl. Discov. **32**(1), 258–286 (2018). DOI 10.1007/s10618-017-0534-x. URL https://doi.org/10.1007/s10618-017-0534-x
92. Karako, C., Manggala, P.: Using image fairness representations in diversity-based re-ranking for recommendations. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP '18, pp. 23–28. ACM, New York, NY, USA (2018). DOI 10.1145/3213586.3226206. URL http://doi.acm.org/10.1145/3213586.3226206
93. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning, pp. 2564–2572. PMLR (2018)
94. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the damage of dataset bias. In: European Conference on Computer Vision, pp. 158–171. Springer (2012)
95. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 656–666 (2017)
96. Kim, J., Ryu, H., Kim, H.: To be biased or not to be: Choosing between design fixation and design intentionality. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, pp. 349–354. ACM, New York, NY, USA (2013). DOI 10.1145/2468356.2468418. URL http://doi.acm.org/10.1145/2468356.2468418
97. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2017)
98. Kobren, A., Saha, B., McCallum, A.: Paper matching with local fairness constraints. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pp. 1247–1257

(2019). DOI 10.1145/3292500.3330899. URL https://doi.org/10.1145/3292500.3330899
99. Koene, A., Dowthwaite, L., Seth, S.: Ieee p7003™ standard for algorithmic bias considerations: Work in progress paper. In: Proceedings of the International Workshop on Software Fairness, FairWare '18, pp. 38–41. ACM, New York, NY, USA (2018). DOI 10.1145/3194770.3194773. URL http://doi.acm.org/10.1145/3194770.3194773
100. Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
101. Kuhlman, C., Rundensteiner, E.: Rank aggregation algorithms for fair consensus. Proceedings of the VLDB Endowment **13**(12), 2706–2719 (2020)
102. Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in Neural Information Processing Systems 30 (NIPS 2017) pre-proceedings **30** (2017)
103. Lahoti, P., Gummadi, K.P., Weikum, G.: ifair: Learning individually fair data representations for algorithmic decision making. In: 35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019, pp. 1334–1345 (2019). DOI 10.1109/ICDE.2019.00121. URL https://doi.org/10.1109/ICDE.2019.00121
104. Lappas, T., Terzi, E.: Toward a fair review-management system. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II, pp. 293–309 (2011). DOI 10.1007/978-3-642-23783-6_19. URL https://doi.org/10.1007/978-3-642-23783-6_19
105. Lauw, H.W., Lim, E.P., Wang, K.: Bias and controversy in evaluation systems. IEEE Transactions on Knowledge and Data Engineering **20**(11), 1490–1504 (2008)
106. Leavy, S.: Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In: 2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering, GE@ICSE, Gothenburg, Sweden, May 28, 2018, pp. 14–16 (2018). URL http://ieeexplore.ieee.org/document/8452744
107. Lee, M.K., Baykal, S.: Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, pp. 1035–1048. ACM, New York, NY, USA (2017). DOI 10.1145/2998181.2998230. URL http://doi.acm.org/10.1145/2998181.2998230
108. Lee, M.K., Kim, J.T., Lizarondo, L.: A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pp. 3365–3376. ACM, New York, NY, USA (2017). DOI 10.1145/3025453.3025884. URL http://doi.acm.org/10.1145/3025453.3025884
109. Leitão, R., Jakobsen, F.: A survey on user-interface design strategies to address online bias. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18, pp. LBW084:1–LBW084:6. ACM, New York, NY, USA (2018). DOI 10.1145/3170427.3188567. URL http://doi.acm.org/10.1145/3170427.3188567
110. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 513–528 (2018)
111. Li, Y., Shi, C., Zhao, H., Zhuang, F., Wu, B.: Aspect mining with rating bias. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II, pp. 458–474 (2016). DOI 10.1007/978-3-319-46227-1_29. URL https://doi.org/10.1007/978-3-319-46227-1_29
