Mitigating Bias in Algorithmic Systems - A Fish-Eye View

KALIA ORPHANOU, Open University of Cyprus, CYPRUS
JAHNA OTTERBACHER and STYLIANI KLEANTHOUS, Open University of Cyprus & CYENS Centre of Excellence, CYPRUS
KHUYAGBAATAR BATSUREN*, National University of Mongolia, MONGOLIA
FAUSTO GIUNCHIGLIA, The University of Trento, ITALY
VERONIKA BOGINA, AVITAL SHULNER TAL, ALAN HARTMAN, and TSVI KUFLIK, The University of HaifaP.O. Box 1212 , ISRAEL

Mitigating bias in algorithmic systems is a critical issue drawing attention across communities within the information and computer sciences. Given the complexity of the problem and the involvement of multiple stakeholders – including developers, end users and third-parties – there is a need to understand the landscape of the sources of bias, and the solutions being proposed to address them, from a broad, cross-domain perspective. This survey provides a "fish-eye view," examining approaches across four areas of research. The literature describes three steps toward a comprehensive treatment – bias detection, fairness management and explainability management – and underscores the need to work from within the system as well as from the perspective of stakeholders in the broader context.

CCS Concepts: • **Information systems** → **Information systems applications**; **Decision support systems**; • **Human-centered computing**; • **Social and professional topics**;

Additional Key Words and Phrases: Algorithmic bias, explainability, fairness, social bias, transparency

ACM Reference Format:
Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. 2022. Mitigating Bias in Algorithmic Systems - A Fish-Eye View. 1, 1 (February 2022), 47 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Long before the widespread use of algorithmic systems driven by big data, Friedman and Nissenbaum [69], writing in the ACM TOIS in 1996, argued that "freedom from bias" should be considered equally alongside

*Work conducted while at The University of Trento.

Authors' addresses: Kalia Orphanou, kalia.orphanou@ouc.ac.cy, Open University of Cyprus, P.O. Box 1212, CYPRUS; Jahna Otterbacher, jahna.otterbacher@ouc.ac.cy; Styliani Kleanthous, styliani.kleanthous@ouc.ac.cy, Open University of Cyprus & CYENS Centre of Excellence, P.O. Box 1212, CYPRUS; Khuyagbaatar Batsuren, National University of Mongolia, Ulan Baatar, MONGOLIA, k.batsuren@unitn.it; Fausto Giunchiglia, The University of Trento, 1 Thørväld Circle, Trento, ITALY; Veronika Bogina; Avital Shulner Tal; Alan Hartman; Tsvi Kuflik, The University of Haifa, \unskipP.O. Box 1212, ISRAEL.

Manuscript submitted to ACM

the criteria of reliability, accuracy and efficiency, when judging the quality of a computer system. Defining biased systems as those that "systematically and unfairly discriminate" against individuals or certain social groups, they emphasized that if a biased system becomes widely adopted in society, that the social biases it perpetuates will have serious consequences. More than 20 years later, the ACM U.S. Public Policy Council (USACM) and the ACM Europe Policy Committee (EUACM) published a joint Statement on Algorithmic Transparency and Accountability,[1] underscoring widespread concerns surrounding computer bias, but this time, focusing on the social consequences of *data-driven algorithmic processes and systems*. The statement puts forward seven principles to be considered in the context of system development and deployment, in working toward mitigating the threat of harm to people posed by biases. Despite that the principles are articulated in a single page, it is clear that the issue of algorithmic bias is extremely complex. Multiple sources of bias (e.g., data, modelling processes) are mentioned, as well as alternative solutions – from simply raising users' awareness of the issue, to enabling the auditing of models by third parties. Furthermore, the principles mention a range of stakeholders (the algorithm's owners, designers, builders, and end users), alluding to their roles in ensuring the ethical development and appropriate use of algorithmic processes.

Despite the recent surge in attention to the topic, addressing algorithmic bias is not a new concern for researchers. For instance, in the 1990s, machine learning researchers were considering problems of *explainability*, or how to interpret models and facilitate their use (e.g., [44], [45], [54]). In the early 2000s, researchers in the data mining community were developing processes for *discrimination discovery* from historical datasets (e.g., [156]). Similarly, around the same time, information retrieval researchers were considering the issue of bias in training datasets (e.g., [23]) and the resulting impact of this bias on ranking algorithms [39]. Thus, while several research communities were tackling various issues related to algorithmic biases earlier on, they were largely disjoint from one another. Furthermore, they addressed the problems from "inside," working exclusively from the perspective of the developer. More recently, multiple perspectives on algorithmic bias have come to light, with the increasing influence of algorithmic systems in society. Arguably, a 2016 article entitled *Machine Bias* [4] played a key role in stimulating widespread discussion, opening up the conversation to other stakeholders beyond those who develop algorithmic processes and systems.

Recently, a number of comprehensive surveys has emerged on algorithmic bias, shedding light on the source(s) of bias and highlighting potential solutions. However, such surveys tend to focus on one source of algorithmic bias and/or one class of solutions. For instance, Olteanu and colleagues [146] reviewed the literature surrounding data biases; in particular, they address social data sources, given their frequent use in the creation of training data sets. Coming from a fair machine learning perspective, Mehrabi and colleagues [137] provided a survey of common problems and solutions, including those focused on data and processes. Addressing explainability, Guidotti et al. contributed a comprehensive survey and a taxonomy of the various methods used to interpret the behaviors of black box models [81]. In addition, there are survey papers providing deep dives into the technical solutions proposed in very well-defined areas. For instance, in [191], the authors focus specifically on gender bias in the natural language processing domain, in [30, 41], the authors consider the technical approaches of mitigating bias in ML while in [8], the authors focus on data bias and data management approaches for mitigating bias.

[1] https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

In this survey, our aim is to help the reader achieve a high-level understanding of the current state of this complex topic, across domains. With a view toward promoting more comprehensive solutions, we present a *fish-eye view* of the literature surrounding algorithmic bias, and provide a methodology that is based on three key aspects, namely problems, domains and stakeholders. By examining the literature along these lines, we can better understand how solutions can and should be used to address algorithmic system bias.

In information visualization, fish-eye views, which balance focus and context (i.e., depth and breadth), are useful for facilitating understanding in information spaces that are very large and diverse [71]. The user maintains perspective of the "big picture," but can still choose when to drill down into further details. Given the diversity of perspectives on algorithmic bias, we argue that a high-level view is much needed, particularly for researchers and practitioners new to the area.

The main contributions of this survey paper are to:

- Provide a methodology for analyzing the work on algorithmic bias, and a "live" repository of articles.
- Document the problems and solutions studied across research communities.
- Map the problems to the solutions across diverse domains, as well as the involved stakeholders.
- Describe opportunities for cross-fertilization between communities, solutions and stakeholders.

The article is organized as follows. Section 2 describes the methodology used for the literature review and the selection of domains and publication venues. Section 3 presents an overview of the problem and solution spaces discovered while analyzing the papers. Following that, we present the detailed analysis of the three categories of solutions described in the literature: Section 4 focuses on Bias Detection, Section 5 details the methods used for Fairness Management, and Section 6 presents a summary of the work within Explainability Management. In each section, we provide specific examples of the respective solution, described in the literature. Each section ends with a table providing a comparison of the specific approaches taken across the four domains studied. Finally, in Sections 7 and 8, we summarize the state-of-the-field, discussing the cross-fertilization among the four communities, the stakeholders and the solutions. These sections also present some open issues for further consideration.

2 METHODOLOGY

We follow a methodology involving both bottom-up and top-down processes for collecting articles relevant to *bias in algorithmic systems*. The methodology can be characterized as an adaptation of the standard facet-based methodology used in information science to carry out book and even product classification [88]. In the first phase, a bottom-up, open search process took place, in which each co-author collected relevant literature, adding it to a shared repository. This initial body of material was then used to guide the choice of research domains and publication venues upon which to focus, as well as to identify a set of properties by which to characterize the problems and solutions described.

2.1 Selection of Domains

An inventory of the initial article repository was taken, to understand which domains (i.e., research communities) had produced a critical mass of publications related to the mitigation of algorithmic biases. We focused on well-established domains within the information and computer sciences, which are investigating

data and knowledge transformation and communication to the user. Based on the initial inventory, four domains emerged – machine learning (ML), human-computer interaction (HCI), recommender systems (RecSys), and information retrieval (IR) – to characterize the problems of algorithmic biases that are being addressed, as well as the solutions being proposed, across domains. Next, we provide additional justification of these four domains.

The widespread application of ML techniques, which in many cases are opaque, led to the issue of potential bias and discrimination of algorithmic systems and processes. Hence, the ML community and ML-related publications naturally emerged as an established area we needed to review. RecSys represents a specific application area and a domain that attracts significant research attention on algorithmic bias. Within RecSys, ML techniques are applied for reasoning on and exploiting user characteristics; thus, within this domain, many challenges have arisen surrounding potential bias and fairness. IR focuses on information delivery to users, often with the use of search and ranking algorithms that are opaque; thus, bias and fairness have long been researched. The above domains cover a substantial amount of applications where the risk of bias and discrimination in the reasoning process exists. Finally, HCI directly considers the end users and their perceptions when interacting directly or indirectly with different applications. In particular, understanding the potential bias, discrimination or fairness issues that might emerge when a user is interacting with information presented through an interface is considered of high importance. It should be noted that an "Other" domain emerged, through the initial repository, where we collected a number of articles published in emerging, cross-disciplinary communities or domains that are not represented in the above main categories. Through "Other" we were able to capture research published in other domains where a mass of publications related to bias, fairness and explainability did not (yet) exist, but important work was published, hence, making this a comprehensive review with applicability in areas other than the main domains that emerged.

2.2 Selection of Publication Venues

Through the exercise of selecting the research domains, a list of high-impact publication venues, including both conferences and journals, was created for each domain, as presented in Table 1. Also, note that some venues publish articles across domains. For instance, while ACM CSCW is generally aligned with the HCI community, some articles describing studies of recommender systems can be found there. Such cases are indicated with parentheses in Table 1.

Domain	Publication Venues Reviewed	# Papers
Machine Learning/AI	AAAI, IJCAI, KDD, SIGKDD, CIDM, ICML, AIES, NIPS, MLSP, ACM Data Mining and Knowledge Discovery Journal	106
Information Retrieval	ACM SIGIR, ACM CIKM, ACM WWW, TOIS, JASIS, IR Journal, (AAAI ICWSM)	68
Recommender Systems	ACM RecSys, AAAI ICWSM, UMUAI, ArXive (ACM CSCW, ACM CIKM, ACM FAccT (formerly FAT*))	46
Human Computer Interaction	ACM CHI, ACM CSCW, ACM CHI Journal, CSCW Journal Journal of Behaviour and Information Technology	34
Other	AAAI HCOMP, ACM FAccT, ICDM,VLDB	57

Table 1. Key publication venues reviewed per domain.

The next step was to review each publication venue's proceedings / published volumes during the twelve-year period 2008 - 2021, resulting in a high-recall search for relevant published articles. The key words used were: "accountability," "bias," "discrimination," "fair(ness)," "explain(able)," and "transparen(cy)." In addition, the articles collected address a particular algorithmic process or system, or a class of system. In other words, articles of a more abstract or philosophical nature were excluded from the survey. Likewise, in the ML category, articles from AI venues (e.g., AAAI, IJCAI) that were not published in the respective ML track, have been excluded, as to focus on algorithmic, data-driven systems.

3 ANALYSIS OF PAPERS

This survey is based on our current repository of over 245 articles.[2] The list of publication venues reviewed is not exhaustive; further venues may be added to our repository in the future. However, the problem and solution spaces discovered and detailed below, have proven to be robust across the articles reviewed to date. In our repository, each article is labeled with its respective domain (ML, HCI, RecSys, IR, Other). After reviewing the article, three additional properties, which shall be explained below, were also recorded:

- The problem(s) identified within the system.
- The attribute(s) affected by the problem(s).
- The solution(s) proposed to address the problem(s).

These four attributes – domain, problematic system component(s), attribute(s) affected by the problem, and proposed solution(s) – are provided as tags in our Zotero repository. Thus, other researchers may use this resource in various ways, e.g., to focus on a specific problem or type of solution. Table 2 provides examples of the manner in which articles in our repository were analyzed; further details are provided in the following subsections.

Domain	Example	Problem(s)	Attribute(s)	Solution(s)
ML	Word embeddings trained on Google News articles were found to perpetuate prevalent gender biases. [18]	Data	gender	Discrimination discovery - indirect
IR	Users of Mendeley search were shown to disproportionately favor articles written by authors sharing their national origin. [196]	User	national origin	Discrimination discovery - direct
RecSys	Profiles of women and people of color in online freelance marketplaces were found to be systematically lower-ranked than others; reasons included bias in training data and lower evaluations by other users. [82]	User, Data Third parties	gender, race	Auditing
HCI	Authors provided various explanations to users about their Facebook feeds. Explanations were found to shape beliefs on how the system works, but not in understanding its specific outputs. [162]	User Output	information	Explainability - model, output
Other	Authors addressed the issue of human bias in computer vision training data, through an algorithm that filters human reporting bias from labels that are visually grounded. [140]	Data	information	Fairness pre-processing

Table 2. Example analyses of articles in the repository.

[2]Available at Zotero - https://www.zotero.org/groups/2450986/cycat_survey_collection_public.

3.1 Problem space

To explore the problem space within the literature addressing algorithmic bias, we characterized, for each article, the system component(s) deemed to be problematic by the authors, as well as the attribute(s) whose values are affected by the bias.

3.1.1 Problematic system component(s). We recorded the macro component(s) of the algorithmic system or process,[3] considered by the author(s) as being the source of the problem. Figure 1 provides a general characterization of an algorithmic system, with its macro components, which we have used to examine the problem space of algorithmic bias. Note that some components are optionally present. This includes a User (U), who interacts directly with the system's inputs and/or outputs. Alternatively, an API may be in place, to allow the system to interact with other systems and applications.

In this generic architecture, the system receives input (I) for an instance of its operation. This is provided by a user (U), or another source (e.g., the result of an automated process). The algorithmic model (M) makes some computation(s) based on the inputs and produces an output (O). The model learns from a set of observations of data (D) from the problem domain. It may also receive constraints from third-party actors (T) and/or internal fairness criteria (F) that modify the operation of the algorithmic model (M). Finally, some systems have direct interaction with a user (U) who, as previously discussed, will bring her own knowledge, background and attitude when interpreting the system's output.

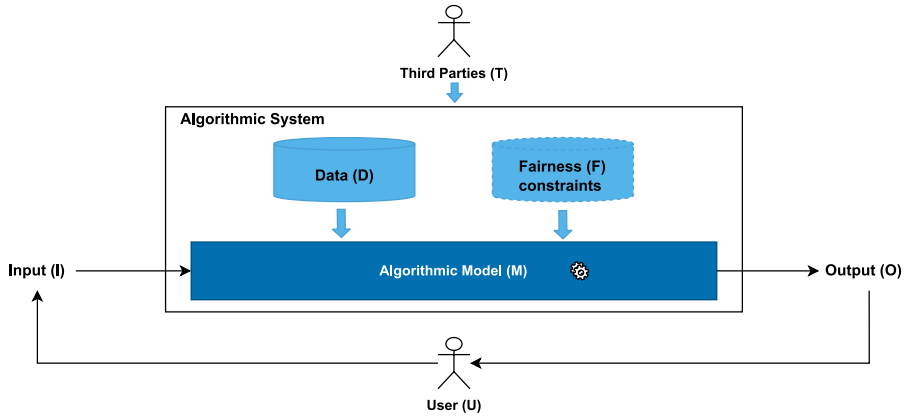

Fig. 1. Generic architecture of an algorithmic system.

Thus, as depicted in Figure 1, bias may manifest and/or be detected in one or more of these components:

- Input (I) - Bias may be introduced in the input data, e.g., as incorrect or incomplete information input by the user.
- Output (O) - Bias may be detected at the outcome (value(s)/label(s)) produced in response to the input.
- Algorithm (M) - Bias can manifest during the model's processing and learning.

[3]Henceforth, we refer to a "system", although as mentioned, we consider articles that describe particular algorithmic processes as well as those describing deployed systems.

