CY. center for
algorithmic
transparency

Document Title	**Literature review and bibliographic referencing system**
Project Title and acronym	Cyprus Center for Algorithmic Transparency (CyCAT)
H2020-WIDESPREAD-05-2017-Twinning	Grant Agreement number: 810105 — CyCAT
Deliverable No.	D3.1
Work package No.	WP3
Work package title	Understanding social and cultural consequences of algorithms
Authors (Name and Partner Institution)	Fausto Giunchiglia (UNITN) Jahna Otterbacher (OUC)
Contributors **(Name and Partner Institution)**	Khuyagbaatar Batsuren (UNITN) Veronika Bogin (UH) Alan Hartman (UH) Styliani Kleanthous (OUC) Tsvi Kuflik (UH) Kalia Orphanou (OUC) Avital Shulner Tal (UH)
Reviewers	Frank Hopfgartner (USFD) Michael Rovatsos (UEDIN)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D3.1_Literature_Review_M12
Date	29 September 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105.

Draft Versions - History of Document				
Version	**Date**	**Authors / contributors**	**e-mail address**	**Notes / changes**
v1.0	16/9/19	J. Otterbacher	jahna.otterbacher@ouc.ac.cy	Initial version - presented at consortium meeting
v2.0	25/9/19	J. Otterbacher	jahna.otterbacher@ouc.ac.cy	Pre-final version submitted for review
v3.0	25/9/19	J. Otterbacher	jahna.otterbacher@ouc.ac.cy	Final version

Abstract

Deliverable D3.1 consists of two parts. First, a publicly accessible repository of published scientific articles related to algorithmic system bias and *Fairness, Accountability and Transparency* (FAT) in algorithmic systems, has been released through Zotero, a freely available, open-source bibliographic reference management software. Secondly, we present a comprehensive review of the literature catalogued in the repository to date, which summarizes the state-of-the-art on algorithmic transparency research, focused on research domains related to information access systems.

Keyword(s): Algorithmic bias, bibliographic referencing system, *Fairness, Accountability and Transparency* (FAT), literature review, state-of-the-art

Contents

1. Executive Summary

D3.1 details our understanding of the state-of-the-art in the emerging field of *Fairness, Accountability and Transparency (FAT) in Algorithmic Systems*, based on 12 months of intensive, collaborative work with the existing published literature. The literature review has been conducted with the primary goals of the CyCAT project in mind, including i) raising awareness of algorithmic bias among various stakeholders (end users, system developers, educators, librarians); ii) finding solutions to the problem of social and cultural biases in information access (IA) systems; and iii) creating and sustaining a distributed, interdisciplinary network of researchers across Europe and Israel.

At the early stages of the literature review, it became clear that *algorithmic transparency,* as originally construed in the CyCAT Grant Agreement*,* is a very complex topic, which is being addressed across a number of diverse research communities, through a variety of methods and approaches. Readers should keep in mind that since the CyCAT proposal was submitted (November 2017) until present, the field has evolved a great deal. In previous years, the literature was scattered across many different research communities, with little interaction between them. For instance, even as early as the 1990s, researchers were considering problems of explainability and interpretability in their models (Craven et. al. 1994; Craven et. al. 1996; Domingos 1998). Similarly, in the early 2000s, researchers in different areas of computer science were considering the social and ethical consequences of their algorithms (e.g., the FairML community; researchers of discrimination discovery in the data mining community) (Pedreschi et. al. 2009; Buckley et. al. 2007; Cho and Roy 2004).

However, in 2019, the field looks very different. For instance, the Association for Computing Machinery (ACM) Fairness, Accountability and Transparency (FAT*) community[1] is an effort to take a more holistic approach to addressing the consequences brought about by extensive use of algorithms and algorithmic systems. Specifically, the community brings together researchers across disciplines - not only from the computer and information sciences, but also from disciplines including the social sciences and law - into an emerging community, which is specifically focused on fairness, accountability, transparency and other ethical issues in socio-technical systems. The effort stems from a growing recognition that algorithmic systems are not merely technical, but rather, are *socio-technical* in nature. Human decisions and biases are present at every step of the development pipeline, not to mention during the interaction with the user. Thus, when analyzing algorithmic bias in complex, networked information access systems, it is necessary to adopt an approach that emphasizes the social dimension(s) of the problem, as well as treating the technical considerations.

We have focused our efforts not only on collecting the most relevant articles that have been published in high-impact computer and information science venues (international, peer-reviewed

[1] It should be noted that community organizers have indicated that a new name and acronym will be announced in early 2020.

conferences and journals) but also to understand the nature of this emerging field. To this end, as will be detailed, our collection efforts focused on five domains of research related to algorithmic systems in general, and in information access systems in particular, as described in the CyCAT Grant Agreement. The five domains considered are: i) Machine Learning (ML); ii) Information Retrieval (IR), iii) Recommender Systems (RecSys), iv) Human-Computer Interaction (HCI), and v) other domains.

As of September 2019 (M12), our repository consists of over 245 articles. However, it should be considered a "living deliverable" and the collection will grow as the project progresses.

2. Repository of articles and bibliographic referencing system

We have created a publicly accessible library, CyCAT Survey Collection[2], within the online, freely available bibliographic referencing system, Zotero.org.

Zotero is an open-source software that can be used either online or downloaded as an application to a computer. Zotero gives the user the opportunity to organize, cite and collect bibliographic references. In Zotero users can save films, web pages, sound recordings, artworks, etc. in addition to bibliographic references. Through the browser, Zotero stores the bibliography in the user's library along with all the metadata such as author name, abstract, date, publisher, and anything else needed to cite the specific item and attached files to the item. A user can manually insert a paper, a book, a journal, etc. among the metadata and Zotero can find the pdf file if it is available online. Another important point of Zotero is that users can save their references with as many tags as they need in order to make it easier to categorize and search for an item in their library. They can also create collections, to save items under the same topic.

In addition to the user's individual library, Zotero has the functionality to create different groups and invite other people of common interests to join and share a library. The creator of each library can choose whether the group will be open to everyone, or it will be by invitation only. The owner of each group is responsible to choose the role of each user.

Our group, named *CyCAT Survey Collection*, is publicly accessible but has closed membership. That means that anyone can view the CyCAT library and benefit from the collection of bibliographic references, but only members of CyCAT consortium can make changes (e.g., add/delete/modify articles) to the library and view any attached file. Zotero registered members can have access to the CyCAT library through the following URL: https://www.zotero.org/groups/2344383/cycat_survey_collection. As explained above, they can view the articles along with their metadata but access to the full article can only be granted through the publisher.

The library currently consists of 245+ bibliographic references. All articles have been categorized into five research domains, which appear as subdirectories of the group library:

[2] https://www.zotero.org/groups/2344383/cycat_survey_collection?

- HCI - Human-Computer Interaction
- IR - Information Retrieval
- Machine_Learning - Machine Learning
- Rec_Sys - Recommender Systems
- Other

These domains represent the expertise of the CyCAT consortium and are the basis for the survey collection we developed. As will be explained in detail in Section 3, each partner in CyCAT was responsible for identifying the bibliography related to their expertise and uploading it to the library under one of the above domains. Then each bibliographic was characterized by a number of tags, describing the content of the paper. The tags represent problems and solutions as detailed in Section 5.

To describe the problem space,[3] the Zotero collection uses six tags: I - Input, D - Data, O - Output, M - Model, T - Third Party, and F - Fairness. As will be detailed in Section 4, these tags correspond to the components of an algorithmic system under study in the particular article (see Figure 1 in Section 4). With those tags, a user can understand the problem that a specific paper discusses before opening it. To describe the solutions proposed, the collection uses the following tags: Auditability, Discrimination Discovery (explicit), Discrimination Discovery (implicit), Explainability Management, Explainability (black box), Explainability (white box), Fairness Management, Fairness Learning, Fairness Certification, Fairness Sampling and Other. As explained above for the problems of interest, again the solutions tags can help the user to easily find what a paper discusses and what is the solution that gives. For example, the journal article "*A causal framework for discovering and removing direct and indirect discrimination*", authored by Zhang Lu, Wu Yongkai and Wu Xintao, is in the Machine Learning category, and is annotated with the problem tag *Data* and solution tags *Discrimination discovery indirect* and *Discrimination discovery direct*. A full presentation of the tags used in the Zotero collection is provided in the Annex.

3. Methodology

<u>Goals</u>: The literature review primarily serves <u>four goals</u>, namely to:

1. Characterize the *problem and solution spaces* in the emerging field of Fairness, Accountability and Transparency in algorithmic systems.
2. Understand the *diversity dimensions* of interest to researchers in the field.[4]
3. Gauge the extent to which various research communities are contributing to / shaping the FAT research.
4. Map the problem space to the solution space, across the research venues/communities.