112. Liao, Q.V., Fu, W.T., Strohmaier, M.: #snowden: Understanding biases introduced by behavioral differences of opinion groups on social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 3352–3363. ACM, New York, NY, USA (2016). DOI 10.1145/2858036.2858422. URL http://doi.acm.org/10.1145/2858036.2858422
113. Lin, Y., Guan, Y., Asudeh, A., Jagadish, H.: Identifying insufficient data coverage in databases with multiple relations. Proceedings of the VLDB Endowment **13**(12), 2229–2242 (2020)
114. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 502–510. ACM (2011)
115. Machanavajjhala, A., He, X., Hay, M.: Differential privacy in the wild: A tutorial on current practices & open challenges. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1727–1730 (2017)
116. Madden, S., Ouzzani, M., Tang, N., Stonebraker, M.: Dagger: A data (not code) debugger. In: CIDR (2020)
117. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Fairness through causal awareness: Learning causal latent-variable models for biased data. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 349–358. ACM (2019)
118. Margarita Boyarskaya, P.I.: Fair payments in adaptive voting. In: Proceedings of HCOMP 2019 (2019)
119. Matsushita, Y., Lin, S., Kang, S.B., Shum, H.Y.: Estimating intrinsic images from image sequences with biased illumination. In: European Conference on Computer Vision, pp. 274–286. Springer (2004)
120. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 220–229 (2019)
121. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019, pp. 220–229 (2019). DOI 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596
122. Mookerjee, V.S.: Debiasing training data for inductive expert system construction. IEEE Transactions on Knowledge and Data Engineering **13**(3), 497–512 (2001)
123. Morgan, J.S., Lampe, C., Shafiq, M.Z.: Is news sharing on twitter ideologically biased? In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13, pp. 887–896. ACM, New York,

NY, USA (2013). DOI 10.1145/2441776.2441877. URL http://doi.acm.org/10.1145/2441776.2441877
124. Moskovitch, Y., Jagadish, H.: Countata: dataset labeling using pattern counts. Proceedings of the VLDB Endowment **13**(12), 2829–2832 (2020)
125. Mouzannar, H., Ohannessian, M.I., Srebro, N.: From fair decision making to social equality. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 359–368. ACM (2019)
126. Mouzannar, H., Ohannessian, M.I., Srebro, N.: From fair decision making to social equality. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 359–368. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287599. URL http://doi.acm.org/10.1145/3287560.3287599
127. Narwal, V., Salih, M.H., Lopez, J.A., Ortega, A., O'Donovan, J., Höllerer, T., Savage, S.: Automated assistants to identify and prompt action on visual news bias. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA '17, pp. 2796–2801. ACM, New York, NY, USA (2017). DOI 10.1145/3027063.3053227. URL http://doi.acm.org/10.1145/3027063.3053227
128. Official Journal of the European Union: REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation . https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (Jan. 2020)
129. Olteanu, A., Kıcıman, E., Castillo, C.: A critical review of online social data: Biases, methodological pitfalls, and ethical boundaries. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, p. 785–786. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3159652.3162004. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3159652.3162004
130. Orr, L., Ainsworth, S., Cai, W., Jamieson, K., Balazinska, M., Suciu, D.: Mosaic: A sample-based database system for open world query processing. In: CIDR (2020)
131. Orr, L., Balazinska, M., Suciu, D.: Sample debiasing in the themis open world database system. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp. 257–268 (2020)
132. Otterbacher, J.: Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 1955–1964. ACM, New York, NY, USA (2015). DOI 10.1145/2702123.2702151. URL http://doi.acm.org/10.1145/2702123.2702151
133. Otterbacher, J.: Social cues, social biases: stereotypes in annotations on people images. In: Sixth AAAI Conference on Human Computation and Crowdsourcing (2018)