- Training Data (D) - Training data may be unbalanced and/or discriminatory toward groups of people. The data may also be based on an unrepresentative set of instances, and may also suffer from inaccuracies in the ground truth.
- Third Party Constraints (T) - Implicit and explicit constraints, given by third parties, may impact the design and performance of the algorithm such as to be discriminatory towards groups of people. These include operators of the system, regulators, and other bodies that influence the use and outcomes of the system.[4]
- User (U) - When users interact directly with a system, they may contribute to bias in a number of ways, such as through the inappropriate use of the system or misinterpretation of system output.

3.1.2 Affected attribute(s). We have also characterized, for a given article, the attribute(s) whose values are influenced by the problematic system behavior. While early technical papers used the generic terms, *"sensitive" or "protected" attributes*, to characterize the features on which a group is unfairly treated by the algorithmic decision [157], recent work has considered a broader set of attributes, including the social, cultural, and political attributes of the content or person being processed, e.g., gender, race, age, income, etc.

The system under study in each article can exhibit different behaviors with respect to the affected attribute, which may or may not be problematic for a given user or observer. It can be noted that while many of the affected attributes concern social and cultural characteristics (i.e., characteristics describing the social world), we also observe dimensions such as the quality / accuracy / credibility of the information provided to the user (i.e., information attribute). By information, we mean the quality (or bias) of the information that is conveyed by or to the user. In other words, the concern here is the extent to which the data used by the algorithm constitute a truthful representation of the world. Note that information bias may also be introduced by the algorithm because of its low classification/predictive performance, i.e., low accuracy. Even though such instances may not represent cases where an algorithmic system's behavior can directly result in discrimination or harm, in many contexts, these issues can indirectly lead to serious consequences for system users (e.g., limited exposure to high-quality sources of information on a given topic because of biased search engine results).

Information is the most studied attribute in our corpus, and is the primary dimension addressed in the ML literature. For instance, in the explainability literature, a primary concern is the extent to which information is effectively conveyed to the user. Likewise, IR articles often consider information as the affected dimension under study; here, the classic example is the large body of work on search engine biases. In contrast, the literature in HCI and RecSys does not often address information as an affected dimension. In these fields, articles on mitigating algorithmic biases more often consider social and cultural dimensions, such as demographics as a general term, or more specific attributes such as gender and race, with a few studies emerging on characteristics such as age, language and physical attractiveness.

[4] An example was described in Table 2 of a RecSys in which other users' ratings of workers affected system performance during a given user's instance. Another example might be a search engine suppressing some ranked results to comply with laws in the user's geographical region.

3.2 Solution Space

The solution space discussed in the literature we surveyed consists of three main steps in mitigating algorithmic bias. Each of these steps involves different stakeholders within each community. Next, we describe the multiple stakeholders who are involved in the solution processes. Afterwards, we give a detailed overview of each of the three steps in bias mitigation.

3.2.1 Stakeholders. The selection of the four domains allows us to capture perspectives and processes involving multiple stakeholders, as also depicted in Figure 2. For instance, while the ML literature focuses primarily on the developer perspective (and thus, *formal* processes), HCI researchers consider the user's interaction with the system, or how the interface might influence the user's perception of fairness (and thus, more *informal* processes). IR and RecSys represent communities focused on end user applications; thus, we can learn the extent to which algorithmic biases have presented challenges to these applications and the nature of the solutions proposed.

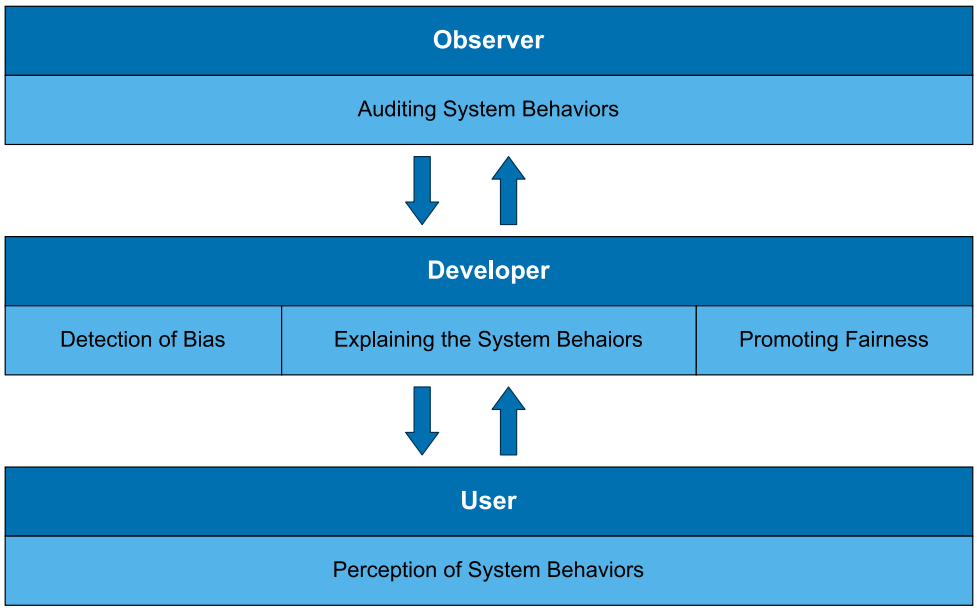

Fig. 2. Processes and stakeholders involved in mitigating algorithmic bias.

Developer(s) can internally detect bias in data and processes, evaluate formal notions of fair treatment of the individuals, groups and/or content affected by algorithmic judgements, as well as implement methods used by the system for explaining its decisions to users and/or third parties. *System Observer(s)*, who may be regulators, researchers or even data journalists, can conduct their own audits of the system behaviors. However, *User(s)* of the system have their own perceptions of the system's behavior, which depend not only on the system itself, but also their own knowledge, experience and attitudes. *Indirect User(s)* are the people who are affected by the algorithmic decision. These are, for instance, defendants evaluated by an

algorithmic risk-assessment system or candidates whose resumes are screened with an algorithmic resume screening system. Indirect users also have their own perceptions of the system's behavior.

3.2.2 Classification of Solutions. The literature suggests that a comprehensive solution for mitigating algorithmic system bias consists of three main steps: *i*) Bias Detection, *ii*) Fairness Management, and *iii*) Explainability Management.

Bias Detection includes techniques that scrutinize the system to detect any type of systematic bias. It can be achieved by *Auditing* and/or *Discrimination Discovery* methods. Auditing involves making cross-system or within-system comparisons, and is typically done by an analyst / observer or a regulator who does not have access to the inner-workings of the system [176]. There are variations in the extent to which the auditing approaches are formalized. In some cases, auditing uses the tools of discrimination discovery (e.g., discrimination/fairness metrics). In this sense, auditing as a term refers to who is doing the discrimination discovery and why, but not to a different set of tools and techniques. In other cases, auditing is used in a more formal way to detect any fairness issues in the system. Discrimination discovery (direct or indirect) approaches include tools and practices for detecting unfair treatment by data / algorithms / systems using statistical metrics, i.e., measuring specific fairness notions.

Fairness Management includes techniques that developers use to mitigate the detected bias and certify that the system is fairness-aware. Fairness management approaches are used by developers to tackle bias in different parts of the system. They are divided into: *Fairness pre-processing*, *Fairness in-processing*, *Fairness post-processing*, *Fairness certification* and *Fairness perception*.

- *Fairness Pre-processing* includes approaches that process the training data in a manner that promotes fairness. Examples are: re-sampling, re-weighting and feature transformation approaches.
- *Fairness In-processing* includes approaches that address discrimination during the training procedure. Examples are: regularization, optimization and learn-to-rank approaches.
- *Fairness Post-processing* includes approaches that ensure that a system is "fair" by changing the output of the learned classifier i.e., changing the label weights and re-ranking approaches.
- *Fairness Certification* includes approaches provided by the developer/observer in the case where no unintended bias has been detected in the system. The developer/observer verifies whether the output satisfies specific fairness constraints, and if so, he or she can certify the algorithmic system as "fair." In general, fairness certification aims to test algorithmic models for possible disparate impact, according to the fairness internal certification and bias detection results, "certifying" those that do not exhibit evidence of unfairness.
- *Fairness Perception* includes approaches that examine the perception of different stakeholders with the decision outcome of the algorithm. Examples are the use of questionnaires and statistical tests.

Explainability Management includes techniques that facilitate transparency and build trust between the end user and the system. Explainability and interpretability contribute to the sense of transparency as well as the perception of fairness [83]. Explainability approaches are used to provide transparency of the system and in that way, enable the detection of any bias or fairness issues in the data and model. In general, explainability-aware techniques can be divided into two main categories: *Model Explainability*, which provides explanations for the training process of the models and *Outcome Explainability*, which provides explanations of the algorithm's decision outcome in an understandable way to the user. Outcome Explainability methods

explain only the output, and they do not provide explanations for the process of the algorithm. This form of explanation is usually helpful when the user of the system is not an expert, such as in the case of RecSys.

Figure 3 aligns the three steps involved in mitigating biases, with the taxonomy of solutions found in the literature surveyed.

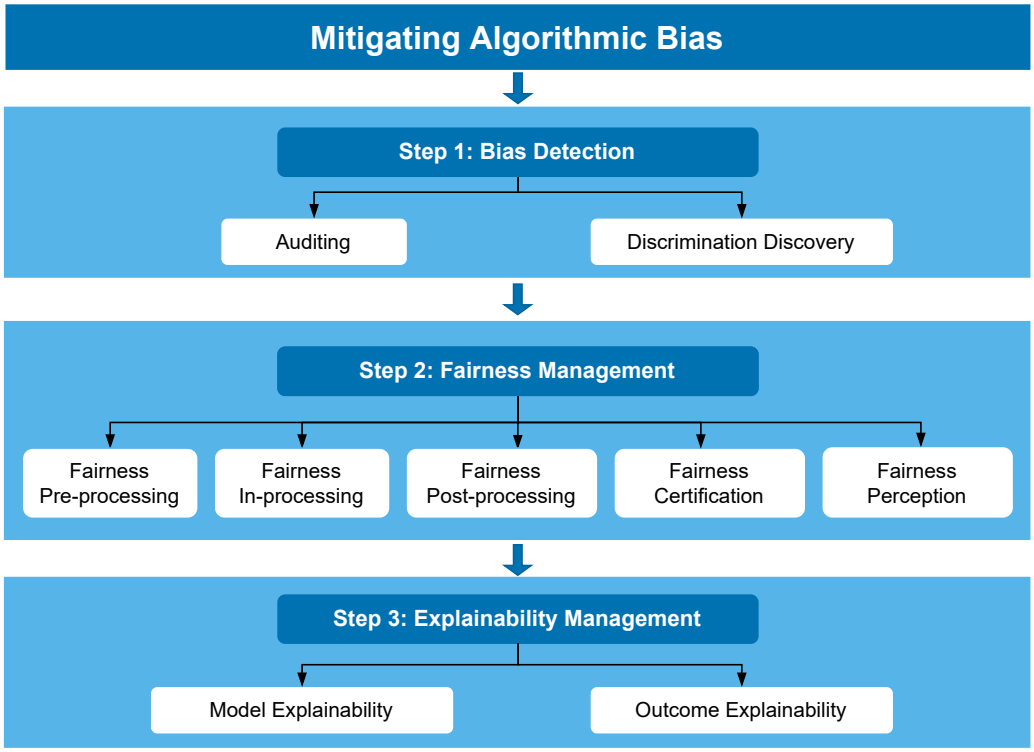

Fig. 3. The solution space - tools for mitigating bias in algorithmic systems.

3.3 Summary

Before describing each set of techniques in detail, we provide a summary overview of the problems and solutions documented within each of the four domains surveyed. These are presented for each of the three steps (i.e., Table 3 presents Bias Detection solutions, while Tables 4 and 5 present solutions for Fairness and Explainability Management, respectively). The distribution of problems addressed across the four domains illustrates the insights gained from our "fish-eye view." As expected, the ML literature addresses problems concerning the training data, the algorithmic model and the system output. The RecSys and IR literature, as user-focused application areas, consider problems both inside and outside the system, while HCI naturally addresses the interactions between the user and the algorithmic system.

Similarly, we find that across domains, researchers are engaged in all three steps in bias mitigation – detection, fairness and explainability management. In the following sections, we shall provide a detailed

overview of specific examples of the approaches of each of the three steps, from across domains, and shall compare the techniques used.

Domain	Problem	Solution	Reference(s)
Bias Detection			
ML	Data/Model	Auditing	[6, 130, 174, 227]
	Data/Model/Output	Discrimination Discovery	[43, 49, 98, 123, 157, 222, 227, 233]
IR	User/Data	Auditing	[92, 105, 118, 122, 132, 147, 204]
	User/Data	Discrimination Discovery	[13, 35, 59, 112, 127, 150, 208, 211–213, 221]
HCI	User/Data	Auditing	[101, 135, 139]
	Third Party/Model/Output	Discrimination Discovery	[11, 48, 78, 161]
RecSys	Data/Output	Auditing	[57, 62, 94]
	Data/Output	Discrimination Discovery	[2, 15, 60, 188, 192]

Table 3. Summary of the problem and bias detection solution space per domain.

Domain	Problem	Solution	Reference(s)
Fairness Management			
ML	Data	Fairness Pre-processing	[26, 100, 102, 121, 227]
	Model	Fairness In-processing	[31, 52, 79, 103, 108–110, 120, 165, 216, 219, 225]
	Model/Output	Fairness Post-processing	[84, 102, 157]
	User/Output	Fairness Perception	[189]
	Data/Model/Output	Fairness Certification	[31, 46, 52, 64, 79, 108–110, 182, 216, 225]
IR	Data	Fairness Pre-processing	[51, 53, 76, 184]
	User/Model	Fairness In-processing	[47, 149, 220]
	User/Output	Fairness Post-processing	[104, 117, 119]
	User	Fairness Certification	[61, 90, 141]
HCI	Data	Fairness Pre-processing methods	[101]
	User/Model	Fairness Perception	[22]
	User/Output	Fairness Certification	[124, 214]
RecSys	Data	Fairness Pre-processing methods	[25, 104, 130, 138, 215, 217, 220]
	User/Model	Fairness In-processing methods	[220]
	User/Output	Fairness Post-processing	[104, 186, 223]

Table 4. Summary of the problem and fairness management solution space per domain.

Domain	Problem	Solution	Reference(s)
Explainability Management			
ML	Model	Model Explainability	[21, 38, 54, 75, 114, 177, 193, 230] [28, 45, 99, 128, 194, 231]
	Output	Outcome Explainability	[49, 160, 167, 167, 168, 190, 200] [17, 68, 87, 180, 185, 203, 218, 230, 232]
HCI	User	Model Explainability	[91]
	User/Output	Outcome Explainability	[16, 58, 70, 162]
RecSys	User/Output	Outcome Explainability	[20, 37, 113, 145, 197, 201, 205]

Table 5. Summary of the problem and explainability management solution space per domain.

4 DETECTION OF BIAS IN ALGORITHMIC SYSTEMS

There are multiple notions of fairness that are important in the context of an algorithmic system as given in [202]. The fairness of an algorithmic model (or classifier), depends on the notion of fairness one wants to adopt. Based on [202], there are three main categories for fairness notions: *i*) Statistical measures, *ii*) Similarity-based measures, and *iii*) Causal reasoning. Before going into a detailed overview of examples for bias detection approaches, we provide the definition of the most popular fairness metrics used in the approaches (discrimination discovery and fairness management) proposed in our corpus.

- *Demographic parity (or Statistical parity)* [233]: Both subjects of the protected and unprotected group have equal probability to be assigned to the positive predicted outcome.
- *Equality of opportunity (or False negative error balance)* [84]: A statistical group fairness notion. The model satisfies this definition if both subjects of protected and unprotected groups have equal false negative rate (FNR), the probability of an individual who is actually in a positive class to be assigned by the classifier a negative predictive value.
- *Disparate mistreatment (or Equalized odds)* [84, 222]: A statistical group fairness notion. The model satisfies this definition of fairness if subjects of both protected and unprotected groups have equal true positive rate (TPR) and false positive rate (FPR).
- *Counterfactual fairness* [120]: A causal reasoning, individual fairness notion. The algorithmic model is considered as fair when the prediction of an individual is the same even if the value of the sensitive variable changes. To validate this type of fairness, a causal model is used.