[3] The motivation and explanation of the conceptual framework is provided in D3.3.

[4] As will be described, diversity dimensions are the aspects upon which the system's behaviours may differ, in ways that may be considered problematic by system observers and users.

Scope: CyCAT Consortium members are experts in various domains within the computer and information sciences. Therefore, it was decided at the outset, to scope our literature review as such. While there is a growing literature on algorithmic biases and FAT issues emerging across disciplines (e.g., within the fields of law, business, philosophy/ethics, and even medicine) we have focused our review within the computer and information sciences. Per the third goal stated above, it was necessary to define the scope of the research communities considered in our review. Since "research communities" themselves are difficult to precisely define, we decided to target high-impact international publication venues (both conferences and journals) across a number of areas related to "intelligent systems in general, and in information access systems in particular" (CyCAT Grant Agreement, p. 21).

Process: We followed a methodology involving both bottom-up and top-down processes for collecting and reviewing articles related to FAT. The methodology is an evolution from the standard facet-based methodology used in information science to carry out book (and even product) classification (Hjørland, 2002).

Bottom-up: At first, a temporary repository was created on the CyCAT project Google Drive, where members could record relevant literature that they had found, through a bottom-up, open search process. Thus, an initial body of material was first examined.

Definition of properties: In February 2019 (M5), a scientific exchange took place in Trento, and members from the teams intensely involved in this processes attended (OUC as coordinator, UNITN as WP3 leader, UH as WP4 leader). The ultimate goal of the meeting was to identify a set of properties by which we could characterize the content of the articles collected. Thus, a concept for understanding social and cultural biases in algorithmic systems was developed. As will be detailed in Section 4, the concept articulates a "diversity lens" for studying the complex problem of bias in algorithmic systems.[5] In addition, it provides a means to characterize each article collected for the literature review, by analyzing the problem(s) presented by the paper as well as the solution(s) proposed or developed.

Top-down: Following the development of the guiding concept, as well as the classification scheme (the problem and solution spaces) a top-down approach was implemented. At M8 (April 2019), an inventory of the article repository was taken, to understand which domains / disciplines (i.e., research communities) had produced a critical mass of publications related to FAT in algorithmic systems (i.e., both problems in systems and their solutions). Through this exercise, a list of key, high-impact publications venues was created for each domain. The domains were divided up by each CyCAT team, according to each team's expertise, as presented in Table 1. Teams were to review each publication venue's proceedings / published volumes during the last 10 years (2008 - 2019), resulting in a high-recall search for relevant published articles. The key words used included: "accountability," "bias," "discrimination," "fairness," "explainable," and "transparency."

[5] D3.3 motivates and develops in detail the "diversity lens" that we will be using in CyCAT, in order to study social and cultural biases in algorithmic systems.

It must be noted that the list of publications in Table 1 is not exhaustive; further publication venues may be added to our repository in the future. However, the problem and solution spaces discovered detailed in Figures 1 and 2 have proven to be robust across the 245+ articles reviewed.

	Key publication venues	Team responsible
Machine Learning	AAAI IJCAI KDD SIGKDD CIDM AIES NIPS MLSP ACM Data Mining and Knowledge Discovery Journal	UNITN
Information Retrieval	AAAI ICWSM ACM CIKM ACM SIGIR ACM WWW TOIS JASIST IR Journal	OUC
Recommender Systems (includes online advertising, freelance marketplaces, shopping, etc.)	AAAI ICWSM CIKM ACM WWW CHI CSCW ACM RecSys ArXive ACM FAT* UMUAI ACM SIGIR	UH
Human-Computer Interaction	ACM CSCW ACM CHI CSCW Journal ACM HCI Journal INTERACT Journal of Behaviour and Information Technology Journal of Big Data and Society	OUC
Other	AAAI HCOMP ACM FAT*	All

Table 1: Domains and publication communities examined in the literature review.

Article analysis: For each article entered into the repository, the bibliographic citation as well as the respective research domain was recorded. After reviewing the article, three additional properties, which shall be explained in detail in Section 4, were also recorded:

- The problem(s) identified
- The solution(s) proposed to address the problem(s)
- The diversity dimension(s) examined in the work

4. Properties examined in the literature review

Here, we describe the three key dimensions that we analyzed, when reading and cataloging each article in our collection: i) the problem posed by the authors, ii) the solution(s), if any, that the authors propose in order to address the particular problem, and iii) the diversity dimension(s) of interest in the study.

4.1 Problem under study

We first characterized the macro components of the algorithmic system, which are cited by the author(s) as being the source of the problem. Figure 1 provides a general characterization of algorithmic systems and their macro components, which we have used to examine the *problem space* of algorithmic system bias.

As shown, a basic system architecture can be described as follows. First, the system receives input (I) for a particular instance of its operation. It operational component (i.e., algorithmic model - M) performs computation based on these inputs, producing an output (O). The algorithmic model (M) learns from a set of observations of data (D) from the problem domain. It may optionally receive constraints from one or more third party actors (T) and/or a set of fairness criteria (F), which may modify the operation of the algorithmic model.

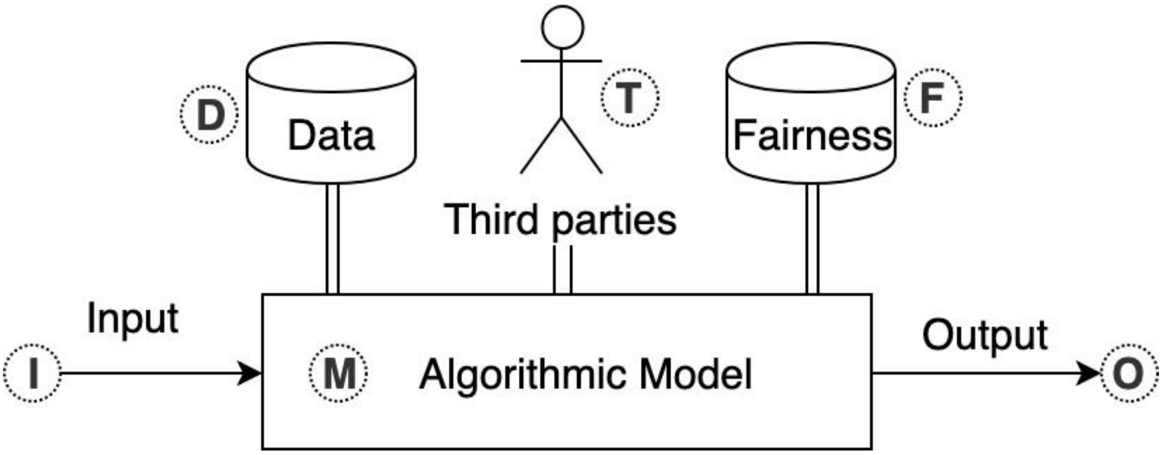

Figure 1: The problem space of algorithmic system bias.

It should be noted that while some studies of FAT in algorithmic systems may examine one and only specific component of that system, other studies address problems that involve more than one macro component of the respective system. In our repository, each article is associated with one or more tags, which indicate the problem(s) examined by the authors in that particular work.

4.2 Solution(s) proposed

Finally, across the articles reviewed, we identified four classes of solutions proposed for promoting Fairness, Accountability, and Transparency in algorithmic systems. These are illustrated in Figure 2, along with specific solutions falling into each class, and are briefly summarized in Table 2.

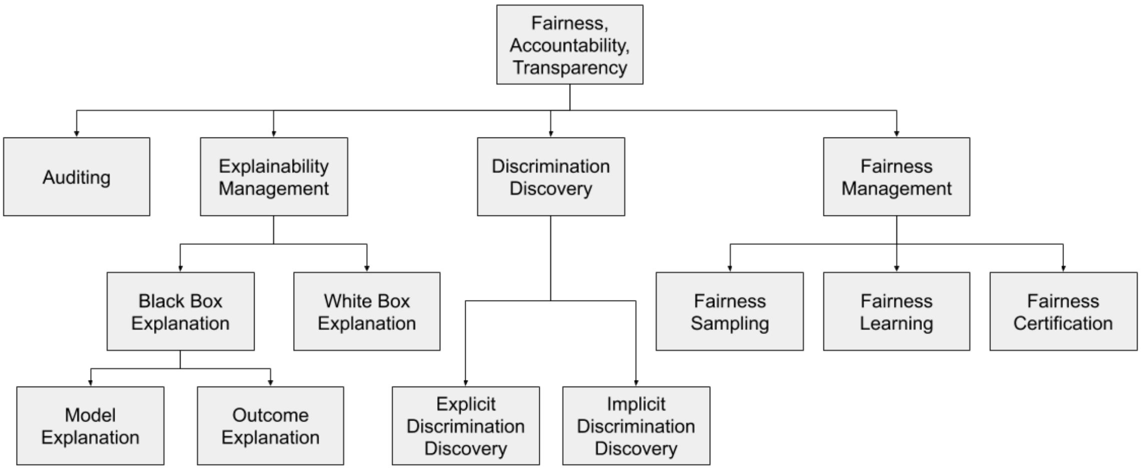

Figure 2: The solution space - tools for promoting Fairness, Accountability and Transparency in algorithmic systems.