134. Otterbacher, J., Barlas, P., Kleanthous, S., Kyriakou, K.: How do we talk about other people? group (un)fairness in natural language image descriptions. In: HCOMP 2019 (2019)
135. Panda, R., Zhang, J., Li, H., Lee, J.Y., Lu, X., Roy-Chowdhury, A.K.: Contemplating visual emotions: Understanding and overcoming dataset bias. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 579–595 (2018)
136. Park, S., Kang, S., Chung, S., Song, J.: Newscube: Delivering multiple aspects of news to mitigate media bias. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, pp. 443–452. ACM, New York, NY, USA (2009). DOI 10.1145/1518701.1518772. URL http://doi.acm.org/10.1145/1518701.1518772
137. Passi, S., Barocas, S.: Problem formulation and fairness. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 39–48. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287567. URL http://doi.acm.org/10.1145/3287560.3287567
138. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 560–568. ACM (2008)
139. Peng, A., Nushi, B., Kiciman, E., Inkpen, K., Suri, S., Kamar, E.: What you see is what you get? the impact of representation criteria on human bias in hiring. HCOMP 2019 **abs/1909.03567** (2019). URL http://arxiv.org/abs/1909.03567
140. Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., Kamar, E.: What you see is what you get? the impact of representation criteria on human bias in hiring. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, pp. 125–134 (2019)
141. Pérez-Suay, A., Laparra, V., Mateo-Garcia, G., Muñoz-Marí, J., Gómez-Chova, L., Camps-Valls, G.: Fair kernel learning. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I, pp. 339–355 (2017). DOI 10.1007/978-3-319-71249-9_21. URL https://doi.org/10.1007/978-3-319-71249-9_21
142. Power, D.J.: Decision support systems: concepts and resources for managers. Greenwood Publishing Group (2002)
143. Power, D.J.: Understanding data-driven decision support systems. Information Systems Management **25**(2), 149–154 (2008)
144. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8227–8236 (2019)
145. Quattrone, G., Capra, L., De Meo, P.: There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, pp. 1021–1032. ACM, New York, NY, USA (2015). DOI 10.1145/2675133.2675235. URL http://doi.acm.org/10.1145/2675133.2675235
146. Rahman, F., Posnett, D., Herraiz, I., Devanbu, P.: Sample size vs. bias in defect prediction. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2013, pp. 147–157. ACM, New York, NY, USA (2013). DOI 10.1145/2491411.2491418. URL http://doi.acm.org/10.1145/2491411.2491418
147. Raji, I.D., Buolamwini, J.: Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In: AAAI/ACM Conf. on AI Ethics and Society, vol. 1 (2019)
148. Ramadan, Q., Ahmadian, A.S., Strüber, D., Jürjens, J., Staab, S.: Model-based discrimination analysis: A position paper. In: Proceedings of the International

Workshop on Software Fairness, FairWare '18, pp. 22–28. ACM, New York, NY, USA (2018). DOI 10.1145/3194770.3194775. URL http://doi.acm.org/10.1145/3194770.3194775

149. Rastogi, A.: Do biases related to geographical location influence work-related decisions in github? In: Proceedings of the 38th International Conference on Software Engineering Companion, ICSE '16, pp. 665–667. ACM, New York, NY, USA (2016). DOI 10.1145/2889160.2891035. URL http://doi.acm.org/10.1145/2889160.2891035
150. Robertson, R.E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., Wilson, C.: Auditing partisan audience bias within google search. Proc. ACM Hum.-Comput. Interact. **2**(CSCW), 148:1–148:22 (2018). DOI 10.1145/3274417. URL http://doi.acm.org/10.1145/3274417
151. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(5), 582–638 (2014)
152. Ruggieri, S., Hajian, S., Kamiran, F., Zhang, X.: Anti-discrimination analysis using privacy attack strategies. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 694–710. Springer (2014)
153. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD) **4**(2), 9 (2010)