In addition to the above fairness notions, in ranking systems, such as RecSys and IR, a common type of bias is the *position bias* where users tend to consider only the items ranked in the top few positions [159]. The fairness notions that consider position bias are the producer or item-side fairness and the consumer or user-side fairness. Producer or item-side fairness focuses on the items that have been recommended so that similar items are ranked/recommended in a similar way. Consumer or user-side fairness focuses on the users who receive the ranking results or the recommendations. A similar group of users should all receive similar recommendations.

Next, we present the details described in the papers collected from the four communities for detecting any type of bias in a system using Auditing and Discrimination Discovery approaches.

4.1 Auditing Approaches for Bias Detection

The most common auditing approach used for bias detection involves humans (external testers, researchers, journalists or the end users) acting as the auditors of the system. In information retrieval systems, researchers usually perform an audit by submitting queries to search engine(s) and analyzing the results. For instance, Vincent et. al. [204] performed an audit on Google result pages, where six types of important queries (e.g., trending, expensive advertising) were analyzed. The goal was to examine the importance of user-generated content (UGC) on the Web, in terms of the quality of information that the search engines provide to users (i.e., if there was a bias in favor/penalizing such content). Similarly, Kay et. al. [105], Magno et. al. [132], and Otterbacher et. al. [147] submit queries to image search engines to study the perpetuation of gender stereotypes, while Metaxa et al. [139] consider the impact of gender and racial representation in image search results for common occupations. They compare gender and racial composition of occupations to that

reflected in image search and find evidence of deviations on both dimensions. They also compare the gender representation data with that collected earlier by Kay et al. [105], finding little improvement over time.

Another example of bias detection in a search engine via auditing is the work of Kilman-Silver et. al. [111] who examine the influence of geolocation on Web search (Google) personalization. They collect and analyze Google results for 240 queries over 30 days from 59 different GPS coordinates, looking for systematic differences. In addition, Robertson et. al. [169] audited Google search engine result pages (SERPs) collected by study participants for evidence of filter bubble effects. Participants in the study completed a questionnaire on their political leaning and used a browser extension allowing the researchers to collect their SERPs.

Kulshrestha et al. [118] propose an auditing technique where queries are submitted on Twitter, to measure bias on Twitter results as compared to search engines. The proposed technique considers both the input and output bias. Input bias allows the researchers to understand what a user would see if shown a set of random items relevant to her query. The output bias isolates the bias of the ranking mechanism. In addition, Johnson et. al. [101] investigate the demographic bias detection in Twitter results using as an auditing technique, the retrieval of geotagged content using Twitter API. Another example where researchers are the auditors is the study of Edelman et. al. [57] where the authors run an experiment to audit Airbnb applications to detect racial bias in ranked results, and more specifically, for African American names.

Another cluster of user-based studies in IR systems concerned the detection of perceived biases about search and/or during a search for information. In these studies, users are the auditors. For instance, Kodama et al. [112] assessed young people's mental models of the Google search engine, through a drawing task. Many informants anthropomorphized Google, and a few focused on inferring its internal workings. The authors called for a better understanding of young people's conceptions of search tools, so as to better design information literacy interventions and programs. In addition, Otterbacher et al. [148] described a study in which participants were the auditors for detecting perceived bias. They were shown image search results for queries on personal traits (e.g., "sensitive person", "intelligent person") and were asked to evaluate the results on a number of aspects, including the extent to which they were "biased."

Auditing approaches using ML algorithms have also been widely used. A situational testing auditing approach has been proposed by Luong et. al. [130], to detect discrimination against a specific group of individuals, using an ML algorithm. K-nearest neighbors was combined with the situation testing approach to identify a group of tuples with similar characteristics to a target individual. Zhang et. al. [226] proposed an improvement over the method of Luong et. al. [130], by engaging Causal Bayesian networks (CBNs), which are probabilistic graphical models used for reasoning and inference. For the development of a CBN, the causal structure of the dataset and the causal effect of each attribute on the decision are used to guide the identification of the similar tuples to a target individual. Robertson et. al. [170], present an auditing approach in the form of an opaque algorithm, called ''recursive algorithm interrogation'' used for detecting bias in search engines. The auto-complete functions of Google and Bing are treated as opaque algorithms. They recursively submitted queries, and their resulting child queries, in order to create a network of the algorithm's suggestions.

Hu et. al. [92] audited Google SERPs snippets, for evidence of partisanship where the generation of snippets is an opaque process. Moreover, Le et. al. [122] audit Google News Search for evidence of reinforcing a user's presumed partisanship. Using a sock-puppet technique, the browser first visited a political Web page, and then continued on to conduct a Google news search. The results of the audit suggested significant

reinforcement of inferred partisanship via personalization. In addition, Eslami et. al. [62] use a cross-platform audit technique that analyzed online ratings of hotels across three platforms, in order to understand how users perceived and managed biases in reviews.

In the HCI literature, auditing often involves characterizing the behavior of the algorithm from a user perspective. For instance, in Matsangidou and Otterbacher [135], the authors consider the inferences on physical attractiveness made by image tagging algorithms on images of people. They audited the output of four image recognition APIs on standardized portraits of people across genders and races. In a more recent work [12], the authors use auditing to understand machine behaviors in proprietary image tagging algorithms. The authors created a highly controlled dataset of people images, imposed on gender-stereotyped backgrounds. Evaluating five proprietary algorithms, they find that in three, gender inference is hindered when a background is introduced. Of the two that "see" both backgrounds and gender, it is the one with output that is most consistent with human stereotyping processes that is superior in recognizing gender. Another example is the work of Eslami et. al. [63], where the authors describe a qualitative study of online discussions about Yelp on the algorithm existence and opacity. The authors further enhanced the results by conducting 15 interviews with Yelp users who acted as auditors of the system, in an attempt to understand how the reviews filtering algorithm works.

Auditing approaches have also been used to detect bias in ML classification systems. For instance, in [24], the authors (developers) audit three automated facial analysis algorithms to detect any gender inequalities in the classification results. They found that males were classified more accurately than females in all the three algorithms and that all the algorithms performed worst on darker female subjects.

Recently, automated methods for auditing have been introduced to detect gender or race bias in the context of online housing advertisements and search engine rankings. Asplund et al. [6] propose the use of controlled "Sock-puppet" auditing techniques, which are automated systems that mimic user behavior in offline audits. They use these techniques to investigate gender-based and race-based discrimination in the context of online housing advertisements and any bias in search-result ranking. The authors use the definition of disparate impact to consider both application systems as fair or not.

4.2 Discrimination Discovery

A common approach for discrimination discovery is to compute metrics in order to detect any direct/indirect discrimination of specific groups in the data. Examples of metrics include absolute measures, conditional measures or statistical tests [233]. Absolute measures define the magnitude of discrimination over a dataset by taking into account the protected characteristics and the predicted outcome. Statistical tests, rather than measuring the magnitude of discrimination, indicate the presence or absence of discrimination at a dataset level. Conditional measures compute the magnitude of discrimination that cannot be explained by any non-protected characteristics of individuals. Fairness notions have also been used in many works as metrics for discrimination.

In Bellogin et. al. [15], the authors detect statistical biases in the evaluation metrics used in recommender systems that affect the effectiveness of the recommendations. They found out that there is sparsity and popularity bias on the evaluation metrics. Many works focus on investigating the racial bias in advertising recommendations systems. For instance, Sweeney [192] investigates the racial bias in advertising recommendations by an ad server when searching for particular names in Google and Reuters search engines. She

finds that ads for services providing criminal records on names were significantly more likely to be served if the name search was on a typically Black first name. Ali et al. [2], Speicher et. al [188] and Imana et al. [94] detected significantly skewed ad delivery on racial lines in Facebook ads for employment, financial services and housing. More specifically, in [94], the authors first build an audience that allows them to infer the gender of the ad recipients on the platforms that do not provide ad delivery statistics along gender lines, i.e., Facebook, Linkedin. They use this audience to distinguish between skew in ad delivery due to protected categories from the skew due to differences in qualifications among people in the targeted audience. Indirectly, they measure the "equality of opportunity" fairness notion.

Another example of bias detection in RecSys and online social networks is the work of Chackraborty et al. [33] who detect demographic bias in the input data of crowds in Twitter who make posts worthy of being recommended as trending. The bias is detected by comparing the characteristics of the trend promoters with the demographics of the general population of Twitter. Apart from demographic bias, political bias is very common in social networks. In Jiang et al. [98], the authors measure the fairness in social media contexts based on the fairness notions: *demographic parity* and the *equalized odds*. The authors detect political bias through content moderation. Bias in the social platform Facebook has also been assessed through reverse engineering of the Facebook API ranking algorithm using logistic regression in [89]. More specifically, the authors identify the features of a post that would affect its odds of being selected. Sentiment analysis reveals that there are significant differences in the sentiment word usage between the selected and non-selected posts.

In information retrieval systems, discrimination discovery is primarily used in user-focused studies. Weber and Castillo [208] conducted a study of user search habits, which involved a large-scale analysis of Web logs from Yahoo!. Using the logs, as well as users' profile information and US-census information (e.g., average income within a given zip code), the authors were able to characterize the typical behaviors of various segments of the population and detect any discrimination related to the users' sensitive demographic attributes. In a similar manner, Yom-Tov [221] used search query logs to characterize the differences in the way that users of different ages, genders and income brackets, formulate health-related queries. His driving concern was the ability to discover users with similar profiles, according to their demographic information (user cohorts), who are looking for the same information e.g., a health condition.

Pal et al. [150] considered the identification of experts in the context of a question-answering community. Their analysis revealed that as compared to other users with less expertise, experts exhibited significant selection biases in their engagement with content. They proposed to exploit this bias in a probabilistic model, to identify both current and potential experts. A method to identify selection bias, IMITATE, has also been proposed in Dost et al. [56]. IMITATE investigates the dataset's probability density, then adds generated points in order to smooth out the density and have it resemble a Gaussian, the most common density occurring in real-world applications.

In a study of information exposure on the Mendeley platform for sharing academic research, Thelwall and Maflahi [196] illustrated a *home-country* bias. Articles were significantly more likely to be read by users in the home country of the authors, as compared to users located in other countries. Chen et al. [35] investigated direct and indirect (implicit) gender-based discrimination in the context of resume search engines, by a system towards its users. Direct discrimination happens when the system explicitly uses the inferred gender or other attributes to rank candidates, while indirect discrimination is when the system unintentionally

discriminates against users (indirectly via sensitive attributes). The results suggested that the system under review indirectly discriminates against females, however, it does not implicitly use gender as a parameter.

Another method for detecting bias in search engine results involves the use of metrics that quantify bias in search engines [142]. A series of articles by Wikie et. al. [211–213] and a paper of Bashir and Rauber [13] investigates the identification retrieval bias in IR systems. Bashir and Rauber study the relationship between query characteristics and document retrievability using the TREC Chemical Retrieval track. In Wilkie and Azzopardi [212], they examined the issue of fairness vs. performance. Wilkie and Azzopardi [211] consider specific measures of retrieval bias and the correlation to the system performance. Wilkie and Azzopardi [213] consider the issue of bias resulting from the process of pooling in the creation of test sets.

A recent study [178] detects gender and race bias in the annotation process of training data of image databases used for facial analysis. The authors found that the majority of image databases rarely contain underlying source material for how those identities are defined. Further, when they are annotated with race and gender information, database authors rarely describe the process of annotation. Instead, classifications of race and gender are portrayed as insignificant, indisputable, and apolitical.

A set of works in the HCI domain analyzes crowdsourced data from the OpenStreetMap to detect any potential biases such as gender and geographic information bias [48, 161]. In a similar vein, two other studies run a crowdsourcing study to detect any bias on human versus algorithmic decision-making [11, 78]. Green and Chen [78] run a crowdsourcing study to examine the influence of algorithmic risk assessment to human decision-making, while Barlas et. al. [11] compared human and algorithmic generated descriptions of people images in a crowdsourcing study in an attempt to identify what is perceived as fair when describing the depicted person. The execution of a crowdsourcing study for detecting bias has also been used in IR systems [59, 127].

Many works study the problem of bias detection in textual data using data mining methods concerning specific protected groups. The typical approach is to extract association or classification rules from the data and to evaluate these rules according to discrimination of protected groups [157, 171]. For instance, Datta et. al. [49] analyse the gender discrimination in online advertising (Google ads) using ML techniques, to identify the gender-based ad serving patterns. Specifically, they train a classifier to learn differences in the served ads and to predict the corresponding gender. Similarly, Leavy et. al. [123] detect gender bias in training textual data by identifying linguistic features that are gender-discriminative, according to gender theory and sociolinguistics. Zhao et al. [229] detect gender bias in coreference resolution systems. They introduce a new benchmark dataset, WinoBias, which focuses on gender bias. They also use a data augmentation approach that in combination with existing word-embedding debiasing techniques, removes the gender bias demonstrated in the data. Madaan et al. [131] detect gender discrimination in movies using knowledge graphs and word embeddings after analyzing the data (using, for example, mentions of each gender in movies, emotions of the actors during the movies, occupation of each gender in the movies, screen time, etc.) In a similar vein, Ferrer et al. [67]propose a data-driven approach to discover and categorize language bias encoded on the vocabulary of online communities in the Reddit platform. They use word embeddings to discover the most biased words towards protected attributes, apply k-means clustering combined with a semantic analysis system to label the clusters, and use sentiment analysis to further specify biased words. Rekabsaz and his colleagues [164] also explore the detection of societal bias in word-embedding models by utilizing the first-order co-occurrence relations between the word and the representative concept words.

Islam et al. [95] introduce a collaborative filtering method to detect gender bias in social media. Their proposed method is called Neural Fair Collaborative Filtering (NFCF). They also use debiasing embeddings, and fairness interventions via penalty term.

A cluster of works in the IR domain addresses the detection of bias such as age-based bias, and text-frequency and stylistic biases in sentiment classification algorithms [51, 163, 184]. Other examples of detecting bias in classifiers that use sentiment analysis are the existence of offensive language or stereotyping of sensitive attributes in automated hate speech detection algorithms [7, 50] and the detection of cultural biases at Wikipedia pages using sentiment analysis [27]. Shandilya et al. [181] also detect the under-representation of sensitive attributes in the summarization algorithms.

Keyes [107] identified the problem of automatic gender recognition in HCI research and how the approaches followed until recently are discriminatory towards trans gendered people. For systems to be fair, Keyes [107] proposed alternative methods and the development of more inclusive approaches in the gender inference process and evaluation. Apart from automatic gender recognition, an additional significant advancement in the field of HCI is that of data-driven personas. Salminen et al. [175] investigated the presence of demographic bias in automatically generated, data-driven personas. They discovered that the more personas they generated, the more diverse the sample became in terms of gender and age representation. Practitioners who use data generated personas should consider the possibility of unintentional bias in the data they use, that consequently is transferred to the personas they generate.

Multiple other approaches have been proposed in ML literature that detect any discrimination in the data or classifier. Choi et al. [40] discover and mine discrimination patterns that refer to whether an individual is classified differently depending on whether some sensitive attributes were observed. The algorithm detects discrimination patterns in a Naive Bayes classifier using branch and bound search and removes them. It learns maximum likelihood parameters based on these parameters. Pedreshi et al. [157] use an opaque predictive model to extract frequent classification rules based on an inductive approach. Background knowledge is used to identify the groups to be detected as potentially discriminated. On the other hand, Zhang et. al. [227] use a causal Bayesian network and a learning structure algorithm to identify the causal factors for discrimination. The direct causal effect of the protected variable on the dependent variable represents the sensitivity of the dependent variable to changes in the discrimination grounds while all other variables are held fixed. They also detect discrimination in the prediction/classification outcome by computing the classification error rate (error bias). In a more recent work, Zucker et al. [234] introduce a new domain-specific programming language, the Arbiter for ML practitioners. It allows users to make guarantees about the level of bias in any produced models.

The notion of divergence [153], which estimates the difference in classification performance measures, has also been proposed as a metric to identify data subgroups in which a classifier performs differently. Pastor et al. [154] introduce the DivExplorer, an interactive visual analytics tool that identifies algorithmic bias using the divergence notion. An interactive system to detect fairness issues in the classifiers has also been proposed in [125]. The system is called DENOUNCER and it allows users to explore fairness issues for a given test dataset, considering different fairness notions. In addition, Nargesian et al. [143] detect the groups in the dataset that are unfairly treated by the classifier by developing an exploration-exploitation based strategy. Their approach captures the cost and approximations of group distributions in the given dataset.