Auditing	Auditing involves the systematic examination of a system's behaviours, by someone other than the system developer. In other words, audits are performed by outside observers who do not have access to a system's inter-workings. Common approaches include the within-system and cross-system audits (Sandvig et al., 2014). Within-system audits consider the changes in an algorithm's output, as a function of the changes in a controlled set of inputs. In contrast, cross-system audits make comparisons between the behaviours of different systems, which serve similar purposes / have similar functions.
Explainability Management	"Explainability" approaches, in contrast to audits, emphasize

	the role of the user, and his or her need to understand the system's behaviours. The key challenge is that in many problems, there is a tradeoff between an algorithmic model's accuracy and its interpretability / transparency from the user perspective. The two approaches summarized below address different scenarios: i) white-box, where interpretability is prioritized; ii) black-box, where complex and opaque models are used in order to achieve greater accuracy / predictive power. In both cases, the researcher is looking to provide some degree of *explanation* to the user, concerning the model's behaviours.
-White-box Explanation	In settings where interpretability is important, a more transparent "white-box" modeling approach can be used. Examples include decision trees and regression models. A white-box modeling approach is also used to explain the prediction/classification outcome of a black-box model e.g. extracting rules from decision trees or a causal model.
-Black-box Explanation	These approaches aim to extract a degree of interpretability from complex, opaque models. For instance, some researchers aim to develop a parallel model that accurately mimics that behaviours of the black-box model (i.e., has high fidelity).
--Model Explanation	Some solutions aim to provide global explanations, or provide insights as to the model's overall behaviours.
--Outcome Explanation	In contrast, other approaches focus on providing local explanations, which help a user to understand why the model results in a particular outcome / decision.
Discrimination Discovery	Discrimination discovery approaches originated within the data mining community. It aims to detect discrimination (i.e., disparate impact) either in historical datasets or in automated decisions (most typically in classification/prediction tasks), against individuals and/or social groups.
-Explicit Discrimination	Also called "direct discrimination," such cases involve a rule or procedure that results in disporportionate burden(s) on a particular group of persons.
-Implicit Discrimination	This type of discrimination also imposes a disporportionate burden on a minority group, however, the rules/procedures involved do not explicitly use the sensitive diversity dimension(s).
Fairness Management	The final class of solutions focus on the ethical concern of treating people and social groups in a fair manner, in the

	context of machine learning and algorithmic systems.
-Fairness Sampling	Often, the source of algorithmic bias has to do with the composition of the training data used to learn a model. Solutions focused on fairness sampling aim to ensure that training data sets are balanced in a manner that promotes fairness.
-Fairness Learning	In contrast, fairness learning solutions consider the role of the learning process in promoting fairness. In other words, solutions in this class typically impose constraints that force the learner to result in fairer models (i.e., in which disparate impact on individuals/groups it mitigated).
-Fairness Certification	Fairness certification solutions aim to test algorithmic models for possible disparate impact, "certifying" those that do not exhibit evidence of unfairness.

Table 2: Summary of the four main classes of solutions for promoting Fairness, Accountability and Transparency in algorithmic systems.

4.3 Diversity dimensions

Finally, having characterized the problem and solution described by a given article, we identified the *diversity dimension(s)*, upon which a problematic system behaviour manifested, as reported by the authors. While it is natural to be most concerned with the "problem" and "solution(s)" examined when reviewing a research article - particularly one that is more technical in nature - we have specifically chosen to consider the relevant diversity dimension as a "first-class citizen" in our literature review. While it is true that early, technical works often treated "sensitive attributes" in a more generic sense (e.g., Pedreschi et al., 2019), more recent work in socio-technical systems addresses particular diversity dimensions including social, cultural, political and information attributes, as will be discussed. Recording these attributes in our review will allow us to eventually consider not only which dimensions are most problematic / frequently studies, but also, how the solutions proposed might differ depending on the particular diversity dimensions being studied in a given system.

Table 3 provides examples of the diversity dimensions discussed in the articles collected. As will be shown, the system under study in each article, can exhibit different behaviours as a function of the diversity dimension, which may or may not be problematic for a given user or observer. It can be noted that while many of the diversity dimensions concern social and cultural attributes, we also observe dimensions such as the quality / accuracy / credibility of the information provided to the user. Even though such instances may not represent cases where an algorithmic system's behaviour 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).

Diversity dimension	**Citation from the repository**	**Explanation / example**
age	Díaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018, April). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 ACM CHI Conference on Human Factors in Computing Systems* (paper 412).	The authors studied word embeddings and sentiment analysis algorithms, and the tendency to perpetuate ageism. They found systematic associations between terms related to older age (e.g., “old”, “elderly”) and negative sentiment, as compared to terms related to younger age (e.g., “youth,” “young”).
gender	Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).	The researchers studied word embeddings that were trained on Google News articles, demonstrating their tendency to perpetuate gender biases (e.g., associating man/woman to computer programmer / homemaker).
information	Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. Information processing & management, 41(5), 1193-1205.	The authors proposed a measure of search engine bias. Using a large set of user-generated search queries, they created a “fair results set,” which was the union of the results retrieved across a number of alternative engines. They then measured the deviation between any given engine’s results and the fair set. Thus, *information diversity* was the dimension of interest.
language / linguistic	Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media.*	The system of interest is a hate speech detector (classifier). Linguistic diversity is the dimension of interest here, as words taken as offensive by some users

		are often (incorrectly) flagged as being hate speech.
minority status	Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73). ACM.	This work addresses the social bias of text classification algorithms for identifying toxic language online. It is shown that terms related to minority status (e.g., “Muslim,” “gay”) can inadvertently become associated with the toxic language label.
national origin	Thelwall, M., & Maflahi, N. (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), 1124-1135.	Problematic user behaviour in an algorithmic bibliographic referencing system was studied, citing a bias towards reading articles produced in the user’s home country.
opinion	Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018, June). Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 165-174). ACM.	This article concerns ranking algorithms for personalized item recommendations, based on user reviews. Authors argue in favor of incorporating a model of the user’s opinion of the target item, as well as of the opinion expressed in reviews.
physical attractiveness	Matsangidou, M., & Otterbacher, J. (2019, September). What Is Beautiful Continues to Be Good. In: Lamas D., Loizides F., Nacke L., Petrie H., Winckler M., Zaphiris P. (eds) Human-Computer Interaction – INTERACT 2019. INTERACT 2019. Lecture Notes in Computer Science, vol 11749. Springer, Cham	In a study of image tagging algorithms’ descriptions of input images of people, it was found that people rated as being more attractive (by human raters) were also systematically more likely to be associated with tags having positive sentiment (e.g., “friendly,” “intelligent”).
political leaning / affiliation	Hu, D., Jiang, S., E Robertson, R., & Wilson, C. (2019, May). Auditing the partisanship of Google search snippets. In *The World Wide Web Conference* (pp. 693-704). ACM.	The authors collected the search engine results pages for queries having left and right political leanings. They compared the Google snippets

		shown to users, to the full pages collected. For both left and right queries, it was found that snippets tend to highlight the most politically extreme text on the page, amplifying partisanship.
sensitive / protected attribute	Weber, I., & Castillo, C. (2010, July). The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 523-530). ACM.	Through an analysis of Web search logs, the authors studied user search behaviors, noting systematic differences based on users' sensitive attributes (e.g., income).
race	Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018, April). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference* (pp. 903-912).	Using the COMPAS system as a case study, the authors surveyed users on their perceptions of this criminal risk prediction system. One of the key questions was whether users consider it fair to use features such as a defendant's race in a decision making scenario.

Table 3: Diversity dimensions appearing in the literature survey on algorithmic system bias and transparency, with an example citation and explanation.