154. Ruggieri, S., Pedreschi, D., Turini, F.: DCUBE: discrimination discovery in databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010, pp. 1127–1130 (2010). DOI 10.1145/1807167.1807298. URL https://doi.org/10.1145/1807167.1807298
155. Ruggieri, S., Turini, F.: A kdd process for discrimination discovery. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 249–253. Springer (2016)
156. Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T., Ghani, R.: Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018)
157. Salimi, B., Cole, C., Li, P., Gehrke, J., Suciu, D.: Hypdb: a demonstration of detecting, explaining and resolving bias in olap queries. Proceedings of the VLDB Endowment **11**(12), 2062–2065 (2018)
158. Salimi, B., Gehrke, J., Suciu, D.: Bias in olap queries: Detection, explanation, and removal. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pp. 1021–1035. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3196914. URL http://doi.acm.org/10.1145/3183713.3196914
159. Salimi, B., Howe, B., Suciu, D.: Data management for causal algorithmic fairness. arXiv preprint arXiv:1908.07924 (2019)
160. Salimi, B., Howe, B., Suciu, D.: Database repair meets algorithmic fairness. SIGMOD Rec. **49**(1), 34–41 (2020). DOI 10.1145/3422648.3422657. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3422648.3422657
161. Salimi, B., Parikh, H., Kayali, M., Getoor, L., Roy, S., Suciu, D.: Causal relational learning. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, SIGMOD '20, p. 241–256. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3318464.3389759. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3318464.3389759
162. Salimi, B., Rodriguez, L., Howe, B., Suciu, D.: Interventional fairness: Causal database repair for algorithmic fairness. In: Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, pp. 793–810. ACM, New York, NY, USA (2019). DOI 10.1145/3299869.3319901. URL http://doi.acm.org/10.1145/3299869.3319901
163. Salman, I.: Cognitive biases in software quality and testing. In: Proceedings of the 38th International Conference on Software Engineering Companion, ICSE '16, pp. 823–826. ACM, New York, NY, USA (2016). DOI 10.1145/2889160.2889265. URL http://doi.acm.org/10.1145/2889160.2889265
164. Salminen, J., Jung, S.G., Jansen, B.J.: Detecting demographic bias in automatically generated personas. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, pp. LBW0122:1–LBW0122:6. ACM, New York, NY, USA (2019). DOI 10.1145/3290607.3313034. URL http://doi.acm.org/10.1145/3290607.3313034
165. Schelter, S., He, Y., Khilnani, J., Stoyanovich, J.: Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. arXiv preprint arXiv:1911.12587 (2019)
166. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 59–68. ACM, New York, NY, USA (2019). DOI 10.1145/3287560.3287598. URL http://doi.acm.org/10.1145/3287560.3287598
167. Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D.: No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536 (2017)
168. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pp. 2219–2228 (2018). DOI 10.1145/3219819.3220088. URL https://doi.org/10.1145/3219819.3220088
169. Sinha, S., Agarwal, M., Vatsa, M., Singh, R., Anand, S.: Exploring bias in primate face detection and recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0 (2018)
170. Solomon, J.: Customization bias in decision support systems. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14, pp. 3065–3074. ACM, New York, NY, USA (2014). DOI 10.1145/2556288.2557211. URL http://doi.acm.org/10.1145/2556288.2557211
171. Sonboli, N., Burke, R.: Localized fairness in recommender systems. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP'19 Adjunct, pp. 295–300. ACM, New York, NY, USA (2019). DOI 10.1145/3314183.3323845. URL http://doi.acm.org/10.1145/3314183.3323845
172. Spillane, B., Lawless, S., Wade, V.: Measuring bias in news websites, towards a model for personalization. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP '17, pp. 387–388. ACM, New York, NY, USA (2017). DOI 10.1145/3079628.3079647. URL http://doi.acm.org/10.1145/3079628.3079647