In IR systems, a common type of bias is the cognitive or perception bias that arises from the manner in which information is presented to users, in combination with the user's own cognition and/or perception. For example, Jansen and Resnick [96] analyzed the behaviors of 56 participants engaged in e-commerce search tasks, with the goal of understanding users' perceptions of sponsored versus un-sponsored (organic) Web links. The links suggested by the search engine were manipulated in order to control content and quality. Even controlling for these factors, it was shown that users have a strong preference for organic Web links. In a similar vein, Bar-Ilan et al. [10] conducted a user experiment to examine the effect of position in a search engine results page (SERPs). Across a variety of queries and synthetic orderings of the results, they demonstrated a strong placement bias; a result's placement, along with a small effect on its source, is the main determinant of perceived quality. User perception is also examined in a study [139] where the authors consider people's impressions of occupations and sense of belonging in a given field when shown search results with different proportions of women and people of color. They find that both axes of representation as well as people's own racial and gender identities impact their experience of image search results. Gezici et al. [74] propose a new evaluation framework to measure bias in the content of SERPs (on political and controversial search topics) by measuring stance and ideological bias. They propose three novel fairness-aware measures of bias based on common IR utility-based evaluation measures.

Ryen White, of Microsoft Research, has published extensively on detecting users' perception bias during and after a search, particularly when trying to find information to answer health-related queries. In an initial work [209], a user study focused on finding yes-no answers to medical questions, showed that pre-search beliefs influence users' search behaviors. For instance, those with strong beliefs pre-search, are less likely to explore the results page, thus reinforcing the above-mentioned positioning bias. A follow-up study by White and Horvitz [210] looked more specifically at users' beliefs on the efficiency of medical treatments, and how these beliefs could be influenced by a Web search. An example of searching for user perception bias in recommender systems was presented in [172], where drivers' perceptions of the Uber application were investigated, taking into consideration drivers' profiles and their history performance.

4.3 Bias Detection Comparison

Table 6 summarizes the methods used for auditing and discrimination discovery within each of the research domains analyzed in this survey. In ML systems, bias detection is mostly done using discrimination or fairness metrics. Auditing in ML systems can be achieved by auditing software tools or when *developers/regulators* act as auditors of the algorithmic system. On the other hand, in IR, HCI and RecSys systems, *users* often act as auditors by submitting different queries in search engines and social networks or by taking the role of crowdworker in the crowdsourcing conducted studies. Discrimination discovery approaches used in IR, HCI and RecSys systems are similar to auditing but with a more concrete methodology on detecting bias.

5 FAIRNESS MANAGEMENT

The second set of tools used in mitigating algorithmic system bias concerns processes of *Fairness Management*. One consideration is to use fairness management approaches to mitigate the bias detected in any part of an algorithmic system. However, in order to make sure that an algorithmic system can be considered "fair," it is not enough to simply mitigate the detected bias – the design of the system should be "fairness-aware."

Domain	Problem	Solution Space	Reference(s)
		Bias Detection	
ML	Data/Model	Auditing	Automatic auditing tool [14, 174]
			Developers as auditors [24, 130, 227]
	Data		Discrimination/Fairness metrics [98, 222, 233]
	Data		Metrics[56, 125, 154]
	Data/Model/Output	Discrimination Discovery	ML methods
			[43, 49, 123, 157, 227, 234]
IR	User/Data/Output	Auditing	Submit queries to search engines/ Twitter
			[92, 105, 118, 122, 132, 147, 204]
	Model/User		Sock-puppet auditing[6]
	User/Data/Output	Discrimination Discovery	Analysis of Web logs
			[13, 35, 94, 150, 208, 211–213, 221]
	User/Data/Output		Word embedding[67, 95, 164]
	User/Third Party/Data		Crowdsourcing studies[59, 127]
	User/Third Party		Direct discrimination of perceived bias
			[10, 96, 209, 210]
HCI	Output/Model/User	Auditing	Analysing system behavior
			[101, 135]
	Data/User/Third Party	Discrimination Discovery	Crowdsourcing studies
			[11, 48, 78, 139, 161]
	Model/User		Use of ML methods[89, 178]
	Data/User		Data-driven personas[175]
RecSys	Data/User	Auditing	Developers as auditors
			[57, 62]
	Model/User		Sock-puppet auditing[6]
	User/Model/Output	Discrimination Discovery	Discrimination detection in advertising
			recommendation systems[2, 188, 192]
	Output/Model		Discrimination detection in evaluation metrics
			[15, 60]
	Output/User		Discrimination in social networks [34]

Table 6. Comparison of Discrimination Detection approaches across the four domains.

In this section, we provide details of the solution approaches proposed in the literature in each of the five fairness management categories.

5.1 Fairness Pre-processing Methods

An approach that is usually used to mitigate bias on the input or training data is the removal of sensitive attributes that may be involved in discrimination. However, in some cases, the inclusion of sensitive characteristics in the data may be beneficial to the design of a fair model [235]. To handle this issue, some approaches remove information about the protected variables from the training data but they also transform the training data using data mining methods. For instance, Kamiran and Calders [102] use a naive Bayes classifier to generate rankings of each observation in the training data based on its probability of belonging to the desirable class category. The outcome variable in the training data is adjusted until there is no remaining association between the protected variable and the intended outcome variable. The drawback of this solution is that it is limited to a binary outcome variable and the transformed training data cannot be used with other outcome variables.

Calders and Verwer [26] eliminate the above drawbacks by presenting three algorithms that transform (i.e., re-weight) the training data based on an objective function that is minimized when the outcomes from a model that fit to the transformed data are independent of the protected variable. This class is also restricted to binary outcome and protected variables. Data transformation approaches have also been proposed by Johndrow and Lum [100] and Zemel [224]. Johndrow et al. [100] suggest a statistical framework where the training data are transformed by mapping individuals to an input space that is mutually independent of specific groupings. In [224], the authors encode the data by mapping each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. The aim of this is to hide any sensitive information that can identify if the individual belongs to a protected group. Percy et al. [158] propose an approach to mitigate gender bias on gambling. The method uses gender data for training only, constructing separate models for each gender and combining trained models into an ensemble that does not require gender data once deployed.

Another frequent pre-processing technique is the use of directed acyclic graphs and causal reasoning that capture the dependencies between the features and their impact on the outcome. For instance, Zhang et al. [227] discover and prevent discrimination bias in decision support systems using a causal Bayesian network (BN) to identify pairs of tuples with similar characteristics from the dataset. Then, they generate a new dataset sampled from the learned BN. Cardoso et al. [121] also use a Bayesian network estimated from real-world data to generate biased data that are learned from real-world data. A data transformation method has also been applied to ensure fairness in RecSys [215]. The authors propose a new graph-model technique, the FairGo model, which ensures fairness for every recommender system by transforming the original embedding of user and items into a filtered embedding space based on the sensitive feature set. FairGo is model-agnostic and can be applied to multiple sensitive attributes.

Rather than adjusting/transforming the observations of the training data, other works use re-labeling. Cardoso et al. [121] propose the use of an auditing tool to repair the dataset by changing attribute labels. Kamiran and Calders [102] massage the data by swapping some of the labels in such a way that a positive outcome for the disadvantaged group is more likely and then they re-train the model. Feldman et al. [66] proposed the *disparate impact removal* solution approach that manipulates individual data dimensions in a way that depends on the protected attribute.

Similar techniques to data transformation, but that consider the selection of features, have been introduced in [32, 173]. Salazar et al. [173] use a multi-objective optimization algorithm for feature construction. They use this approach to generate more features that lead to both high accuracy and fairness by applying human understandable transformations. Celis et al. [32] develop a novel approach that takes as input a visibly diverse control set of images of people and uses this set as part of a procedure to select a set of images of people in response to a query. The goal is to have a resulting set that is more visibly diverse in a manner that emulates the diversity depicted in the control set.

Other popular fairness pre-processing methods are the re-sampling methods that generate a balanced dataset that will not under- or over-represent a particular protected group [51, 53, 76, 101, 184]. A similar approach is used in RecSys where a re-sampling method is used to balance the neighborhoods before producing recommendations [25] or re-balance the input data according to the protected attributes (e.g., gender) to produce a fair training dataset [130]. A re-sampling method has also been used by Sharma et al. [183]. They use a data augmentation technique that adds synthetic data for removing bias in the data.

The technique selectively adds only a subset of the synthetic points to create new augmented dataset to meet the fairness criteria while maintaining accuracy.

5.2 Fairness In-processing Methods

One category of the in-processing approaches is the use of an optimization method. Xiao et al. [217] suggest a multi-objective optimization framework optimizing fairness and social welfare simultaneously on group recommendation. The goal was to maximize the satisfaction of each group member while minimizing the unfairness between them. The results show that considering fairness in group recommendation can enhance the recommendation accuracy. A multi-objective optimization approach has also been proposed in [144] for fair allocations using two criteria, maximum fairness and efficiency. They propose a dynamic programming algorithm to construct an appropriate Pareto set.

Optimization approaches with fairness weights have also been used in recommender systems for two-sided marketplaces [138]. In that scenario, the developed recommendation systems should be fair on both the demand and supplier sides. The authors propose different recommendation policies that jointly optimize the relevance of recommendations to consumer (i.e., user) and fairness of representation of suppliers. Kusner et al. [120] focus on satisfying the counterfactual fairness as the notion of fairness. They capture the social biases that may arise towards individuals based on sensitive attributes. They provide optimization of fairness and prediction accuracy of the classifier using a causal model.

A second category of in-processing methods is the use of regularization methods. Yan and Howe [219] introduce the FairST, a fairness-aware demand prediction model for spatiotemporal urban applications. Two spatio-temporal fairness metrics have been introduced as a form of regularization to reduce discrimination in demographic groups. Kamishima et al. [103], also use a regularization approach that can be applied to any algorithmic model (classifier). They introduce a prejudice remover regularizer that enforces classifier's independence from sensitive attributes.

Instead of applying a regularization method, Rezaeil et al. [165] mitigate bias detected in any classifier by re-building the classifier and incorporating fairness constraints to the predictor. The method reshapes the predictions (output) for each group to satisfy the fairness constraints that consider the protected groups.

In ranking systems, IR and RecSys, in-processing approaches primarily explore the mitigation of bias in the ranking algorithms using learn-to-rank methods. For instance, Dai and his colleagues [47] propose a novel framework, Adversarial Imitation Click Model (AICM), which is based on imitation learning to address the exposure bias in click-models. Click-models rely on learning-to-rank, by studying how users interact with a ranked list of items. In [149], the authors consider both the selection and position bias in the rank-based results. They frame the problem as a counterfactual problem and adapt Heckerman's (rank) approach by combining it with position bias correction methods to correct both the selection and position bias. Yang and Ai [220] propose a fair and unbiased ranking method named Maximal Marginal Fairness (MMF) for dynamic learning to rank, to achieve both fairness and relevance in top-k results.

In a recent work [116], the authors introduce a fair rank aggregation framework for aggregating multiple rankings in a database, which can be applied to the databases of the ranking systems. It uses pairwise discordance to both compute closeness among consensus and base rankings and measure the advantage given to each group of candidates. Another fairness issue in ranking systems concerns the mitigation of bias in the PageRank algorithm. The authors in [199] provide a parity-based definition of fairness that

imposes constraints on the proportion of PageRank allocated to the members of each group. They validate the fairness notion of local and personalized PageRank fairness.

5.3 Fairness Post-processing Methods

The most well-known post-processing method used in ML literature is the re-labeling of the decision outcome. Pedreschi et al. [157] alter the confidence of classification rules inferred by the CPAR algorithm, whereas Kamiran et al. [102] re-label the class that is predicted at the labels of a decision tree. In [84], the authors propose a new fairness definition to optimally adjust any learned predictor so as to remove discrimination. Their framework constructs classifiers from any Bayes optimal regressor following a post-processing step that minimizes the loss in utility.

Additionally, in the literature of rank-based systems (i.e. IR and RecSys), post-processing methods focus on the re-ranking of the recommended or search results. In [126], the authors provide a re-ranking approach to mitigate the unfairness problem between active and inactive users by adding constraints over evaluation metrics. Experiments show that their approach improves group fairness of users in recommender systems, and also achieves better overall recommendation performance.

A re-ranking method has also been proposed by Karako and Manggala [104] who introduce a fairness-aware variation of the Maximal Marginal Relevance (MMR). The proposed method incorporates fairness in a recommender or search system by choosing a sample of labeled images, based on gender when retrieving untagged images similar to an input image or query. Mitigation of gender bias on image tagging has been proposed in [195] where the authors introduce the Guided Attention Image Captioning model (GAIC). The GAIC pipeline encourages the model to provide correct gender identification with high confidence when gender evidence in image is obvious. When gender evidence is vague or occluded, GAIC tends to describe the person with gender neutral words, such as "person" and "people." In addition, Zehlike et al. [223] and Singh and Joachims [186] propose fair top-k ranking algorithms for RecSys that makes the recommendations subject to group fairness criteria and constraints.

"Other" works in ranking systems propose methods to achieve the fairness of the general ranking results, rather than focusing on the top-k ranking. Patro and colleagues [155] propose the FairRec algorithm, which validates exhibiting the desired two-sided fairness, both consumer and producer fairness, by mapping the fair recommendation problem to a fair allocation problem. Kuhlman et al. [117] use an auditing methodology FARE (Fair Auditing based on Rank Error) for error-based fairness assessment of the ranking results. They propose three error-based fairness criteria, which are rank-appropriate, to assess the correctness of the rankings. In addition, Kirnap et al. [119] estimate four fair ranking metrics by acquiring group membership annotations for a sample of documents in its corpus.

5.4 Fairness Certification

Fairness certification methods are used to certify the fairness of the system using some constraints [31, 52, 110, 225] or by introducing new fairness metrics such as FACE and FACT [108], feature-apriori fairness, feature accuracy fairness and feature-disparity fairness [79]. Wu et al. [216] propose a framework that uses many of these fairness metrics as convex constraints that are directly incorporated into the classifier. They first present a constraint-free criterion (derived from the training data) that guarantees that any learned classifier will be fair according to the specified fairness metric. Thus, when the criterion is satisfied, there is

no need to add any fairness constraint into the classifiers. When the criterion is not satisfied, a constrained optimization problem is used to learn fair classifiers.

Hu et al. [93] propose a metric-free individual fairness based on the gradient contextual bandit algorithm that aims to maximize fairness. In [73], the authors use multi-objective clustering algorithms to maximize both accuracy and fairness and also to introduce diversity and transparency as constraints. Counterfactual fairness is another well-known metric used for certifying the fairness of the system. In [182], counterfactual explanations evaluate fairness with respect to a particular individual as well as the fairness of the model towards groups of individuals. They define the metric "burden" to evaluate group fairness. The burden is computed taking into consideration how much the fitness measure differs for counterfactuals generated for specific groups of individuals. Cruz Cortes et al. [46] use population fairness metrics: Predictive Parity and Error Rate Balance. They propose a simple agent-based model to detect any discrimination inequalities in an arrest-sentence system. Group fairness has been used as a definition in recommender systems for group recommendations as well. Kaya et al. [106] define a new metric for group fairness called Group Fairness Aware Recommendations (GFAR) considering the fairness of the top-N ranked items. GFAR defines top-N ranking as fair when the relevance of each of the top-N items, to the group members is 'balanced' across the group members.