5. Literature review

This section provides a review of the articles in our Zotero repository, for each of the five categories of research upon which we focused our efforts. Within each research domain, we aim to characterize the problems identified and the solution(s) proposed for the identified problems. We also discuss the diversity dimensions of each area of the research. The review for each domain is structured as follows: first, we summarize the key problems and solutions examined by researchers, organized by the algorithmic component most relevant to the problem/solution (see Figure 1); secondly, we discuss the range of diversity dimensions that are being addressed by the research in each domain.

5.1 Machine Learning

Although the focus of CyCAT is on *algorithmic systems for information access*, we begin with a review of the FAT literature in Machine Learning (ML) for two key reasons. First, modern information access systems use applied ML techniques extensively, for a range of tasks, from data cleaning and augmentation, to training ranking mechanisms, to inferring user models for personalization. Secondly, the ML community has been studying problems related to FAT -- although it was not called that at the time -- for several decades now.

5.1.2 Problems examined / solutions proposed in ML

The articles included in our survey of the machine learning literature concern three problems: i) the *discrimination discovery* / prevention in the classification problems, ii) the *fairness computation* in decision making systems, and iii) promoting the *interpretability* of the model and/or outcome.

5.1.2.1 Discrimination Discovery/Prevention Problem/Fair ML

Although originally, the research focusing on discrimination discovery and fairness in ML emerged from distinct communities of researchers, it is important to recognize the inherent link between these lines of research. In particular, the approaches used to solve the discrimination discovery / prevention problem, as well as the fairness computation problems, are both required in order to develop fairness-aware ML algorithms. Therefore, in our review of the ML literature, discrimination discovery and fairness approaches are reviewed together.

These approaches are divided into: pre-processing (i.e., data-focused), in-processing (i.e., model-focused) and post-processing methods (i.e., output-focused). The pre-processing methods modify the input datasets so that the outcome of the algorithm applied to the data will be fair. The in-processing methods are applied during the learning phase of the model and their goal is to modify an existing algorithm or create a new one that will be fair applied to any input. The post-processing techniques modify the output of the model to be fair.

Data (Pre-processing Methods)

Many of the articles that concern the discrimination discovery/prevention and fairness problems use pre-processing methods to remove the discrimination bias of the input training data. The proposed solutions include implicit and explicit discrimination discovery, fairness sampling and auditing (when performed by an outside party).

A frequently used technique for fairness sampling of the data is to generate a new dataset using a causal Bayesian network. For instance, Zhang et. al. (2016, 2017) discover and prevent discrimination bias in decision support systems using a causal Bayesian network to identify pair of tuples with similar characteristics from the dataset. By learning the BN structure, the authors identify the causal factors for discrimination. Also, they remove any discrimination bias from the dataset by generating a new dataset. Cardoso et. al. (2019) also use a Bayesian network estimated from real-world data to generate biased data that are learned from real-world data and fairness metrics such as disparate impact and disparate mistreatment to assess discrimination. Also, Johndrow (2019) identify fairness constraints in the training datasets of machine learning algorithms and apply them into the training data in order to remove discrimination bias. However, their approach can be applied in cases where there is only one protected variable. Kilbertus et. al.

(2018) provide fairness training and certification in machine learning using an encrypted version of sensitive data, privacy constraints and decision verification using secure multi-party computation (MPC) methods.

As an alternative to fairness sampling, Romei et. al. (2013) use auditing as an approach to discover discrimination where auditors (testers) search through the dataset. Also, they propose situational and corresponding testing approaches which are special cases for auditing. Cardoso et. al. (2019) propose the use of black-box auditing to repair the dataset by changing attribute labels. Similarly, Pedreshi et. al (2009) use a black-box 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. In addition, Kuhlman et. al. (2019) identify fairness specifically in ranking algorithms used for decision making. The authors use an auditing methodology FARE (Fair Auditing based on Rank Error) for error-based fairness assessment of ranking. They proposed three error-based fairness criteria which are rank-appropriate.

Another approach in terms of a data-focused solution is the implicit and explicit discrimination discovery. For instance, Rudinger et. al. (2017) discover discrimination bias in natural language processing (NLP) data by searching for overgeneralization at the level of word co-occurrences considering the sensitive attributes e.g., age, genre, and ethnicity, using an association metric, the pointwise mutual information.

Some other works use pre-processing methods as a solution for discrimination discovery applied to specific diversity domains. For instance, Datta et. al (2015) analyse the gender discrimination in online advertising (Google ads). They use machine learning 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. (2018) detect gender bias in NLP data by identifying linguistic features that are gender-discriminative. Zhao et. al. (2018) 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. (2018) detect gender discrimination in movies using knowledge graph and word embedding for bias detection and removal after analysing the data (i.e., mentions of each gender in movies, emotions of the actors during the movies, occupation of each gender in the movies, screen time.)

These approaches are summarized below in Table 4, in terms of approach.

Problem	Implicit and Explicit Discrimination	Auditing	Fairness Sampling
Zhang et. al. (2016) aim to discover and remove discrimination bias in decision making systems.			Their approach is to search for pairs of tuples from the dataset with similar characteristics using a Causal Bayesian network and the associated causal inference as a guideline.
Cardoso et. al. (2019) detect and remove discrimination in machine learning models concerning ethical and legal implications.		One approach suggested by the authors is the black-box auditing. A pre-processing process to repair the dataset by changing attribute labels.	They learn the structure of a Bayesian network (BN) automatically from real-world data. Data were sampled from the estimated BN. Fairness metrics used to assess discrimination include disparate impact and disparate mistreatment.
Zhang et. al. (2017) identify and prevent discrimination in decision support systems.			They build a causal model to identify the causal factors for discrimination and then remove any discrimination bias from the dataset by generating a new dataset

Rudinger et. al. (2017) discover overgeneralization in natural language processing data which leads to bias among individuals.	They use an association metric, pointwise mutual information to search for over-generalisation in bias at the level of word co-occurrences considering the sensitive attributes e.g. age, genre, and ethnicity. They found out that the dataset represent gender, racial, religious and age-based stereotypes.		
In the survey paper of Romei et. al. (2013), the authors describe different approaches for discrimination discovery in data mining.		The authors use auditing as an approach to discover discrimination where auditors (testers) search through the dataset. Also, they propose situational and corresponding testing approaches which are special cases for auditing.	
Kuhlman et. al. (2019) identify fairness in ranking algorithms used for decision making.		The authors use an auditing methodology FARE (Fair auditing based on rank error) for error-based fairness assessment of ranking. They proposed three error-based fairness criteria which are rank-appropriate.	
Kilbertus et. al. (2018) provide fairness training and certification in machine learning.			The authors provide an encrypted version of sensitive data, privacy constraints and decision verification using secure multi-party

			computation (MPC) methods.
Leavy et. al. (2018) detect gender bias in machine learning models	Identify linguistic features that are gender-discriminative through text training data.		
Zhao et. al. (2018) detect gender bias in a 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 genre bias demonstrated in the data.		
Datta et. al (2015) analyse the gender discrimination in online advertising (Google ads)	They use machine learning 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.		
Madaan et al. (2018) detect gender discrimination in movies			Use knowledge graph and word embedding for bias detection and removal after analysing the data (i.e. mentions of each gender in movies, emotions of the actors during the movies, occupation of each gender in the movies, screen time..)

Johndrow (2019) identify fairness constraints in the training datasets of machine learning algorithms.			Their approach works only for one protected variable. The authors construct a new dataset by removing the ‘race’ variable to achieve data privacy and anonymization.
Heindorf et. al. (2019) measure and reduce bias against edits by anonymous and newly registered editors in wikidata.	Their approach is to omit user-related features and to develop features that purely encode the content of an edit, rather than any meta information.		
Pedreschi et. al. (2009) discover patterns of direct and systematic discrimination	They use a black-box 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		
Feldman et. al. (2015) measure computational fairness and link it to the legal notion of disparate impact.	Their pre-processing approach modifies each attribute (but not the training labels) in the dataset so that the marginal distributions based on the subsets of that attribute with a given sensitive value are all equal. They use a disparate impact remover		

Table 4: Pre-processing (data-focused) methods for discrimination discovery / prevention and fairness

Model Training (In-processing Methods)

The in-processing methods proposed consider the problem of discrimination discovery and fairness in the algorithm itself. Therefore, the methods modify the classification/predictive algorithm mainly by introducing some fairness constraints (Zhang et. al. (2018), Celis et. al

(2019), Kleinberg et. al (2016), Dimitrakakis et. al. 2018) or by introducing new fairness metrics such as FACE and FACT (Khademi et. al (2019)), feature-apriori fairness, feature accuracy fairness and feature-disparity fairness (Grgic-Hlaca et. al. (2018)). In addition, Kamishima et. al. (2012) propose a regularization approach by introducing a fairness-focused regularization term and apply it to a logistic regression classifier. Kusner et. al. (2017) measure counterfactual fairness on decision support systems. They provide optimization of fairness and prediction accuracy of the classifier using a causal model. Speicher et al. (2018) propose the Aequitas auditing tool which tests models for several bias and fairness metrics. These approaches are summarized below in Table 5, in terms of approach.