173. Springer, A., Garcia-Gathright, J., Cramer, H.: Assessing and addressing algorithmic bias-but before we get there... In: 2018 AAAI Spring Symposium Series (2018)
174. Srivastava, M., Heidari, H., Krause, A.: Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pp. 2459–2468 (2019). DOI 10.1145/3292500.3330664. URL https://doi.org/10.1145/3292500.3330664
175. Stock, P., Cisse, M.: Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 498–512 (2018)
176. Stoyanovich, J., Howe, B.: Nutritional labels for data and models. Data Engineering p. 13 (2019)
177. Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., Weikum, G.: Fides: Towards a platform for responsible data science. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM '17, pp. 26:1–26:6. ACM, New York, NY, USA (2017). DOI 10.1145/3085504.3085530. URL http://doi.acm.org/10.1145/3085504.3085530
178. Stoyanovich, J., Howe, B., Jagadish, H.: Responsible data management. Proceedings of the VLDB Endowment **13**(12), 3474–3488 (2020)
179. Stoyanovich, J., Yang, K., Jagadish, H.V.: Online set selection with fairness and diversity constraints. In: Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018., pp. 241–252 (2018). DOI 10.5441/002/edbt.2018.22. URL https://doi.org/10.5441/002/edbt.2018.22
180. Stratigi, M., Kondylakis, H., Stefanidis, K.: Fairness in group recommendations in the health domain. In: 33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017, pp. 1481–1488 (2017). DOI 10.1109/ICDE.2017.217. URL https://doi.org/10.1109/ICDE.2017.217
181. Sühr, T., Biega, A.J., Zehlike, M., Gummadi, K.P., Chakraborty, A.: Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pp. 3082–3092 (2019). DOI 10.1145/3292500.3330793. URL https://doi.org/10.1145/3292500.3330793
182. Sun, C., Asudeh, A., Jagadish, H., Howe, B., Stoyanovich, J.: Mithralabel: Flexible dataset nutritional labels for responsible data science. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2893–2896 (2019)
183. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E.M., Chang, K., Wang, W.Y.: Mitigating gender bias in natural language processing: Literature review. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 1630–1640 (2019). URL https://www.aclweb.org/anthology/P19-1159/
184. Tae, K.H., Roh, Y., Oh, Y.H., Kim, H., Whang, S.E.: Data cleaning for accurate, fair, and robust models: A big data - ai integration approach. In: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, DEEM'19, pp. 5:1–5:4. ACM, New York, NY, USA (2019). DOI 10.1145/3329486.3329493. URL http://doi.acm.org/10.1145/3329486.3329493
185. Thebault-Spieker, J., Venkatagiri, S., Mitchell, D., Hurt, C., Luther, K.: Pairwise: Mitigating political bias in crowdsourced content moderation. In: Proceedings of HCOMP 2019 (2019)
186. Torralba, A., Efros, A.A., et al.: Unbiased look at dataset bias. In: CVPR, vol. 1, p. 7. Citeseer (2011)
187. Udeshi, S., Arora, P., Chattopadhyay, S.: Automated directed fairness testing. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, pp. 98–108. ACM, New York, NY, USA (2018). DOI 10.1145/3238147.3238165. URL http://doi.acm.org/10.1145/3238147.3238165
188. Vandenhof, C.: A hybrid approach to identifying unknown unknowns of predictive models. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, pp. 180–187 (2019)
189. Vasconcelos, M., Cardonha, C., Gonçalves, B.: Modeling epistemological principles for bias mitigation in ai systems: An illustration in hiring decisions. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, pp. 323–329. ACM, New York, NY, USA (2018). DOI 10.1145/3278721.3278751. URL http://doi.acm.org/10.1145/3278721.3278751
190. Veale, M., Van Kleek, M., Binns, R.: Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp. 440:1–440:14. ACM, New York, NY, USA (2018). DOI 10.1145/3173574.3174014. URL http://doi.acm.org/10.1145/3173574.3174014
191. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, FairWare '18, pp. 1–7. ACM, New York, NY, USA (2018). DOI 10.1145/3194770.3194776. URL http://doi.acm.org/10.1145/3194770.3194776