In information retrieval systems, researchers often focus on user evaluation to certify the fairness of the system. Mitra et al. [141] presented the first large-scale study of users' interactions with the auto-complete function of Bing. Through an analysis of query logs, they found evidence of a position bias (i.e., users were more likely to engage with higher-ranked suggestions). They were also more likely to engage with auto-complete suggestions after having typed at least half of their query. In a follow-up study, Hofmann et al. [90] conducted an eye-tracking study with Bing users. In half of their queries, users were shown the ranked auto-complete suggestions while in the other half of queries, the suggestions were random. The authors confirmed the position bias in the auto-complete results, across both ranking conditions. They found that the quality of the auto-complete suggestions affected search behaviors; in the random setting users visited more pages in order to complete their search task. Another popular fairness certification method is simply to raise users' awareness. Epstein et al. [61] develop solutions for the Search Engine Manipulation Effect (SEME), citing recent evidence of its impact on the views of undecided voters in the political context. In a large-scale online experiment with 3,600 users in 39 countries, they showed that manipulating the rankings in political searches can shift users' expressed voting preferences by up to 39%. However, providing users with a "bias alert," which informed them that "the current page of search rankings you are viewing appears to be biased in favor of [name of candidate]," reduced the shift to 22%. They found that this could be reduced even further when more detailed bias alerts were provided to users. Nonetheless, they reported that SEME cannot be completely eliminated with this type of intervention, and suggested instituting an "equal-time" rule such as that used in traditional media advertisements.

"Other" works use alternative approaches rather than the computation of specific metrics to certify an ML system. For instance, Fang et al. [64] certify the fairness of a classifier by constructing fairgroups, considering the feature importance to the decision variable. Individuals with similar features are grouped into clusters. This approach adopts the notion of fairness related to disparate impact, which affects individuals with at least one protected feature. In addition, Kilbertus et al. [109] provide fairness learning and certification without access to users' sensitive data. To achieve this, they use an encrypted version of sensitive data,

privacy constraints and decision verification by employing secure multi-party computation (MPC) methods. The use of techno-moral graphs for certifying ML algorithmic systems was also suggested in [97]. The authors argue that a three-dimensional conceptual space can be used to map machine learning algorithmic projects in terms of the morality of their respective and constitutive ground-truth practices. Such techno-moral graphs may, in turn, serve as equipment for greater governance of machine learning algorithms and systems.

5.5 Fairness Perception

Woodruff et al. [214] explore, in a qualitative study, the perception of algorithmic fairness by populations that have been marginalized. In particular, they consider how race and low socioeconomic status was used in stereotyping and adapting services to those involved. Most participants were not aware of algorithmic unfairness even though they had experienced discrimination in their daily lives. Brown et al. [22] also present a qualitative study for understanding the public's perspective on algorithmic decision-making in public services. They discovered that many participants mentioned discrimination and bias based on race, ethnicity, gender, location, and socioeconomic status. A descriptive approach for identifying the notion of perceived fairness for machine learning was suggested by Srivastava et al. [189]. They argued that the perceived fairness of the user is the most appropriate notion of algorithmic fairness. Their results show that the formal measurement, demographic parity, most closely matches the perceived fairness of the users and that in cases when the stakes are high, accuracy is more important than equality.

Perceived fairness on algorithmic decision-making is also explored in [207] where the authors conduct an online experiment to better understand perceptions of fairness, focusing on three sets of factors: algorithm outcomes, algorithm development and deployment procedures, and individual differences. They find that people rate the algorithm as more fair when the algorithm predicts in their favor, even surpassing the negative effects of describing algorithms that are very biased against particular demographic groups. This effect is moderated by several variables, including participants' education level, gender, and several aspects of the development procedure. These findings suggest that systems that evaluate algorithmic fairness through users' feedback must consider the possibility of "outcome favorability" bias.

In another study, the authors identify perception bias in borderline fact-checking messages [152]. The authors conduct both a quantitative and qualitative study by conducting semi-supervised user interviews to learn the user experience and perception of different fact-checking conditions. In a recent work [134], the authors introduce a network-centric fairness perception function that can be viewed as a local measure of individual fairness.

In addition, Maxwell et al. [136] investigated the influence of result diversification on users' search behaviors. Diversification can reduce search engine biases by exposing users to a broader coverage of information on their topic of interest. A within-subject study with 51 users was performed, using the TREC AQUAINT collection. Two types of search tasks - ad hoc versus aspectual - are assigned to each user using a non-diversified IR system as well as a diversified system. Results indicated significant differences in users' search behaviors between the two systems, with users executing more queries, but examining fewer documents when using the diversified system on the aspectual (i.e., more complex) task.

5.6 Fairness Management Comparison

As displayed in Table 7, in ML algorithmic systems, the most popular techniques are data re-sampling, removal of sensitive attributes and data transformation to mitigate bias in the data, optimization and regularization approaches to mitigate bias during the model training and re-labeling of the outcome decision to mitigate bias on the output of the system. In ranking systems such as RecSys and IR, the most popular approaches are re-sampling for mitigating data bias, learning to rank methods to mitigate bias in the ranking algorithms and re-ranking methods as for modifying the ranking outcomes. Two approaches that are common in RecSys and ML communities are the data transformation (fairness pre-processing) and optimization approaches (fairness in-processing). In the HCI community, since the *user* is the main stakeholder, most of the papers examine the user perception on fairness. Approaches to mitigate bias referred to the use of a human-in-the-loop on the decision-making [22]. Fairness certification techniques use fairness constraints or defining new fairness notions, i.e., counterfactual fairness and metrics for certifying the fairness of systems in all the four research domains. In IR, some studies also use user evaluation to certify the fairness of the system.

6 EXPLAINABILITY MANAGEMENT

The increasing use of algorithms in decision-making – especially for critical applications – has lead to policies requiring clearer accountability for algorithmic decision-making, such as the European Union's General Data Protection Regulation, and its "Right to Explanation" [77]. Doshi-Velez and Kim [55] argue that interpretability can help us evaluate if a model is biased or discriminatory by explaining the incompleteness that produces some kind of unquantified bias. On the other hand, Selbst and Barocas [179] and Kroll et al. [115] have demonstrated that even if a model is fully transparent, it might be hard to detect and mitigate bias due to the existence of correlated variables.

According to Eslami et al. [63], full transparency is neither necessary nor desirable in most systems. One reason is that full transparency may negatively affect users' information privacy [37]. Moreover, users often need to be provided with details on the decisions made, and not simply with explanations of the outcome. A good example is a qualitative study [22] in which participants requested not only information concerning how the algorithm under study took decisions, but also the parameters upon which the decisions were taken.

Friedrich and Zanker [70] classify explainability into two types: *white-box* and *black-box*. *How* explanations are white-box explanations of the input, output and the process leading to the particular outcome. They provide information focusing on the system's reasoning and data source, which enhances the user satisfaction of the system. *Why* explanations treat the systems as non-transparent and they do not provide any information on how a system works. Instead, they give justifications for outcomes and explain the motivations behind the system, to fill the gap between the user's needs and the system's goals. Rader et al. [162] proposed two additional types of explainability, "What" and "Objective". *What* explanations only reveal the existence of algorithmic decision-making without providing any additional information on how the system works. This type of explainability aims to raise the users' awareness of the algorithm. *Objective* explains the process of the development of the system and its potential improvement with the objective of preventing or mitigating bias in the system.

Fairness Management			
Domain	**Problem**	**Solution Space**	**Reference(s)**
ML	Data	Fairness Pre-processing	Removal of protected attributes & Data Transformation [26, 100, 158, 224]
			Causal BN[121, 227]
			Data Re-labeling [66, 102]
			Re-sampling methods [101, 182]
	Model	Fairness In-processing	Regularization approach [103, 219]
			Optimization approach [144, 173]
			Constraints[165]
	Model/Output		Counterfactual fairness [120]
	Third Party/Output	Fairness Post-processing	Altering of labels [84, 102, 157]
	User/Third Party	Fairness Perception	[134, 189]
	Data/Model/Output	Fairness Certification	Fairgroups [64]
			Counterfactual Fairness [109, 182]
			Techno-moral graphs[97]
			Fairness Constraints/Metrics [31, 46, 52, 79, 108, 110, 216, 225]
IR	Data	Fairness Pre-processing	Data sampling [51, 53, 76, 184]
	Model	Fairness In-processing	Learn-to-rank methods [47, 116, 149, 220]
	Output	Fairness Post-processing	Re-ranking[104, 117, 119, 126]
	Model/Output/User	Fairness Certification	[61, 90, 141]
	User/Output	Fairness Perception	[136, 152]
HCI	Data	Fairness Pre-processing	Data sampling [101]
			Data transformation [32]
	Output	Fairness Perception	Human-in-the-loop [22]
	User/Output		Metrics[206]
	Output/User	Fairness Certification	[124, 214]
RecSys	Data	Fairness Pre-processing	Data sampling [25, 104, 130]
			Data transformation [215]
	Model/Output		Optimization approaches [138, 217]
	Model	Fairness In-processing	Learn-to-rank [116, 220]
	Output	Fairness Post-processing	Re-ranking[104, 155, 186, 223]
	Model/Output	Fairness Certification	Metrics [106]

Table 7. Comparison of fairness management methods in the different domains.

Important aspects for personalized explanations in algorithmic systems include the presentation format of the different types of explanations (e.g., graphical, textual, bullet points), the length of each explanation, and the adopted vocabulary if natural language is used for the explanations. The range of explanations is based on the domain; for example, decisions in the health domain are more critical than in movie recommendations and may need a wider range of explanations of how a system derives its predictions/classifications. Regarding the presentation format, Eiband et al. [58] proposed a participatory design methodology for incorporating transparency in the design of user interfaces such as to make intelligent systems more transparent and explainable. The process used in the design methodology consists of two main parts. The first part defines the content of an explanation (what to explain) while the second focuses on the presentation format of the explanation (how to explain). In a similar vein, Binnis et al. [16] classify a set of explanation styles into four categories based on the type of information they would like to present to the end user:

- **Input influence style**: A set of input variables are presented to the user along with their positive or negative influence on the outcome.
- **Sensitivity style**: A sensitivity analysis shows how much each of the input values would have to differ in order to change the outcome (e.g., class).
- **Case-based style**: A case from the model's training data that is most similar to the decision outcome is presented to the user.
- **Demographic style**: Using this style, the system presents to the user statistics regarding the outcome classes for people in the same demographic categories as the decision subject, e.g., based on age, gender, income level or occupation.

Recent surveys on interpretable machine learning methods and techniques can be found in [3, 81]. In the following sections, we briefly identify the main explainability approaches used in ML and RecSys systems.

6.1 Model Explainability

Model explainability techniques are primarily used to explain the process of an opaque ML model such as a neural network or a deep learning model. These techniques usually use a transparent model to mimic the model's behavior and be interpretable by humans. For instance, some works use a decision tree to mimic the behavior of a non-transparent model such as a neural network [21, 45, 99, 114] and tree ensemble models [54, 75, 177, 193, 230]. The use of decision trees for explaining neural networks was first presented in [45] where the *TREPAN* network implements the algorithmic process of the neural network and returns the representations of the model. Chipman et al. [38] use decision trees as an interpetable predictor model for tree ensemble models by summarizing the forest of trees through clustering, and use the associated clusters as explanation models. A similar technique is the use of decision rules to explain a non-transparent model, for instance, by extracting rules from a trained model such as a neural network (NN), and then using the NN to refine existing rules [45, 99, 231].

More recent works use ontologies to represent and integrate knowledge to the model in order to enhance human understandability. An example is the recent extension of TREPAN [42] that uses and integrates knowledge in the form of ontologies in the decision tree extraction to enhance human understandability on decision trees. In addition, Ribeiro et al. [166] use ontologies to explain neural networks (NN). They build small classifiers that map a neural network model's internal state to concepts from an ontology, enabling the generation of symbolic justifications for the output of NN. An alternative approach has been proposed in [19], where the authors present the Bayes-TREX framework, which uses Bayesian inference techniques to explain NN based on the whole dataset, not only the test data. Bayes-TREX takes as input the whole data and finds in-distribution examples that trigger various model behaviors across several contexts.

In addition to the aforementioned approaches for explainability of non-transparent ML algorithms, many articles, especially in the domain of recommender systems, propose some approaches for interpreting the ranking (recommender) algorithms. In such systems, the authors aim to provide explanations based on user opinions and evaluation of previous purchases, rather than on the analysis of the ranking algorithm [205]. The aim is to provide personalized explanations by selecting the most appropriate explanation style. Nunes et al. [145] presented a systematic review on explanations for recommendations in decision support systems where they proposed a taxonomy of concepts that are required for providing explanation. According to

Tintarev and Masthoff [197], there are seven purposes for providing explanations in a recommender system: transparency, scrutability, trust, effectiveness, persuasiveness, satisfaction and efficiency. Park et al. [151] introduce the J-RECS, a recommendation model-agnostic method that generates personalized justifications based on various types of product and user data (e.g., purchase history and product attributes). Although most of the surveyed works in RecSys provide explanations based on user data, a recent work [72] propose some metrics for measuring explainability and transparency of the ranking algorithm.

A good example that shows the connection of explainability with fairness perception is the recent work of Anik et al. [5]. In this work, the authors explore the concept of data-centric explanations for ML systems that describe the training data to end users. They first investigate the potential utility of such an approach, including the information about training data that participants find most compelling. In a second study, the authors investigate reactions to the explanations across four different system scenarios. Their results suggest that data-centric explanations have the potential to impact how users judge the trustworthiness of a system and to assist users in assessing fairness.

6.2 Outcome (or Post-hoc) Explainability

Outcome explainability approaches attempt to provide an interpretation for the outcome generated by the model. A recent work focuses on providing both local and pedagogical explanations for the output of ML models [133]. Pedagogical explanations are those that teach something about how the model works rather than attempting to represent it directly. Outcome explanations are divided into: *visual explanations*, *local explanations* and *feature relevance explanations* techniques.

Local explanation approaches are the Local Interpretable Model-Agnostic Explanations (LIME) [167] and its variations [80, 168, 200]. The explanations in LIME are only provided through linear models and their respective feature importance. Anchors is another local explanation method proposed by Ribeiro et al. [168] that uses decision rules for explaining the model sufficiently.

A post-hoc global explainability method has been proposed in [9]. A SEPA framework has been introduced that incorporates post-hoc global explanation methods for image classification tasks. SEPA uses understandable semantic concepts (entities and attributes) that are obtained via crowd-sourcing from local interpretability saliency maps.

An example of feature relevant explanation approach is the *ExplainD*, a framework presented in [160] for interpreting the outcome of any non-transparent model. ExplainD uses generative additive models (GAM) to weight the importance of the input features. A unified framework of the class of six existing additive feature importance methods, the SHAP (SHapley Additive exPlanations) has also been introduced in [129]. SHAP assigns each feature an importance value for a particular prediction to interpret the predictions.

According to Slack et al. [187], post-hoc explanation techniques that rely on the input, such as LIME and SHAP, are not reliable since they do not take into consideration the bias in the model. In [187], the authors proposed a scaffolding technique that scaffolds any biased classifier in a way that its input data remain biased but the generated post-hoc explanations do not reflect the underlying bias.

Other examples of feature relevant explanations include the approach used in Horne et al. [91] for explaining the spread of fake news and misinformation online. They used an AI assistance framework for providing these explanations to users. This was been shown to improve the user perception of bias and reliability on online news consumption. In another approach, Henelius et al. [87] search for a group of

attributes of which the interactions affect the predictive performance of a given classifier, and they evaluate the importance of each group of attributes using the fidelity metric. In addition, Vidovic et al. [203] propose the measure of feature importance (MFI), which is model-agnostic and can be applied to any type of model. Feature-relevant explanations are also used in [1], where the authors suggest the DIFF operator, a declarative operator that unifies explanation and feature selection queries with relational analytics workloads.

Another widely known category of outcome explainability approaches is the use of counterfactual explanations, which is a special case of feature-related explanations [182, 206]. In [182], they propose the CERTIFAI model-agnostic technique that provides counterfactual explanations using a genetic algorithm. The user can use counterfactual explanations to understand the importance of the features. In [206], the authors introduce Lewis, an open-source software that provides counterfactual explanations for the decision-making of an algorithm at the global, local and contextual level. For individuals negatively impacted by the algorithm's decision, it provides actionable resources to change the outcome of the algorithm in the future.

Visualization model-specific techniques are used to inspect the training process of a deep neural network (DNN) behavior on images [17, 218, 232]. In these works, a Saliency Mask (SM) is used as the interpretable local predictor e.g., a part of an image. Similarly, Fong et al. [68] propose a framework of explanations as meta-predictors for explaining the outcome of deep learning models. The meta-predictor is a rule that predicts the response of the model to certain inputs such as highlighting the salient parts of an image. Another set of works use saliency masks to incorporate the DL network activation into their visualizations [180, 185, 230].