Problem	**Fairness constraints**	**Optimization of Fairness metrics**	**Other approaches**
Khademi et. al (2019) measure fairness in decision making systems.		The authors proposed two fairness metrics (FACE and FACT) and with the use of causal models. They analyse the cause-effect relationships to detect and quantify discrimination on sensitive attributes.	
Zhang et. al. (2018) detect discrimination in decision making systems.	They use a causal explanation formula to evaluate fairness and explain the total observed disparity of decisions through different discriminatory mechanisms. They use fairness constraints and counterfactual measurements for causal explanations of the discrimination.		

Kamishima et. al. (2012) proposes three causes of unfairness: prejudice, underestimation and negative legacy.			They propose a regularization approach by introducing a fairness-focused regularization term and apply it to a logistic regression classifier.
Kusner et. al. (2017) measure counterfactual fairness on decision support systems.	Optimization of fairness and prediction accuracy of the classifier using a causal model. The authors propose algorithms to take into account the different social biases that may arise towards an individual based on ethically sensitive attributes and compensate for these biases effectively rather than removing the attributes		
Celis et al. (2019) measure individual/preference/procedural fairness in classification algorithms.	They propose a meta-algorithm for classification with (nonconvex) linear-fractional constraints. Linear fractional constraints capture many existing fairness definitions in the literature.		

Grgic-Hlaca et. al. (2018) measure fairness in the input features of the algorithm conditioned on its impact on outcomes.		They proposed three measures of process fairness: feature-apriori fairness, feature accuracy fairness and feature-disparity fairness. They also focus on human judgements to quantify process fairness of each of the individual features	
Speicher et al. (2018) measure bias and fairness in algorithmic decision making.			They propose the Aequitas auditing tool which tests models for several bias and fairness metrics.
In the survey paper of Romei et. al. (2013), the authors describe different approaches for discrimination discovery in data mining.	The in-processing techniques reviewed in this survey paper is to modify the classification algorithm by integrating with anti-discrimination criteria. Some methods train a separate model for each protected group. For decision trees, the entropy-based splitting criterion in decision tree induction to take account attributes denoting protected groups.		

Kleinberg et. al (2016) detect fairness in the risk assessment of group of individuals in any application domain (e.g. google ads, considering the genre)	The authors propose three fairness constraints that any algorithm should take into consideration while assessing the risk for individuals divided into multiple groups (e.g. females, males).		
Dimitrakakis et al. (2018) consider the problem of fairness in decision making when the underlying probabilistic model of the world is uncertain.			The authors deploy a Bayesian fairness-aware algorithm to explicitly incorporate parameter uncertainty and fairness constraints to decision making problems

Table 5: In-processing (model-focused) methods for discrimination discovery / prevention.

Output (Post-processing Methods)

The post-processing methods concern the modification of the output of the classifier. As discussed in the survey paper of Romei et. al. (2013), examples of post-processing techniques include the re-labelling of the predicted class or altering the confidence of classification rule. Furthermore, in Hardt et. al. (2016), the authors propose a framework to construct classifiers from any Bayes optimal regressor following a post-processing step which avoids to modify the training process. They discover discrimination against a specified sensitive attribute in supervised learning. Zhang et Wu (2017) proposed an indirect discrimination approach using a causal model where they detect discrimination in the prediction/classification outcome by computing the classification error rate (error bias).

5.1.2.2 Promoting the Interpretability / Explainability of the Model or Outcome

The third approach used within the ML literature for promoting FAT in algorithmic systems, is that of promoting the explainability / interpretability of the learned models. These problems fully concern the model modification, either with in-processing or post-processing techniques. Below, we outline a summary of the approaches proposed in the reviewed papers to solve the particular problem.

Model Explainability

Most of the reviewed papers concern the interpretability of black-box models such as neural networks either by measuring the feature importance or with the support of a white-box model (white-box explainability). Some of the methods reviewed in the survey paper of Guidotti et. al. (2018), they use global and local interpretability metrics such as the model complexity, accuracy and fidelity. Other proposed methods are sensitivity analysis, feature importance and salience mask for images. Ribeiro et. al. (2018) improve the interpretability of black-box models using anchors whereas Tan et. al. (2017) propose a model distillation to train a transparent student model to mimic the black-box model and then comparing the transparent mimic model to a transparent model trained using the same features on true outcomes instead of the labels predicted by the black-box model.

Lu et. al. (2005) propose the Neurorule framework which adds classification rules using a GP to a neural network. Similarly, Zhou et. al (2003) also use classification rules to improve the interpretability of a black-box model. They propose the REFNE framework that extracts symbolic rules from trained neural network ensembles. Another approach is to extract decision trees from trained neural networks (Boz 2002, Craven et. al. (1996)).

Regarding the explainability of white-box models, there are only two works in the reviewed papers. Schetinin et. al. (2017) find an interpretable classification model for medical domains. They propose a Bayesian averaging over ensemble of decision trees classifier. A selection procedure was proposed for extracting confident decision trees from the Bayesian decision tree ensemble. In addition, Cowgill and Tucker (2017) propose a counterfactual evaluation method where causal inference models are used for quantifying changes in bias from a new algorithm.

Moreover, there are some methods proposed in the papers that can be applied to any classifier (either black-box or white-box model) concerning the feature importance. For instance, Henelius et. al. (2014) search for a group of attributes whose 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. (2016) propose the measure of feature importance (MFI) which can be applied both to white-box and black-box models.

The relevant papers from our repository are summarized in Table 6, in which we distinguish between white- and black-box explainability, and also detail the importance of particular features.

Problem	**White-box Explainability**	**Black-box Explainability**	**Feature Importance**
Henelius et. al. (2014) search for a group of attributes whose interactions affect the			The authors use a fidelity metric to measure the importance of each group of attributes but they also take

predictive performance of a given classifier.			classification accuracy into consideration.
Schetinin et. al. (2017) find an interpretable classification model for medical domains.	They propose a Bayesian averaging over an ensemble of decision trees classifier. A selection procedure was proposed for extracting confident decision trees from the Bayesian decision tree ensemble.		
Cowgill and Tucker (2017) measure transparency and interpretability of any classification algorithm.	They propose a counterfactual evaluation method where causal inference models are used for quantifying changes in bias from a new algorithm.		
In the survey paper of Guidotti et. al. (2018), the authors discuss different techniques for measuring the interpretability in black-box models.		Some of the methods reviewed in the survey paper, involve the use of global and local interpretability metrics such as the model complexity, accuracy and fidelity.	Other proposed methods are sensitivity analysis, feature importance and salience mask for images.
Ribeiro et. al. (2018) improve the interpretability of black-box models using anchors.		The anchors enable users to predict how a model would behave on unseen instances with much less effort and higher precision as compared to existing techniques for model-agnostic explanation or no explanations.	
Tan et. al. (2017) measure transparency in black-box models.		The authors propose a model distillation to train a transparent student model to mimic the black-box model and then comparing the transparent mimic model to a transparent model trained using the same features on true outcomes instead of the labels predicted by the	

		black-box model.	
Lu et. al. (2005) improve interpretability of neural networks using classification rules.		They propose the Neurorule framework which adds classification rules using a GP to a neural network. In Neurorule, the rule extraction process is three-hold and generate the relations between inputs and outputs to get ultimate production rules.	
Boz (2002) extract decision trees from trained neural networks in order to improve the interpretability of any neural network.		They extract decision trees from any neural network and prunes the tree in order to maximize fidelity between the tree and the neural network using a fidelity pruning algorithm.	
Zhou et. al. (2003) improve the comprehensibility of any neural network ensemble classifier.		They propose the REFNE framework that extracts symbolic rules from trained neural network ensembles. It utilizes ensembles to generate a number of instances and then extract rules from those instances.	
Craven et. al. (1996) improve the comprehensibility of neural networks.		They use the TREPAN algorithm for extracting comprehensible, symbolic representations from trained neural networks. TREPAN queries a given network to induce a decision tree that describes the concept represented by the network. They measure the comprehensibility of the network using the fidelity metric.	

Vidovic et. al. (2016) measure the feature importance in machine learning and deep learning models.	They propose the measure of feature importance (MFI) which can be applied both to white-box and black-box models. The metric can be used for both for a general explanation of the prediction model and for a data instance specific explanation. MFI can detect features that exhibit their importance only through interactions with other features.		