192. Vorvoreanu, M., Zhang, L., Huang, Y.H., Hilderbrand, C., Steine-Hanson, Z., Burnett, M.: From gender biases to gender-inclusive design: An empirical investigation. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, pp. 53:1–53:14. ACM, New York, NY, USA (2019). DOI 10.1145/3290605.3300283. URL http://doi.acm.org/10.1145/3290605.3300283
193. Wang, Y., Redmiles, D.F.: Implicit gender biases in professional software development: an empirical study. In: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019, pp. 1–10 (2019). DOI 10.1109/ICSE-SEIS.2019.00009. URL https://doi.org/10.1109/ICSE-SEIS.2019.00009
194. Weng, J., Shen, Z., Miao, C., Goh, A., Leung, C.: Credibility: How agents can handle unfair third-party testimonies in computational trust models. IEEE Transactions on Knowledge and Data Engineering **22**(9), 1286–1298 (2009)
195. Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warshaw, J.: A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp. 656:1–656:14. ACM, New York, NY, USA

(2018). DOI 10.1145/3173574.3174230. URL http://doi.acm.org/10.1145/3173574.3174230

196. Yahav, I., Shehory, O., Schwartz, D.: Comments mining with tf-idf: The inherent bias and its removal. IEEE Transactions on Knowledge and Data Engineering **31**(3), 437–450 (2018)
197. Yamada, M., Sigal, L., Raptis, M.: No bias left behind: Covariate shift adaptation for discriminative 3d pose estimation. In: European Conference on Computer Vision, pp. 674–687. Springer (2012)
198. Yan, A., Howe, B.: Fairness in practice: A survey on equity in urban mobility. Data Engineering p. 49 (2019)
199. Yang, K., Gkatzelis, V., Stoyanovich, J.: Balanced ranking with diversity constraints. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI (2019)
200. Yang, K., Huang, B., Stoyanovich, J., Schelter, S.: Fairness-aware instrumentation of preprocessing pipelines for machine learning. In: HILDA workshop at SIGMOD (2020)
201. Yang, K., Loftus, J.R., Stoyanovich, J.: Causal intersectionality for fair ranking (2020)
202. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, p. 547–558. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3351095.3375709. URL https://doi.org/10.1145/3351095.3375709
203. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, SSDBM '17, pp. 22:1–22:6. ACM, New York, NY, USA (2017). DOI 10.1145/3085504.3085526. URL http://doi.acm.org/10.1145/3085504.3085526
204. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., Miklau, G.: A nutritional label for rankings. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD '18, pp. 1773–1776. ACM, New York, NY, USA (2018). DOI 10.1145/3183713.3193568. URL http://doi.acm.org/10.1145/3183713.3193568
205. Yuan, S., Wu, X., Xiang, Y.: A two phase deep learning model for identifying discrimination from tweets. In: Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016., pp. 696–697 (2016). DOI 10.5441/002/edbt.2016.92. URL https://doi.org/10.5441/002/edbt.2016.92
206. Zhang, J., Wu, X., Sheng, V.S.: Imbalanced multiple noisy labeling. IEEE Transactions on Knowledge and Data Engineering **27**(2), 489–503 (2014)
207. Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, pp. 1335–1344 (2017). DOI 10.1145/3097983.3098167. URL https://doi.org/10.1145/3097983.3098167
208. Zhang, L., Wu, Y., Wu, X.: Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. IEEE Transactions on Knowledge and Data Engineering (2018)
209. Zhang, Y., Bellamy, R.K., Kellogg, W.A.: Designing information for remediating cognitive biases in decision-making. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 2211–2220. ACM, New York, NY, USA (2015). DOI 10.1145/2702123.2702239. URL http://doi.acm.org/10.1145/2702123.2702239
210. Zheng, Y., Dave, T., Mishra, N., Kumar, H.: Fairness in reciprocal recommendations: A speed-dating study. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP '18, pp. 29–34. ACM, New York, NY, USA (2018). DOI 10.1145/3213586.3226207. URL http://doi.acm.org/10.1145/3213586.3226207
211. Zliobaite, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Discov. **31**(4), 1060–1089 (2017). DOI 10.1007/s10618-017-0506-1. URL https://doi.org/10.1007/s10618-017-0506-1