In RecSys, outcome explainability approaches are used to explain the recommendations to the user. One category of explainability techniques for RecSys are the ones that explain the latent factors that contribute to the decision outcome based on the collection of users' interests and items' characteristics such as Explicit Factor Models [228] and Tensor factorization [36]. Other approaches for explaining recommendations are based on the use of knowledge graphs that relate the items' characteristics and users' behavior, based on their past interactions with the items [29, 85]. Visual explanations have recently been used in RecSys to justify the recommendation process in combination with giving more control to the user in a specific context of an interactive social recommender system [198]. By conducting a user study, the authors investigate how the addition of user control and explainability affect the user perception, user experience, and user engagement. Based on the results, the best user experience happens when there is full explainability and control.

6.3 Explainability Management Comparison

Table 8 provides a comparison of the solutions focusing on Explainability Management. Explainability approaches have primarily been developed in the context of ML algorithms and systems. The best known methods for explaining the model decision-making process use interpretable models to mimic the behavior of black-box models, i.e., decision trees, decision rules and ontologies. Methods for explaining the decision outcome include feature-relevance, local and global explainability and visualization methods.

There is also a growing literature on explainability within the HCI community. These works suggest that explainability and judgement of the outcome or decision of the system should be provided in order to enhance the trust of the end user in the system. Also in HCI, we found a few works that connect explainability to fairness perception. Finally, explainability approaches have also been widely discussed in RecSys and IR systems. The difference between these approaches and the ones used in ML are that they take into

consideration the user's perception and have the specific goal of increasing the trust of the end user in the system. The most popular explainability techniques in the RecSys and IR literature are the visualization methods (outcome explainability) that have been applied to justify the ranking results.

Explainability Management			
Domain	**Problem**	**Solution Space**	**Reference(s)**
ML	Model	Model Expainability	Use of decision tree [38, 54, 75, 114, 177, 193, 230]
	Model		Use of decision rules [45, 99, 128]
	Model		Ontologies [19, 42, 166]
	Output	Outcome Explainability	Local explanations [167, 168, 200]
	Output/User		Visualization methods [17, 68, 180, 185, 218, 230, 232]
	Output/User		Counterfactual explanations[182, 206]
			Feature-relevance explanations[1, 87, 187, 203]
IR	Output/User	Outcome Explainability	Global explanaions[9]
HCI	User/Data	Model Explainability	data-centric explanations[5]
	Output/Data	Outcome Explainability	Feature-relevance explanation [91]
	User/Output		Taxonomy of explanations & Styles[16, 58, 70]
	User/Output		Raise user awareness [162]
RecSys	Model/User	Model Explainability	Taxonomy of concepts [145]
	Model/User		Based on user opinions [37, 205]
	Output/User		Personalized explanations [151]
	Output/User		Knowledge graph [29, 86]
	Output/User	Output Explainability	Visualization methods [20, 113, 198, 201]

Table 8. Comparison of explainability management approaches for the different research domains.

7 BRINGING IT ALL TOGETHER

Our survey was intentionally broad, as we aimed to provide a "fish-eye view" of this complex topic. We did not restrict our review to the literature on fairness and/or discriminatory bias in a social sense; rather, we considered articles describing the problems and solutions surrounding bias, which affect any number of attributes including the quality of information provided by a system.

Sources of Bias. Articles reviewed in our survey mentioned at least one of seven problematic components and/or points at which bias can be detected. These are shown in Figure 4, which also groups them into four main types: data bias, user bias, processing bias, and human bias. In reality, all biases are at least indirectly *human biases*; for instance, datasets and processing techniques are created by humans. However, we believe that it is helpful to distinguish the biases that are directly introduced into the system by humans, such as third-party biases, those resulting from conflicting fairness constraints, as well as those due to the choices of the developer. User bias is distinguished from other human biases in our framework; as detailed in the literature, users can both introduce bias (e.g., in biased input), but can also perceive bias in the output. Finally, Figure 4 also incorporates, at a high level, the three steps in a comprehensive solution to mitigate algorithmic bias: bias detection, fairness management and explainability. Figure 4 presents an overview of the problem and solution spaces revealed by the survey. This framework integrates the concepts presented

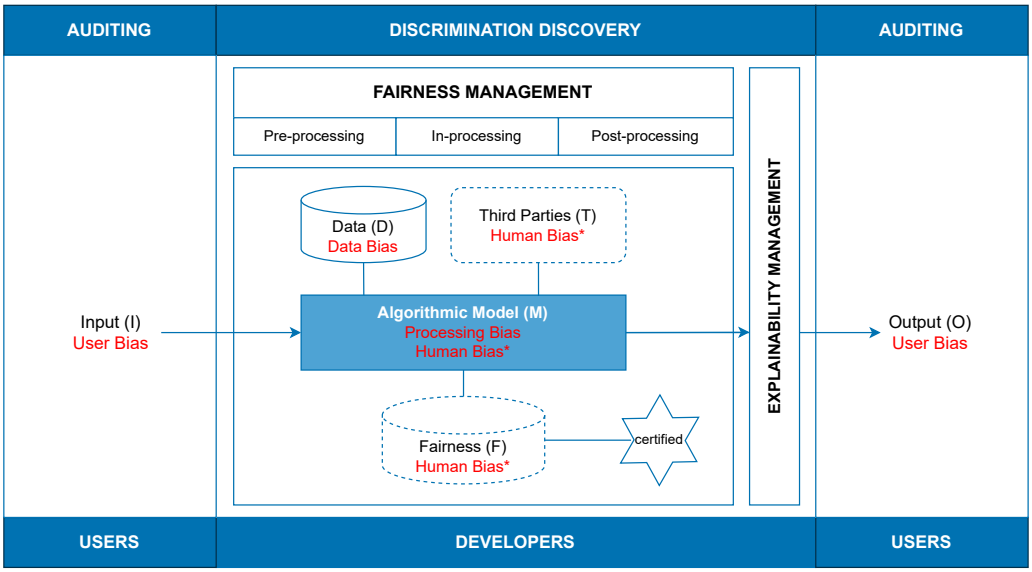

Fig. 4. A fish-eye view of mitigating algorithmic bias: problems, stakeholders, solutions.

earlier on, the components of a system that can be problematic (Figure 1), and the solutions described across communities (Figure 3).

Next, Table 9 depicts the cross-fertilization between the four communities that we reviewed, in terms of realizing comprehensive solutions for mitigating bias. All four communities use all three steps for mitigating bias in different parts of an algorithmic system. However, the interrelationships between the communities is primarily based on the stakeholders involved in implementing each solution. In addition, there are similarities and differences across the specific solution approaches used in each step of mitigating bias within the different communities. For instance, data re-sampling is an approach used in all four communities for fairness pre-processing, while learn-to-rank (fairness in-processing) and re-ranking (fairness post-processing) are approaches used for fairness management only in the communities of IR and RecSys, which are concerned with ranking systems. In the following paragraphs, we give a more detailed view of the interrelationships of the four communities in each step of mitigating bias, considering both the solution approaches and the stakeholders involved in implementing them.

Bias Detection: In most of the articles in our repository, across all four communities, auditing is typically done by the observers of the system. It should also be noted that within ML, beyond involving the model, inputs and outputs, auditing can also involve the generation of biased datasets for conducting a black-box audit. As presented in Table 9, in three domains, bias detection, in general (both auditing and discrimination discovery), is done by observers. The exception is ML, where developers implement automated auditing and discrimination discovery tools, and also observers use auditing to detect fairness issues in the system. In addition, in the RecSys community, developers sometimes implement the auditing process or use discrimination (or fairness) metrics to detect bias as in the ML community.

	Stakeholders			
Domains	**Developers**	**Users**	**Observers**	**Indirect Users**
ML	Bias Detection Fairness management Model Explainability	 Fairness Perception Outcome Explainability	Auditing	 Fairness Perception Outcome Explainability
IR	 Fairness Pre-processing Fairness In-processing Fairness Post-processing 	 Perceived Fairness Outcome Explainability	Bias Detection Fairness Certification	Auditing Perceived Fairness Outcome Explainability
RecSys	Bias detection Fairness Pre-processing Fairness In-processing Fairness Post-processing Model Explainability	 Fairness Perception Fairness Certification Outcome Explainability	Bias Detection	Auditing Fairness Perception Outcome Explainability
HCI	 Model Explainability	 Fairness Perception Outcome Explainability	Bias Detection Fairness Certicifation Fairness Pre-processing	Auditing Fairness Perception Outcome Explainability

Table 9. Cross-fertilization between research communities.

Fairness Management: The issue of ensuring that people and/or groups of people are treated fairly by an algorithmic system was found to be of interest to researchers across all domains considered. However, the tools stakeholders have at their disposal vary. For instance, in three communities (ML, RecSys, IR), developers are the ones who implement pre-processing, in-processing and post-processing methods to mitigate fairness issues in different parts of the algorithmic system, as also presented in Table 9. Specifically, in ML systems, developers are involved both in the development of the system and manage fairness of the system by developing inside the box. In contrast, in HCI, the system observers manage fairness by observing the system's behavior or the output of the system. In addition, the users of the system participate in the conducted studies for managing system fairness concerning the users' perception. In IR and RecSys, apart from the developers, observers are also involved in certifying that the algorithm is fair.

Explainability Management: With respect to the transparency of the algorithmic system, a set of explainability approaches has been introduced in the literature, to encourage trust in the system by the end user, which primarily concerns the HCI and ML communities. In HCI articles, the most appropriate presentation and format of explainability is examined for enriching the transparency of the systems and the trust of the end user. Moreover, multiple papers study specific explainability approaches for explaining the matching/ranking algorithm in RecSys and IR ranking systems. As shown in Table 9, in ML systems, the developers implement algorithms or methods for providing transparency for the black-box model and outcome whereas in RecSys and IR ranking systems, personalized explanations focus on the user and indirect users of the system. In HCI, the observer, in collaboration with the user, conducts experimental studies using various explanation presentation styles and in some cases, personalized explanations for providing the user with some transparency of the system. The exception is one HCI approach, where the developer(s) provides data-centric explanations (Model Explainability).

Affected Attributes. From the articles reviewed in this survey, we can conclude that there are two types of attributes that are affected by the bias and fairness issues in an algorithmic system:

- Attributes describing the *social world*; in particular, socio-cultural characteristics of people such as gender, age, language and national origin.
- Attributes describing *information*, with the critical question being how well the attributes describe real-world events and phenomena, i.e., the quality and/or credibility of information provided as input to the algorithm, or as output to the user.

As mentioned, the attributes describing information are most clearly connected to the explainability management approaches. The other solutions (auditing, discrimination discovery and fairness management) typically address bias that concerns attributes of the real-world and in some cases, information as well. This is the case because in explainability management, people are interested in the process by which information is built while in the other cases, they are interested in the actual discrimination. Based on that, we can also conclude that the three steps of mitigating bias are complementary and can be applied to address different facets of the problem within an algorithmic system.

Limitations. We must note some challenges faced when reviewing the literature on mitigating algorithmic bias. First, the field is becoming highly interdisciplinary. It was often difficult to categorize the articles we collected into one domain; for instance, RecSys researchers often publish in HCI venues, or even ACM FAccT. Thus, while we aimed to collect articles from across four domains, one should keep in mind that there is some overlap between them. Thus, it was more difficult than expected to characterize how each community has contributed to the work on addressing algorithmic bias. This challenge, however, does not affect the development of a "fish-eye view" on the field. In addition, the classifications of solutions that we provide is driven by empirical evidence as we discovered it by the extensive, state-of-the-art works reviewed in the survey. Still, there are cases we do not capture, which are outside of our classification. Any classification scheme has its own foundation issues, which have long-term effects, as they influence the validity of the classification in the long term. Thus, it becomes obvious that the entire issue of bias and the solution(s) to bias should be placed into the context of diversity, taking into account local cultures and problems [65], which will be examined in a future work.

Secondly, the framework presented in Figure 4 does not yet explicitly incorporate *accountability* into the solutions for mitigating algorithmic bias. Because we focused on literature in the information and computer sciences, studying articles describing particular algorithms and/or systems, the issue of accountability was not often discussed. Going forward, the literature search could be expanded into law and the social sciences as to further investigate the role of the Observer/ Regulator in the landscape of solutions.

8 CONCLUSION

In this survey, we provided a "fish-eye view" of research to date on the mitigation of bias in any type of algorithmic system. With the aim of raising awareness of biases in user-focused, and algorithm-focused systems, we examined studies conducted in four different research communities: information retrieval (IR), human-computer interaction (HCI), recommender systems (RecSys) and machine learning (ML). We outlined a classification of the solutions described in the literature for detecting bias as well as for mitigating the risk of bias and managing fairness in the system. Multiple stakeholders, including the developer (or anyone

involved in the pipeline of a system's development), and various system observers (i.e., stakeholders who are not involved in the development, but who may use, be affected by, oversee, or even regulate the use of the system) are involved in mitigating bias. In future work, we aim to further refine the various roles of individual stakeholders and the relationships between them.

A second consideration to be explored, is that while many solutions described in the literature have been formalized (e.g., discrimination detection methods, fairness management, internal certification), there are many other issues surrounding *perceived fairness*. The perceived fairness of the user is somewhat subjective and it is not clear how the internal, formal processes relate to users' perceptions of the systems and their value judgements. To this end, it is important to emphasize the particular role of explainability management for bias mitigation. Specifically, in this context, explainability can be viewed as a means rather than an end; complex algorithmic systems can become more transparent to users, the more interpretable their models and outcomes are. Clearly, explainability has a tight relationship to the user's perception of fairness.

Finally, in this survey, we recorded the attribute(s) affected by the problematic system in each of the reviewed domains and found that there are two key types of attributes affected by the problematic system: attributes describing the world and attributes describing information. Based on that, explainability management solutions mitigate bias that only affects information, while bias detection and fairness management mitigate bias that affects the attributes describing the social world. In future work, we aim to treat the two types of bias (social world, information) independently.

ACKNOWLEDGMENTS

This project is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105 (CyCAT). Otterbacher and Kleanthous are also supported by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under agreement No. 739578 (RISE).

REFERENCES

[1] Firas Abuzaid, Peter Kraft, Sahaana Suri, Edward Gan, Eric Xu, Atul Shenoy, Asvin Ananthanarayan, John Sheu, Erik Meijer, Xi Wu, Jeff Naughton, Peter Bailis, and Matei Zaharia. 2018. DIFF: a relational interface for large-scale data explanation. *Proc. VLDB Endow.* 12, 4 (Dec. 2018), 419–432. https://doi.org/10.14778/3297753.3297761

[2] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv:1904.02095 [cs]* (April 2019). http://arxiv.org/abs/1904.02095 arXiv: 1904.02095.

[3] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. [n.d.]. Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery* n/a, n/a ([n. d.]), e1424. https://doi.org/10.1002/widm.1424 _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1424.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445736

[6] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 24–35. https://ojs.aaai.org/index.php/ICWSM/article/view/7276

[7] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-based Generalizations. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 49–59. https://doi.org/10.1145/3308558.3313504 event-place: San Francisco, CA, USA.

[8] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. 2021. 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. *The VLDB Journal* 30, 5 (Sept. 2021), 739–768. https://doi.org/10.1007/s00778-021-00671-8

[9] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2021. What do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1937–1948. https://doi.org/10.1145/3442381.3450069

[10] Judit Bar' Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results—A user study. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 135–149. https://doi.org/10.1002/asi.20941

[11] Pinar Barlas, Styliani Kleanthous, Kyriakos Kyriakou, and Jahna Otterbacher. 2019. What Makes an Image Tagger Fair?. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '19)*. ACM, New York, NY, USA, 95–103. https://doi.org/10.1145/3320435.3320442 event-place: Larnaca, Cyprus.

[12] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (Jan. 2021), 232:1–232:31. https://doi.org/10.1145/3432931

[13] Shariq Bashir and Andreas Rauber. 2011. On the relationship between query characteristics and IR functions retrieval bias. *Journal of the American Society for Information Science and Technology* 62, 8 (2011), 1515–1532. https://doi.org/10.1002/asi.21549

[14] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943 [cs]* (Oct. 2018). http://arxiv.org/abs/1810.01943 arXiv: 1810.01943.