Table 6: Methods used for promoting model explainability / interpretability.

Output (Outcome Explainability)

In contrast to model explainability, some approaches attempt to provide a local interpretation, focusing on explaining a particular outcome generated by the model. A general method for explaining the output of a classifier (either a black-box or white-box model) is by using only the input and output of the model to decompose the changes in the algorithm's prediction outcome into contributions of individual feature values. These contributions correspond to known concepts from coalitional game theory (Strumbelj et. al. (2010)).

More specific methods for black-box explainability were proposed by Krishnan et. al. (1999) and more recently by Card et. al. (2019). Krishan et. al (1999) explain the outcome of a black-box model by extracting decision trees from the data. A genetic algorithm was applied to predict membership queries to the trained neural network and obtain prototypes to control the size of the decision tree. Card et. al. (2019) use transparent explanations for classification decisions as well as an intuitive notion of the credibility of each prediction using a new measure of non-conformity. They also develop a deep weighted averaging classifier replacing softmax in order to provide a transparent version of any successfully developed deep learning architecture.

Articles in our repository that describe such approaches are described in Table 7.

Problem	White-box Explainability	Black-box Explainability	Feature Importance
Strumbelj et. al. (2010) interpret the prediction outcome in the classification models focusing on the importance of each feature.			Use of game theory applied to any classifier. Using only the input and output of a classifier, the authors decompose the changes in its prediction into contributions of individual feature values. These contributions correspond to known concepts from coalitional game theory.
Card et. al. (2019) measure calibration, robustness and interpretability of deep learning models		The authors provide transparent explanations for classification decisions as well as an intuitive notion of the credibility of each prediction using a new measure of non-conformity. They also develop a deep weighted averaging classifier replacing softmax in order to provide a transparent version of any successfully developed deep learning architecture.	
Krishnan et. al. (1999) explain the outcome of a black-box model using decision trees.		Decision trees are generated from the generated input of the trained neural network instead from extracting them directly from data. A genetic algorithm was applied to predict membership queries to the trained neural network and obtain prototypes to control the size of the decision tree.	

Table 7: Methods used for promoting outcome explainability / interpretability.

5.1.3 Diversity dimensions in the ML literature

As previously mentioned, as a research domain ML differs significantly from the others we examined in our review (e.g., information retrieval, recommender systems) as ML researchers' primary focus is typically on methods for learning a model, rather than the development of a model in the context of a particular system to be used by end users (as is the case in IR and RecSys). Therefore, it is often the case in ML that "sensitive attributes" more generally, rather than a particular diversity dimension of a social or cultural nature, is the dimension of interest.

Furthermore, information as a diversity dimension is very common in the ML literature. In particular, the articles that concern the interpretability of the algorithmic classification model focus on *information* biases regarding the learning phase of the classifier or the output of the classifier. Nonetheless, below we present an analysis of the diversity dimensions that appear in our collection of ML articles.

In contrast to the information dimension, algorithmic biases in social and cultural dimensions in the reviewed ML papers are mainly described in articles concerning the problem of discrimination discovery in decision support systems. Most of the reviewed articles in ML, do not specify particular social/cultural diversity dimensions but instead they focus on sensitive attributes more generally. The social/cultural diversity dimensions examined in the collection of ML articles are summarized below.

- Gender (8 articles)

Gender is the most frequently mentioned social/culture diversity dimension in the collection of ML articles. For instance, Madaan et al. (2018) discover algorithmic bias concerning the gender discrimination in movies. In order to detect the gender bias, the authors analyse the differences between the gender in the mentions in movies, in the emotions expressed by the actors in the movie, the occupation of the actors and the screen time for each actor. They remove bias using knowledge graphs and word embedding.

- Race (5)

Johndrow and Lum (2019) employed an algorithm to remove the sensitive information from the training data. They applied their proposed algorithm in a dataset including the criminal histories of individuals with the aim to predict re-arrest. The “race” is the protective attribute in the dataset, thus their algorithm removes racial disparities from predictions without affecting the classification accuracy.

- Nationality (ethnicity) (1):

Rudinger et. al. (2017) measure the bias in natural language processing (NLP) data concerning the attributes of gender, age and ethnicity. Their aim is to demonstrate the existence of stereotypes of various forms in NLP training data.

5.1.4 Summary of ML literature

In summary, we can make the following observations concerning the research trends to date in the ML literature:

- The problems underlined in the ML literature concerned the discrimination discovery and prevention in the decision support systems as well as the transparency/interpretability of classification/predictive models (both white-box and black-box models).

- The solutions for removing bias from decision support systems are divided into three categories: pre-processing methods that modify the data in order to remove bias (e.g. fairness sampling), in-processing methods that modify the learning process of the algorithm and post-processing methods that modify the output of the algorithm.
- In terms of the diversity dimensions of interest, ML research most often concerns the information dimension, as well as a generic notion of a "sensitive attribute."

5.2 Information Retrieval

In total, 56 articles from the target IR publications were archived. The publication venues having the greatest number of articles concerning algorithmic system bias were the journal JASIST (11 articles) and the international conference, ACM WWW (seven articles). Section 5.2.1 details the problems and solutions examined in the IR literature, presented by the relevant system component. Following that, Section 5.2.2 explores the diversity dimensions most relevant to the IR community's recent research.

5.2.1 Problems examined / solutions proposed in IR

Data

Several of the reviewed articles (13) reported that the problem was related - in part or in full - to the characteristics of the system's training data. Three of the above-mentioned four classes of solutions were discussed across these articles - *auditing* (i.e., developing a technique to examine system behaviours), *discrimination discovery* (i.e., the solution proposed in the article focused on how to reveal the bias in the data) and *fairness sampling*, or how to somehow modify the dataset in order to result in a less biased algorithmic model. As described in Table 8, some use more than one approach.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Social/Cultural diversity dimensions			
Koolen and van Cranenburgh (2017) study text categorization for author gender attribution. They note a lack of fairness as well as transparency in such algorithms' decisions, citing dataset bias and interpretation bias as problems.		They compare the linguistic features of texts written authors across genders, in two corpora of Dutch literary works. They explain that confounding factors such as text genre, topic / domain, and target audience affect the words used by authors.	The researchers argue that dataset bias can be overcome by taking into consideration factors such as genre/topic/audience, and balancing the data appropriately in order to control for these effects.
Herdagdelen and Baroni (2011) ultimately aim to enrich a common sense repository (OMCS), which describes everyday actions, with gender		Using a Twitter corpus containing content and profiles, they identified OMCS actions (phrases) in the Twitter corpus, to compute gender	

expectations on actions. In this sense, they exploit implicit gender stereotypes for enriching the OMCS.		correlations. They are able to associate phrases such as “my husband/wife” as being used more often by female/male Twitter users, respectively.	
Dixon and colleagues (2018) consider the use of classification for identifying “toxic” language in Wikipedia talk pages. The problem is that words associated with minority status (e.g., “gay”) tend to be more often classified as being toxic, regardless of the ground truth.		The authors discovered that imbalances in the training data, in which sensitive words were overrepresented in positive examples of toxic text, were the culprit.	By rebalancing the dataset, this form of bias could be mitigated.
Shen et al. (2018) investigated stylistic biases in sentiment analysis algorithms. The problem of interest was the use of markers of African-American English (AAE).		Texts containing markers of AAE were more likely to be scored as conveying negative sentiment, as compared to texts of similar context, without markers of AAE.	The authors “translated” the texts with markers of AAE, before performing the sentiment analysis, in order to mitigate the racial bias of the algorithms.
Callahan and Herring (2011) considered cultural biases at Wikipedia, comparing the descriptions of famous people from Poland and the US at the respective language editions.		Through an analysis of the content and writing style of the articles, the authors detected systematic differences in the descriptions of the same persons. For instance, differences in content (e.g., personal information shared) and tone/sentiment were documented.	
Information as a diversity dimension			
Vincent et al. (2019) aimed to show the importance of user-generated content (UGC) on the Web, in terms of the quality of information that the search engines provide to users.	An audit was performed on Google result pages, across a large set of queries. Six types of important queries (e.g., trending, expensive advertising) were analyzed, to understand how prominent Wikipedia and other UGC was in these important results. Results showed that UGC was returned to users in 80% of results.		
Buckley et al. (2007) considered data bias resulting from the process		Typically, only a subset of documents in a dataset will be judged for	The authors argue in favor of modifications to the traditional pooling process