[15] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. 2017. Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Inf. Retr.* 20, 6 (Dec. 2017), 606–634. https://doi.org/10.1007/s10791-017-9312-z

[16] Reuben Binns, Max Van Kleek, Veale Michael, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *CHI '18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM New York, NY, USA '©2018. https://doi.org/10.1145/3173574.3173951

[17] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba. 2016. Visualbackprop: visualizing cnns for autonomous driving. *arXiv:1611.05418 [cs]* (Nov. 2016). http://arxiv.org/abs/1611.05418 arXiv: 1611.05418.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *In Advances in neural information processing systems*. Curran Associates Inc. , USA '©2016, Barcelona, Spain, pp. 4349–4357.

[19] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah. 2021. BAYES-TREX: a Bayesian Sampling Approach to Model Transparency by Example. (2021), 11423–11432. https://ojs.aaai.org/index.php/AAAI/article/view/17361

[20] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 150–159. https://doi.org/10.1145/3287560.3287583 event-place: Atlanta, GA, USA.

[21] Olcay Boz. 2002. Extracting Decision Trees from Trained Neural Networks. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*. ACM, New York, NY, USA, 456–461. https://doi.org/10.1145/775047.775113 event-place: Edmonton, Alberta, Canada.

[22] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 41:1–41:12. https://doi.org/10.1145/3290605.3300271 event-place: Glasgow, Scotland Uk.

[23] C. E. Buckley, Darrin L. Dimmick, Ian M. Soboroff, and Ellen M. Voorhees. 2007. Bias and the Limits of Pooling for Large Collections | NIST. *Information Retrieval* (July 2007). https://www.nist.gov/publications/bias-and-limits-pooling-large-collections

[24] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

https://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[25] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*. PMLR, 202–214.

[26] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[27] Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology* 62, 10 (2011), 1899–1915. https://doi.org/10.1002/asi.21577

[28] Dallas Card, Michael Zhang, and Noah A. Smith. 2019. Deep Weighted Averaging Classifiers. ACM New York, NY, USA ’©2019, Atlanta, GA, USA, pp. 369–378. https://doi.org/10.1145/3287560.3287595

[29] Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi, and William Cohen. 2017. Explainable Entity-based Recommendations with Knowledge Graphs. *arXiv:1707.05254 [cs]* (July 2017).

[30] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[31] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. ACM New York, NY, USA ©2019, Atlanta, GA, USA, Pp. 319–328. https://doi.org/10.1145/3287560.3287586

[32] L. Elisa Celis and Vijay Keswani. 2020. Implicit Diversity in Image Summarization. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020), 139:1–139:28. https://doi.org/10.1145/3415210

[33] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. 2017. Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In *Eleventh International AAAI Conference on Web and Social Media*. pp. 22–31. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15680

[34] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. 2019. Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19)*. ACM, New York, NY, USA, 129–138. https://doi.org/10.1145/3287560.3287570 event-place: Atlanta, GA, USA.

[35] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, New York, NY, USA, 651:1–651:14. https://doi.org/10.1145/3173574.3174225

[36] Tsai-Wei Chen and S. Shyam Sundar. 2018. This App Would Like to Use Your Current Location to Better Serve You: Importance of User Assent and System Transparency in Personalized Mobile Services. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, New York, NY, USA, 537:1–537:13. https://doi.org/10.1145/3173574.3174111 event-place: Montreal QC, Canada.

[37] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C. Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Trans. Inf. Syst.* 37, 2 (Jan. 2019), 16:1–16:28. https://doi.org/10.1145/3291060

[38] Andrea Chipman, Edward I. George, R. E.F, and McCullochDepartment. 2007. Making sense of a forest of treesH.

[39] Junghoo Cho and Sourashis Roy. 2004. Impact of Search Engines on Page Popularity. In *Proceedings of the 13th International Conference on World Wide Web (WWW ’04)*. ACM, New York, NY, USA, 20–29. https://doi.org/10.1145/988672.988676 event-place: New York, NY, USA.

[40] YooJung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy Van den Broeck. 2020. Learning Fair Naive Bayes Classifiers by Discovering and Eliminating Discrimination Patterns. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 06 (April 2020), 10077–10084. https://doi.org/10.1609/aaai.v34i06.6565 Number: 06.

[41] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[42] Roberto Confalonieri, Tillman Weyde, and Tarek R Besold. 2020. TREPAN Reloaded: A Knowledge-driven Approach to Explaining Black-box Models. *Santiago de Compostela* (2020), 8.

[43] Bo Cowgill and Catherine Tucker. 2017. *Algorithmic bias: A counterfactual perspective*. Technical Report. Working Paper: NSF Trustworthy Algorithms. 3 pages.

[44] Mark Craven and Jude W. Shavlik. 1994. Using Sampling and Queries to Extract Rules from Trained Neural Networks. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning (ICML’94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 37–45. http://dl.acm.org/citation.cfm?id=3091574.3091580 event-place: New Brunswick, NJ, USA.

[45] Mark Craven and Jude W. Shavlik. 1996. Extracting Tree-Structured Representations of Trained Networks. In *Advances in Neural Information Processing Systems 8*. MIT Press, 24–30. http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

[46] Efrén Cruz Cortés and Debashis Ghosh. 2020. An Invitation to System-wide Algorithmic Fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 235–241. https://doi.org/10.1145/3375627.3375860

[47] Xinyi Dai, Jianghao Lin, Weinan Zhang, Shuai Li, Weiwen Liu, Ruiming Tang, Xiuqiang He, Jianye Hao, Jun Wang, and Yong Yu. 2021. An Adversarial Imitation Click Model for Information Retrieval. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 1809–1820. https://doi.org/10.1145/3442381.3449913

[48] Maitraye Das, Brent Hecht, and Darren Gergle. 2019. 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)*. ACM, New York, NY, USA, 563:1–563:14. https://doi.org/10.1145/3290605.3300793 event-place: Glasgow, Scotland Uk.

[49] A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. 598–617. https://doi.org/10.1109/SP.2016.42

[50] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Eleventh International AAAI Conference on Web and Social Media*. Montreal, Canada, 512 – 515. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

[51] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 412:1–412:14. https://doi.org/10.1145/3173574.3173986 event-place: Montreal QC, Canada.

[52] Christos Dimitrakakis, Yang Liu, David Parkes, and Goran Radanovic. 2018. Bayesian Fairness. https://hal.inria.fr/hal-01953311

[53] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729 event-place: New Orleans, LA, USA.

[54] Pedro Domingos. 1998. Knowledge discovery via multiple models. *Intelligent Data Analysis* 2, 1-4 (Jan. 1998), 187–202. https://doi.org/10.1016/S1088-467X(98)00023-7

[55] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[56] Katharina Dost, Katerina Taskova, Patricia Riddle, and Jörg Wicker. 2020. Your Best Guess When You Know Nothing: Identification and Mitigation of Selection Bias. IEEE Computer Society, 996–1001. https://doi.org/10.1109/ICDM50108.2020.00115

[57] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics* 9, 2 (April 2017), 1–22. https://doi.org/10.1257/app.20160213

[58] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961 event-place: Tokyo, Japan.

[59] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654 event-place: Marina Del Rey, CA, USA.

[60] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[61] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 42:1–42:22. https://doi.org/10.1145/3134677

[62] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. AAAI Press, 62–71. https://experts.illinois.edu/en/publications/be-careful-things-can-be-worse-than-they-appear-understanding-bia

[63] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes Towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 494:1–494:14.

https://doi.org/10.1145/3290605.3300724 event-place: Glasgow, Scotland Uk.

[64] Boli Fang, Miao Jiang, Pei-yi Cheng, Jerry Shen, and Yi Fang. 2020. Achieving Outcome Fairness in Machine Learning Models for Social Decision Problems. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 444–450. https://doi.org/10.24963/ijcai.2020/62

[65] Jahna Otterbacher Tim Draws Fausto Giunchiglia, Styliani Kleanthous. 2021. Transparency Paths - Documenting the Diversity of User Perceptions. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (2021-06-21) *(UMAP '21)*. Association for Computing Machinery. https://doi.org/10.1145/3450614.3463292

[66] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311 event-place: Sydney, NSW, Australia.

[67] Xavier Ferrer, Tom van Nuenen, Jose M. Such, and Natalia Criado. 2021. Discovering and Categorising Language Biases in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 2021), 140–151. https://ojs.aaai.org/index.php/ICWSM/article/view/18048

[68] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3449–3457. https://doi.org/10.1109/ICCV.2017.371 arXiv: 1704.03296.

[69] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[70] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* 32, 3 (2011), 90–98.

[71] George W Furnas. 2006. A fisheye follow-up: further reflections on focus+ context. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 999–1008.

[72] Abraham Gale and Amélie Marian. 2020. Explaining monotonic ranking functions. *Proc. VLDB Endow.* 14, 4 (Dec. 2020), 640–652. https://doi.org/10.14778/3436905.3436922

[73] Sainyam Galhotra. 2020. Reliable Clustering with Applications to Data Integration. In *PhD@VLDB*.

[74] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Inf Retrieval J* 24, 2 (April 2021), 85–113. https://doi.org/10.1007/s10791-020-09386-w

[75] Robert D. Gibbons, Giles Hooker, Matthew D. Finkelman, David J. Weiss, Paul A. Pilkonis, Ellen Frank, Tara Moore, and David J. Kupfer. 2013. The CAD-MDD: A Computerized Adaptive Diagnostic Screening Tool for Depression. *The Journal of clinical psychiatry* 74, 7 (July 2013), 669–674. https://doi.org/10.4088/JCP.12m08338

[76] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2010. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. 1–9. https://doi.org/10.1109/INFCOM.2010.5462078

[77] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a Right to Explanation. *AI Magazine* 38, 3 (Oct. 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741 Number: 3.

[78] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563 event-place: Atlanta, GA, USA.

[79] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 903–912. https://doi.org/10.1145/3178876.3186138 event-place: Lyon, France.

[80] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. *Local Rule-Based Explanations of Black Box Decision Systems*. Technical Report. http://arxiv.org/abs/1805.10820 arXiv: 1805.10820.

[81] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (Aug. 2018), 93:1–93:42. https://doi.org/10.1145/3236009

[82] Aniko Hannak, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1914–1933. https://doi.org/10.1145/2998181.2998327 event-place: Portland, Oregon, USA.

[83] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, and et al. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and

Explainability in the Cloud. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Aug 2021). https://doi.org/10.1145/3447548.3467177

[84] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in neural information processing systems* (Oct. 2016). http://arxiv.org/abs/1610.02413 arXiv: 1610.02413.

[85] Reinhard Heckel and Michail Vlachos. 2016. Interpretable recommendations via overlapping co-clusters. *CoRR* abs/1604.02071 (2016). arXiv:1604.02071 http://arxiv.org/abs/1604.02071

[86] Reinhard Heckel, Michail Vlachos, Thomas Parnell, and Celestine Dunner. 2017. Scalable and interpretable product recommendations via overlapping co-clustering. *arXiv:1604.02071 [cs]* (May 2017). http://arxiv.org/abs/1604.02071 arXiv: 1604.02071.

[87] Andreas Henelius, Kai Puolamaki, Henrik Bostrom, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery* 28, 5-6 (Sept. 2014), 1503–1529. https://doi.org/10.1007/s10618-014-0368-8

[88] Birger Hjørland. 2002. Domain analysis in information science. *Journal of documentation* (2002).

[89] Justin Chun-Ting Ho. 2020. How biased is the sample? Reverse engineering the ranking algorithm of Facebook's Graph application programming interface. *Big Data & Society* 7, 1 (Jan. 2020), 2053951720905874. https://doi.org/10.1177/2053951720905874 Publisher: SAGE Publications Ltd.

[90] Kajta Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. 2014. An Eye-tracking Study of User Interactions with Query Auto Completion. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 549–558. https://doi.org/10.1145/2661829.2661922 event-place: Shanghai, China.

[91] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adali. 2019. Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone?. In *In Proceedings of the International AAAI Conference on Web and Social Media (1)*, Vol. 13. 247 – 256.

[92] Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 693–704. https://doi.org/10.1145/3308558.3313654 event-place: San Francisco, CA, USA.

[93] Q. Hu and H. Rangwala. 2020. Metric-Free Individual Fairness with Cooperative Contextual Bandits. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE Computer Society, Los Alamitos, CA, USA, 182–191. https://doi.org/10.1109/ICDM50108.2020.00027

[94] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 3767–3778. https://doi.org/10.1145/3442381.3450077

[95] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3779–3790. https://doi.org/10.1145/3442381.3449904

[96] Bernard J. Jansen and Marc Resnick. 2006. An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links During Ecommerce Web Searching. *J. Am. Soc. Inf. Sci. Technol.* 57, 14 (Dec. 2006), 1949–1961. https://doi.org/10.1002/asi.v57:14

[97] Florian Jaton. 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society* 8, 1 (Jan. 2021), 20539517211013569. https://doi.org/10.1177/20539517211013569 Publisher: SAGE Publications Ltd.

[98] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020. Reasoning about Political Bias in Content Moderation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (April 2020), 13669–13672. https://doi.org/10.1609/aaai.v34i09.7117 Number: 09.

[99] Ulf Johansson and Lars Niklasson. 2009. Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, Nashville, TN, USA, 238–244. https://doi.org/10.1109/CIDM.2009.4938655

[100] James E. Johndrow and Kristian Lum. 2019. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (March 2019), 189–220. https://doi.org/10.1214/18-AOAS1201

[101] Isaac Johnson, Connor McMahon, Johannes Schoning, and Brent Hecht. 2017. The Effect of Population and "Structural" Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1167–1178. https://doi.org/10.1145/3025453.3026015 event-place: Denver, Colorado, USA.

[102] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[103] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*. Springer Berlin Heidelberg, 35–50.

[104] Chen Karako and Putra Manggala. 2018. 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)*. ACM, New York, NY, USA, 23–28. https://doi.org/10.1145/3213586.3226206 event-place: Singapore, Singapore.

[105] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520 event-place: Seoul, Republic of Korea.

[106] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/3383313.3412232

[107] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357

[108] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. *The World Wide Web Conference on - WWW '19* (2019), 2907–2914. https://doi.org/10.1145/3308558.3313559 arXiv: 1903.11719.

[109] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. *arXiv:1806.03281 [cs, stat]* (June 2018). http://arxiv.org/abs/1806.03281 arXiv: 1806.03281.

[110] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017*. http://arxiv.org/abs/1609.05807

[111] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 121–127. https://doi.org/10.1145/2815675.2815714 event-place: Tokyo, Japan.

[112] Christie Kodama, Beth St Jean, Mega Subramaniam, and Natalie Greene Taylor. 2017. There'sa creepy guy on the other end at Google!: engaging middle school students in a drawing activity to elicit their mental models of Google. *Springer Netherlands* 20, 5 (Oct. 2017), 403–432. https://doi.org/10.1007/s10791-017-9306-x

[113] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 379–390. https://doi.org/10.1145/3301275.3302306 event-place: Marina del Ray, California.

[114] R. Krishnan, G. Sivakumar, and P. Bhattacharya. 1999. Extracting decision trees from trained neural networks. *Pattern Recognition* 32, 12 (Dec. 1999), 1999–2009. https://doi.org/10.1016/S0031-3203(98)00181-2

[115] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable Algorithms. *University of Pennsylvania Law Review* 165 (2016), 633–706. https://heinonline.org/HOL/P?h=hein.journals/pnlr165&i=648

[116] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank aggregation algorithms for fair consensus. *Proc. VLDB Endow.* 13, 12 (Aug. 2020), 2706–2719. https://doi.org/10.14778/3407790.3407855

[117] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443 event-place: San Francisco, CA, USA.

[118] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 417–432. https://doi.org/10.1145/2998181.2998321 event-place: Portland, Oregon, USA.

[119] Omer Kurnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 1065–1075. https://doi.org/10.1145/3442381.3450080

[120] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4066–4076. http://papers.nips.cc/paper/6995-

counterfactual-fairness.pdf

[121] Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. 2019. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. ACM, New York, NY, USA, 437–444. https://doi.org/10.1145/3306618.3314262 event-place: Honolulu, HI, USA.

[122] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring Political Personalization of Google News Search. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 2957–2963. https://doi.org/10.1145/3308558.3313682 event-place: San Francisco, CA, USA.

[123] Susan Leavy. 2018. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (GE '18)*. ACM, New York, NY, USA, 14–16. https://doi.org/10.1145/3195570.3195580 event-place: Gothenburg, Sweden.

[124] Min Kyung Lee and Su Baykal. 2017. 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)*. ACM, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230 event-place: Portland, Oregon, USA.

[125] Jinyang Li, Yuval Moskovitch, and H. V. Jagadish. 2021. DENOUNCER: detection of unfairness in classifiers. *Proc. VLDB Endow.* 14, 12 (July 2021), 2719–2722. https://doi.org/10.14778/3476311.3476328

[126] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866

[127] Y. L. Lin, C. Trattner, P. Brusilovsky, and D. He. 2015. The impact of image descriptions on user tagging behavior: A study of the nature and functionality of crowdsourced tags. *Journal of the Association for Information Science and Technology* 66 (Sept. 2015), 1785–1798. http://d-scholarship.pitt.edu/25927/

[128] Jianjun Lu, Shozo Tokinaga, and Yoshikazu Ikeda. 2006. Explanatory rule extraction based on the trained neural network and the genetic programming. https://doi.org/10.15807/jorsj.49.66

[129] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[130] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. 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 (KDD '11)*. Association for Computing Machinery, New York, NY, USA, 502–510. https://doi.org/10.1145/2020408.2020488

[131] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Conference on Fairness, Accountability and Transparency*. 92–105. http://proceedings.mlr.press/v81/madaan18a.html

[132] Gabriel Magno, Camila Souza Araujo, Wagner Meira Jr., and Virgilio Almeida. 2016. Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. *arXiv:1609.05413 [cs]* (Sept. 2016). http://arxiv.org/abs/1609.05413 arXiv: 1609.05413.

[133] David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. 2011. Performance of classification models from a user perspective. *Decision Support Systems* 51, 4 (2011), 782–793.

[134] Farzan Masrour, Pang-Ning Tan, and Abdol-Hossein Esfahanian. 2020. Fairness Perception from a Network-Centric Perspective. *arXiv:2010.05887 [cs]* (Oct. 2020). http://arxiv.org/abs/2010.05887 arXiv: 2010.05887.

[135] Maria Matsangidou and Jahna Otterbacher. 2019. What Is Beautiful Continues to Be Good. In *Human-Computer Interaction '€“ INTERACT 2019 (Lecture Notes in Computer Science)*. Springer International Publishing, 243–264.

[136] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2019. The impact of result diversification on search behaviour and performance. *Information Retrieval Journal* (May 2019), 1 – 25. https://doi.org/10.1007/s10791-019-09353-0

[137] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[138] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off Between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 2243–2251. https://doi.org/10.1145/3269206.3272027 event-place: Torino, Italy.

[139] Danaë Metaxa, Michelle A. Gan, Su Goh, Jeff Hancock, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 26:1–26:23. https://doi.org/10.1145/3449100

[140] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 2930–2939. https://doi.org/10.1109/CVPR.2016.320

[141] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*. ACM Press, Gold Coast, Queensland, Australia, 1055–1058. https://doi.org/10.1145/2600428.2609508

[142] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring Search Engine Bias. *Inf. Process. Manage.* 41, 5 (Sept. 2005), 1193–1205. https://doi.org/10.1016/j.ipm.2004.05.005

[143] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2021. Tailoring data source distributions for fairness-aware data integration. *Proc. VLDB Endow.* 14, 11 (July 2021), 2519–2532. https://doi.org/10.14778/3476249.3476299

[144] Trung Thanh Nguyen and Jörg Rothe. 2020. Approximate Pareto Set for Fair and Efficient Allocation: Few Agent Types or Few Resource Types. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 290–296. https://doi.org/10.24963/ijcai.2020/41

[145] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[146] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2, 3 (2019). https://doi.org/10.3389/fdata.2019.00013

[147] J. Otterbacher, J. Bates, and P. D. Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025727

[148] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 933–936. https://doi.org/10.1145/3209978.3210094 event-place: Ann Arbor, MI, USA.

[149] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-rank Systems. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1863–1873. https://doi.org/10.1145/3366423.3380255

[150] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. 2012. A Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Transactions on Information Systems (TOIS)* 30, 2 (Jan. 2012), 10. https://doi.org/10.1145/0000000.0000000

[151] Namyong Park, Andrey Kan, Christos Faloutsos, and Xin Luna Dong. 2020. J-Recs: Principled and Scalable Recommendation Justification. *arXiv:2011.05928 [cs]* (Nov. 2020). http://arxiv.org/abs/2011.05928 arXiv: 2011.05928.

[152] Sungkyu Park, Jamie Yejean Park, Hyojin Chin, Jeong-han Kang, and Meeyoung Cha. 2021. An Experimental Study to Understand User Experience and Perception Bias Occurred by Fact-checking Messages. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 2769–2780. https://doi.org/10.1145/3442381.3450121

[153] Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD/PODS '21)*. Association for Computing Machinery, New York, NY, USA, 1400–1412. https://doi.org/10.1145/3448016.3457284

[154] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. 2021. How divergent is your data? *Proc. VLDB Endow.* 14, 12 (July 2021), 2835–2838. https://doi.org/10.14778/3476311.3476357

[155] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1194–1204. https://doi.org/10.1145/3366423.3380196

[156] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Integrating Induction and Deduction for Finding Evidence of Discrimination. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL '09)*. ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/1568234.1568252 event-place: Barcelona, Spain.

[157] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*. 581–592.

[158] Christian Percy, AS d'Avila Garcez, Simo Dragicevic, and Sanjoy Sarkar. 2020. Lessons learned from problem gambling classification: Indirect discrimination and algorithmic fairness. In *Proc. AAAI Fall Symposium, AI for Social Good, Washington DC, USA*.

[159] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (Oct. 2021). https://doi.org/10.1007/s00778-021-00697-y

[160] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. 2006. Visual Explanation of Evidence in Additive Classifiers. In *In Proceedings of the National Conference on Artificial Intelligence*, Vol. 21. 8.

[161] Giovanni Quattrone, Licia Capra, and Pasquale De Meo. 2015. 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)*. ACM, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2675133.2675235 event-place: Vancouver, BC, Canada.

[162] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 103:1–103:13. https://doi.org/10.1145/3173574.3173677 event-place: Montreal QC, Canada.

[163] Abdelhalim Rafrafi, Vincent Guigue, and Patrick Gallinari. 2012. Coping with the Document Frequency Bias in Sentiment Classification. In *ICWSM*.

[164] Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021. Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 2021), 549–560. https://ojs.aaai.org/index.php/ICWSM/article/view/18083

[165] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. Fairness for Robust Log Loss Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (April 2020), 5511–5518. https://doi.org/10.1609/aaai.v34i04.6002 Number: 04.

[166] Manuel de Sousa Ribeiro and Joao Leite. 2021. Aligning Artificial Neural Networks and Ontologies towards Explainable AI. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (May 2021), 4932–4940. https://ojs.aaai.org/index.php/AAAI/article/view/16626 Number: 6.

[167] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, San Francisco, CA, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778 arXiv: 1602.04938.

[168] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High Precision Model-Agnostic Explanations. In *In Thirty-Second AAAI Conference on Artificial Intelligence*. 9.

[169] Ronald E. Robertson, Lisa Friedland, Kenneth JOSEPH, David Lazer, Christo Wilson, and Shan Jiang. 2018. Auditing Partisan Audience Bias within Google Search. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. 1–22. https://doi.org/10.1145/3274417

[170] Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. ACM, New York, NY, USA, 235–244. https://doi.org/10.1145/3292522.3326047 event-place: Boston, Massachusetts, USA.

[171] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (Nov. 2014), 582–638. https://doi.org/10.1017/S0269888913000039

[172] Alex Rosenblat and Luke Stark. 2016. *Algorithmic Labor and Information Asymmetries: A Case Study of Uber'€™s Drivers*. SSRN Scholarly Paper ID 2686227. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2686227

[173] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. 2021. Automated feature engineering for algorithmic fairness. *Proc. VLDB Endow.* 14, 9 (May 2021), 1694–1702. https://doi.org/10.14778/3461535.3463474

[174] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv:1811.05577 [cs]* (Nov. 2018). http://arxiv.org/abs/1811.05577 arXiv: 1811.05577.

[175] Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2019. Detecting Demographic Bias in Automatically Generated Personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, LBW0122:1–LBW0122:6. https://doi.org/10.1145/3290607.3313034 event-place: Glasgow, Scotland Uk.

[176] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association*. Seattle, WA.

[177] Vitaly Schetinin, Jonathan E. Fieldsend, Derek Partridge, Timothy J. Coats, Wojtek J. Krzanowski, Richard M. Everson, Trevor C. Bailey, and Adolfo Hernandez. 2007. Confident Interpretation of Bayesian Decision Tree Ensembles

for Clinical Applications. *IEEE Transactions on Information Technology in Biomedicine* 11, 3 (May 2007), 312–319. https://doi.org/10.1109/TITB.2006.880553

[178] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 058:1–058:35. https://doi.org/10.1145/3392866

[179] Andrew D. Selbst and Solon Barocas. 2018. *The Intuitive Appeal of Explainable Machines*. SSRN Scholarly Paper ID 3126971. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3126971

[180] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, 618–626. https://doi.org/10.1109/ICCV.2017.74

[181] Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of Extractive Text Summarization. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 97–98. https://doi.org/10.1145/3184558.3186947 event-place: Lyon, France.

[182] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York NY USA, 166–172. https://doi.org/10.1145/3375627.3375812

[183] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York NY USA, 358–364. https://doi.org/10.1145/3375627.3375865

[184] Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M. Rush. 2018. Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Stockholm, Sweden.

[185] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]* (Dec. 2013). http://arxiv.org/abs/1312.6034 arXiv: 1312.6034.

[186] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088 event-place: London, United Kingdom.

[187] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830

[188] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabriccio Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In *FAT 2018 - Conference on Fairness, Accountability, and Transparency*, Vol. 81. New-York, United States, 1–15. https://hal.archives-ouvertes.fr/hal-01955343

[189] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. 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 '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[190] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research* 11 (Jan. 2010), 1–18. https://doi.org/10.1145/1756006.1756007

[191] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[192] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Queue* 11, 3 (March 2013), 10:10–10:29. https://doi.org/10.1145/2460276.2460278

[193] Hui Fen Tan, Giles Hooker, and Martin T. Wells. 2016. Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. *arXiv:1611.07115 [cs, stat]* (Nov. 2016). http://arxiv.org/abs/1611.07115 arXiv: 1611.07115.

[194] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2017. Detecting Bias in Black-Box Models Using Transparent Model Distillation. *ArXiv* abs/1710.06169 (2017).

[195] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating Gender Bias in Captioning Systems. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 633–645. https://doi.org/10.1145/3442381.3449950

[196] Mike Thelwall and Nabeil Maflahi. 2015. Are scholarly articles disproportionately read in their own country? An analysis of mendeley readers. *Journal of the Association for Information Science and Technology* 66, 6 (June 2015), 1124–1135. https://doi.org/10.1002/asi.23252

[197] N. Tintarev and J. Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[198] Chun-Hua Tsai and Peter Brusilovsky. 2021. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 591–627.

[199] Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Panayiotis Tsaparas, Ilias Kleftakis, and Nikos Mamoulis. 2021. Fairness-Aware PageRank. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3815–3826. https://doi.org/10.1145/3442381.3450065

[200] Ryan Turner. 2016. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. 1–6. https://doi.org/10.1109/MLSP.2016.7738872

[201] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. 351–362.

[202] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[203] Marina M.-C. Vidovic, Nico Görnitz, Klaus-Robert Muller, and Marius Kloft. 2016. Feature Importance Measure for Non-linear Learning Algorithms. In *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Barcelona, Spain. http://arxiv.org/abs/1611.07567 arXiv: 1611.07567.

[204] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. In *In Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 505–5016.

[205] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 165–174. https://doi.org/10.1145/3209978.3210010 event-place: Ann Arbor, MI, USA.

[206] Paul Y. Wang, Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Demonstration of generating explanations for black-box algorithms using Lewis. *Proc. VLDB Endow.* 14, 12 (July 2021), 2787–2790. https://doi.org/10.14778/3476311.3476345

[207] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[208] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA, 523–530. https://doi.org/10.1145/1835449.1835537

[209] Ryen W. White. 2014. Belief dynamics in web search. *Association for Information Science and Technology* 65, 11 (Nov. 2014), 2165–2178. https://doi.org/10.1002/asi.23128

[210] Ryen W. White and Eric Horvitz. 2015. Belief Dynamics and Biases in Web Search. *ACM Transactions on Information Systems* 33, 4 (May 2015), 1–46. https://doi.org/10.1145/2746229

[211] Colin Wilkie and Leif Azzopardi. 2014. Best and Fairest: An Empirical Analysis of Retrieval System Bias. In *In European Conference on Information Retrieval (LNCS)*, Vol. 8416. Springer, Cham, 13–25. https://doi.org/10.1007/978-3-319-06028-6_2

[212] Colin Wilkie and Leif Azzopardi. 2014. A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. Shanghai, China, 81–90. https://doi.org/10.1145/2661829.2661948

[213] Colin Wilkie and Leif Azzopardi. 2017. Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable?. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Singapore, 2375–2378. https://doi.org/10.1145/3132847.3133135

[214] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 656:1–656:14. https://doi.org/10.1145/3173574.3174230 event-place: Montreal QC, Canada.

[215] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning Fair Representations for Recommendation: A Graph-based Perspective. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 2198–2208. https://doi.org/10.1145/3442381.3450015

[216] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. In *The World Wide Web Conference*. ACM New York, NY, USA ©2019, San Francisco, CA, USA, 3356–3362. https://doi.org/10.1145/3308558.3313723

[217] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 107–115. https://doi.org/10.1145/3109859.3109887 event-place: Como, Italy.

[218] Kelvin Xu, Jimmy Ba Lei, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, S. Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Vol. 37. Lille, France, 2048–2057. https://dl.acm.org/citation.cfm?id=3045336

[219] An Yan and Bill Howe. 2020. Fairness-Aware Demand Prediction for New Mobility. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (April 2020), 1079–1087. https://doi.org/10.1609/aaai.v34i01.5458 Number: 01.

[220] Tao Yang and Qingyao Ai. 2021. Maximizing Marginal Fairness for Dynamic Learning to Rank. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 137–145. https://doi.org/10.1145/3442381.3449901

[221] Elad Yom-Tov. 2019. Demographic differences in search engine use with implications for cohort selection | SpringerLink. *Springer Netherlands* (2019), 1–11. https://doi.org/10.1007/s10791-018-09349-2

[222] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[223] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938 event-place: Singapore, Singapore.

[224] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *International Conference on Machine Learning*. 325–333. http://proceedings.mlr.press/v28/zemel13.html

[225] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making: The Causal Explanation Formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949

[226] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-based Discrimination Discovery: A Causal Inference Approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2718–2724. http://dl.acm.org/citation.cfm?id=3060832.3061001 event-place: New York, New York, USA.

[227] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. (2017). https://arxiv.org/abs/1611.07509

[228] Yongfeng Zhang. 2015. Incorporating Phrase-level Sentiment Analysis on Textual Reviews for Personalized Recommendation. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 435–440. https://doi.org/10.1145/2684822.2697033

[229] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2)*. http://arxiv.org/abs/1804.06876 arXiv: 1804.06876.

[230] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Olivia, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929. https://arxiv.org/abs/1512.04150

[231] Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. 2003. Extracting symbolic rules from trained neural network ensembles. *AI Communications - Artificial Intelligence Advances in China* 16, 1 (Jan. 2003), 3–15. https://dl.acm.org/citation.cfm?id=1218644

[232] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. 12. https://arxiv.org/abs/1702.04595

[233] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv:1511.00148 [cs, stat]* (Oct. 2015). http://arxiv.org/abs/1511.00148 arXiv: 1511.00148.

[234] Julian Zucker and Myraeka d'Leeuwen. 2020. Arbiter: A Domain-Specific Language for Ethical Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 421–425. https://doi.org/10.1145/3375627.3375858

[235] Indrė Žliobaitė and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (June 2016), 183–201. https://doi.org/10.1007/s10506-016-9182-5