of pooling data, common in the construction of IR data sets, such as the AQUAINT test collection used in the TREC tasks.		relevance by (human) annotators, for each topic (i.e., query) in the IR task. It is shown that when pools are small relative to the size of the entire set of documents (i.e., database or documents on the Web), pooled sets are typically biased in favor of documents that contain the words in the topic title.	in order to construct unbiased datasets. In particular, samples/pools should include not only top-ranked (for relevance) documents but also those deeper into the systems' rankings.
Urbano (2016) also considers problems of bias in IR test collections. However, his focus is on the metrics commonly used to quantify the accuracy of a test collection of topics/documents.		Simulation experiments were conducted using several ad hoc and statistical measures of test collection reliability. It was found that most ad hoc measures underestimate test set reliability; thus, resulting in excessive investment in created annotated collections. Statistical measures were found to be more accurate, although on small test collections they also tended to underestimate reliability.	
Eickhoff (2018) considers the problem of cognitive biases in relevance judgments. He conducts crowdworker experiments using three public IR document collections.		The experiments test for evidence of four types of cognitive biases: anchoring, bandwagon effect, ambiguity and decoy effect. In particular, it is shown that bandwagon and decoy biases can occur unintentionally in crowdsourced relevance judgements, with consequences for the quality of the data.	
Lin et al. (2015) were interested in the use of crowdsourcing to generate descriptions of images. In particular, they studied the nature of the task - tagging images where an initial description was offered to the worker versus the task without an image description.		Experiments showed that the nature of the data collected differed in the two settings. Tags generated in the "with description" setting tended to be more diverse and specific. However, users were more likely to reuse their generated tags in the "no description" setting.	
Niu et al. (2015) aim to identify the inherent characteristics of human-labelled training		The authors investigate two characteristics of training data (document pair noise, document noise	Guidelines are then suggested for the data labelling strategy, based on which learning

data that make them robust to noise (i.e., not problematic to learning-to-rank algorithms).		ratio) and show how they correlate to three types of learning-to-rank algorithms (pointwise, pairwise, listwise methods).	mechanism one will use to train the ranking model.

Table 8: Data-based problems and solutions in IR articles.

Model

Many (i.e., 30) of the collected articles focus on the algorithmic model as a key (if not the only) problem. Just as in the case of the data-focused articles, the solutions proposed fall into three classes: auditing, discrimination discovery and fairness. However, fairness solutions include both fairness *sampling* and *learning.*

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling / learning**
Social/Cultural diversity dimensions			
Kay et al. (2015), Magno et al. (2016), and Otterbacher et al. (2017) study the perpetuation of gender stereotypes in image search engines.	In all three studies, sets of queries are submitted to search engines and the characteristics of the outputs are studied.	Kay et al. document gender-based biases in Google's portrayal of the professions. Magno et al. study the role of global/local factors in perpetuating stereotypes surrounding physical attractiveness. Otterbacher et al. document gender-based biases in Bing's portrayal of character traits.	
Shandilya et al. (2018) consider extractive text summarization algorithms and the selection of textual units that reference sensitive attributes (political leaning, gender)		The authors show that summarization algorithms often violate notions of fairness; some attributes are underrepresented in the textual units selected for inclusion.	
Kilman-Silver et al. (2015) considers the influence of geolocation on Web search (Google) personalization.	The authors collected Google results for 240 queries over 30 days from 59 different GPS coordinates. They showed that personalization results differ as a function of geographical distance, although this was dependent on the nature of the query.		
Badjatiya et al. (2019) examined hate speech detection algorithms and		Methods were proposed to quantify the extent to which sensitive words	Step two involves a data correction method, in which the amount of

the stereotyping of sensitive attributes.		(e.g., related to race, religion, sexual orientation, etc.) were stereotyped in the models. Step one of their solution is to detect sensitive words.	information available to the classifier is reduced. Essentially, input text is converted to a simpler form.
Diaz et al. (2018) studied age-based bias in text sentiment classification algorithms.		The algorithms were found to systematically associate texts using words related to older age, with negative sentiment, as compared to texts using words related to youth.	The bias was addressed by balancing examples of positive/negative texts making references to young/old age.
Language as diversity dimension			
Rafrafi et al. (2012) addressed bias in sentiment classification algorithms. The problem is that terms occurring frequently in the training data end up with overweighted polarity scores, relative to their actual subjectivities.			A method is proposed that penalizes document frequencies in the training data regularization process. The result is increased model accuracy.
Davidson et al (2017) considers the problem of automated hate speech detection (classification) and the inadvertent influence of offensive words (that are not actually hate speech).			The authors introduce a three-way classification, which separates offensive languages from hate speech and neutral language.
Information as diversity dimension			
Germano et al. (2019) focuses on popularity-based ranking mechanisms for news articles and their biases.		Through simulations and experiments with users, the authors uncover a surprising effect of popularity ranking - a “few-get-richer” effect. Items with a given signal / class (e.g., a particular political leaning) receive more traffic when there are fewer highly-ranked items with the signal. Authors conclude that this “few-get-richer” effect results in a systematic ranking bias - items from a smaller class are better ranked than those of the larger class.	
Bashir and Rauber (2011)		Using the TREC Chemical	

investigates bias quantification in retrieval functions.		Retrieval track, the authors study the relationship between query characteristics and document retrievability. Although the authors note that all retrieval functions have some bias, several characteristics of queries are found to increase/decrease bias of retrieval functions.	
Kulshrestha et al. (2017) presents a method to measure bias in searches for information on Twitter.	The proposed auditing technique considers both the input bias 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.		
A series of articles by Wilkie et al. considered the retrievability biases - where certain document characteristics make them more/likely to be found by the user - of retrieval systems. In Wilkie and Azzopardi (2014a), they examined the issue of fairness vs. performance. Wilkie and Azzopardi (2014b) considers specific measures of retrieval bias and the correlation to system performance. Wilkie and Azzopardi (2017) considers the issue of bias resulting from the process of pooling in the creation of test sets.		Wilkie and Azzopardi (2014a) shows a strong negative correlation between fairness in retrievability and system performance. Wilkie and Azzopardi (2014b), through extensive experiments on five TREC test sets and 10 performance measures, shows a negative correlation between bias and performance. The 2017 article investigated the retrievability of non-relevant and relevant documents. It is shown that good systems make more relevant documents more retrievable, but that this can occur at the expensive of diversity.	
Cho and Roy (2004) and Cho and Adams (2005) considered whether Google, though its use of PageRank, creates a bias for newly created web pages.		Cho and Roy (2004) document the impact of Google's PageRank on newly created pages' popularity. In Cho and Adams (2005), the authors propose formal definitions and metrics for page quality. They then conduct an experiment, which illustrates the bias against newly created, high quality Web pages.	

Jiang et al. (2016) considers the problem of time-related bias in ranking mechanisms for the retrieval of scientific literature.			The authors propose to exploit heterogeneous sources of information from intra- and inter-network sources in ranking scientific literature as well as scientists with respect to one's interests. The experiments show that the method outperforms PageRank and other approaches.
Ortega and Aguillo (2014) compared two platforms for academic searches: Microsoft Academic Search and Google Scholar for finding relevant scientists.		771 Scholar profiles common to both platforms were compared. It was found that Google Scholar contained more information (on publications and citations), however, it was slanted toward the information and computing sciences, as compared to Microsoft.	
Robertson et al. (2019), treated Google and Bing's auto-completion functions as black-box algorithms.	They presented an auditing technique called "recursive algorithm interrogation." They recursively submitted queries, and their resulting child queries, in order to create a network of the algorithm's suggestions.		
Mowshowitz and Kawaguchi (2005) developed a method as well as metrics for quantifying search engine bias.		Using a large set of search queries, they constructed a "fair results set," which consisted of the distribution of URLs obtained for a given query, across a number of different search engines. Bias of a given engine could then be quantified, as the distance between the distribution of URLs returned for a given query, and that of the "fair" set.	
Political affiliation / leaning as diversity dimension			
Robertson et al. (2018) audited Google search engine result pages (SERPs) of 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.	The researchers found little evidence of filter bubble effects, but they did find that left-leaning articles tended to be lower ranked in SERPs and that Google's rankings favored right-leaning content.	

Hu et al. (2019) audited Google SERPs snippets, for evidence of partisanship. The generation of snippets is a black-box process.	Using a collection of partisan search queries, the researchers collected SERPs with the snippets, as well as the source Web pages. Comparing the pages to the snippets, the researchers found evidence that snippets amplify partisanship.		
Le et al. (2019) 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.		
Jiang et al. (2019) aimed to investigate the process of comment moderation (both automated and human) on YouTube videos with partisan content.		The authors analyzed a dataset of 258 political videos, with a total of 84k comments. They found that overall, comments on right-leaning videos were more likely to receive moderation. However, when controlling for other factors such as the presence of hate speech in comments, no political bias was found.	

Table 9: Model-based problems and solutions in IR articles.

Third Party

Many IR articles focus on the influence of human factors in resulting information biases. However, IR researchers tend to focus on the end-user of an IR system or tool, rather than third parties who influence the system behaviours more generally (for other users). These shall be analyzed below (Input/Output).

Fairness

We have collected only one IR publication that focuses on the problem of the fairness constraints that might be used to improve the system. Interestingly, this article, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems," (Mehrotra et al., 2018), while published in a venue we considered to represent an IR community (ACM CIKM), has both an IR and recommendation systems "flavor."

This paper investigates the difficult issue of two-sided platforms, which should satisfy not only the end-user requesting information (e.g., a consumer looking for a purchase recommendation), but also the supplier of the information or product in question. In other words, the problem is to identify the appropriate fairness constraints on the system's recommendations to the user. The solution is *fairness learning*; a conceptual and computational framework is proposed by the authors, in order to evaluate different policies. In particular, the authors consider trade-offs between the relevance of the recommendation to the user, and fairness to both parties involved.

Input/Output

Several of the IR articles collected considered aspects of the user's behaviour (as evidenced from the input/query or her interaction with the system output) might be problematic. *Discrimination discovery (implicit)*, was almost exclusively the solution being proposed in the articles focusing on the user behaviour as the research problem.

Systematic differences in what people search

Two articles focused on characterizing significant trends in the types of information people search, as a function of their sensitive and/or demographic attributes. In a sense, these patterns represent "user bias" that can be exploited for improving system performance. For instance, noting the need for focused Web advertising and improved personalization, Weber and Castillo (2010) 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 behaviours of various segments of the population.

Similarly, Yom-Tov (2019) 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 user cohorts - users with similar profiles who are looking for the same information, in this case, information concerning a health condition. The analysis across three query data sets, revealed significant differences in query formulation, with women in particular executing significant more queries than men, and also using significant longer queries. Given these differences, Yom-Tov noted the importance of careful creation of user cohorts, which are demographically representative of the users that the models will eventually serve.

Pal and colleagues (2012) 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. Finally, in a study of information exposure on the Mendeley platform for sharing academic research, Thelwall and Maflahi (2015) illustrated a "home-country" bias. Articles were significant more likely to be read by users in the home country of the authors, as compared to users located in other countries.

Cognitive and/or perception biases in selecting / attending to information

Other articles in our collection focus on biases that arise 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 (2006) studied the behaviours of 56 participants engaged in e-commerce search tasks. The study was non-invasive as the researchers analyzed the log data collected, with the goal of understanding users' perceptions of sponsored versus unsponsored (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. (2009) conducted a user experiment to examine the effect of position in a search engine results page. 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.

Nikolov et al. (2019) used a Yahoo! Toolbar dataset to study biases in users' exposure to information. In particular, they consider and develop metrics for, homogeneity bias (i.e., the tendency to limit one's set of information sources) and popularity bias (i.e., the tendency to rely primarily on top sites). Exposure biases were characterized across popular Web platforms. For instance, social media and new sites tend to suffer from more popularity bias, while search engines helped to expose users to more diverse sources of information.

Finally, Yu et al. (2017) noted that a novelty bias can impact a researcher's ability to accurately interpret user search log data. They argued that users are more likely to click on novel documents rather than those viewed as redundant, even when they are not high quality / of high relevance. To this end, they model user click behaviour using utility theory, and illustrate in their experiments that this improves click-stream data interpretation.

User beliefs about / during search

Another cluster of user-based studies concerned the impact of users' beliefs about search and/or during a search for information. Kodama et al. (2017) assessed young people's mental models of the Google search engine, through a drawing task. Many informants anthropomorphized Google, and 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.

Ryen White, of Microsoft Research, has published extensively on the dynamic interaction between users' beliefs before, during and after a search, particularly when trying to find information to answer health-related queries. In an initial study (White, 2014), a user study focused on finding yes-no answers to medical questions, showed that users' pre-search beliefs are rarely changed; when they are changed it is typically in the direction of "yes." It was also clear that pre-search beliefs influence users search behaviours. 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 (2015) looked more specifically at users' beliefs on the efficacy of medical treatments, and how these beliefs could be influenced by a Web search. The Cochrane collection of medical meta-analyses was used as ground truth on the efficacy judgments. The results echoed the previous study, in that users' pre-search beliefs shaped their information behaviours. To that end, they provided insights on how to model users' beliefs during a search for medical treatments, to provide better, more personalized search results.

Finally, Otterbacher et al. (2018) described a user study in which participants were shown image search results for queries on personal traits (e.g., “sensitive person,” “intelligent person”). Some result sets were relatively gender-neutral, while others were heavily skewed toward reinforcing the predominant prescriptive gender stereotype of the warm / agentic woman/man. Users were asked to evaluate the search results on a number of aspects, including the extent to which they were “biased.” They also completed a common psychological test to measure of sexism. It was shown that more sexist users (both men and women) were less likely to report a heavily gender-biased results set.

User / system interaction

Several studies considered the interaction between the user and a system, or a particular system component, as possible insight in solving information biases. Mitra et al. (2014) presented the first large-scale study of users’ interaction 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 after having typed at least half of their query. In a follow-up study, Hofmann et al. (2014) conducted an eye-tracking study with Bing users. In half of their queries, users were shown ranked auto-complete suggestions; in the other half the suggestions were random. The authors again 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 behaviours; in the random setting users visited more pages in order to complete their search task.

Epstein et al. (2017) aimed to 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 searchers 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 more when more detailed bias alerts were provided to users. Nonetheless, they reported that SEME cannot be completely eliminated with this type of intervention, and suggest instituting an “equal-time” rule such as that used in traditional media advertisements.

Finally, Maxwell et al. (2019) investigated the influence of result diversification on users’ search behaviours. Diversification is meant to 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 - were assigned to each user, and each performed tasks using a non-diversified IR system as well as a diversified system. Results indicated significant differences in users’ search behaviours 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.

5.2.2 Diversity dimensions in the IR literature

As previously mentioned, although the CyCAT remit is to study social and cultural biases in algorithmic systems, we included in our survey articles focusing on *information* biases. Perhaps unsurprisingly, more than half of the articles (31) we found on biases in IR systems concerned the nature of the information presented to users and how certain characteristics of the information was impacted by algorithmic mitigation, without any consideration of the user's (or the information's) social or cultural characteristics. In such articles, we considered information itself to be the diversity dimension of interest in the research.

For instance, an early work by Mowshowitz and Kawaguchi (2005), proposed a method of quantitatively measuring search engine bias, by comparing the results retrieved across a number of search engines, based on a large set of queries. As a second example, Hofmann and colleagues (2014) considered users' interactions with the query auto-complete feature in a search engine. They identified a position bias, which resulted in users more often following the first suggestions, regardless of their actual quality. In summary, much of the work on algorithmic biases in the IR literature focuses on how algorithmic processes can result in systematic biases in terms of the information accessed by users.

Algorithmic biases rooted in social and cultural dimensions are less often the subject of research in the area of IR. Nonetheless, we identified several other dimensions being described in the articles in our repository. Below, we summarize the diversity dimensions examined in the collection of IR articles, and provide an example of each.[6]

- Age (2)

Yom-Tov (2019) was concerned with the ability of a search engine to identify cohorts of users, based on their input queries. In this study, a cohort referred to individuals having the same health condition, searching for relevant information. The author detected systematic differences in the textual queries, based on user age, gender and income, which pose challenges for the creation of user cohorts that are balanced and representative, and ultimately, could affect the perceived fairness of search engine results.

- Culture (1)

Callahan and Herring (2011), considering Wikipedia as a data source for training algorithmic processes, studied cultural biases across language editions. In particular, they compared a set of articles on famous persons (both Polish and American), described in the English vs. Polish language editions of Wikipedia. Despite the fact that Wikipedia has a 'neutral point of view' policy, numerous cultural biases were found, such as using a more positive tone for in-group persons, and a tendency for the English Wikipedia to include more personal information about the target person.

- Gender (8)

[6] Note that because some articles address multiple diversity dimensions, the total number of cases exceeds the number of articles studied.

In a study of Google image search, Kay and colleagues (2015) found gender-based biases in terms of the portrayal of the professions. For a given profession (e.g., “doctor,” “nurse”), they compared the gender distribution in retrieved images, based on the profession query, to published U.S. labour statistics. They demonstrated that gender stereotypes were not only perpetuated in the image search results, they were even more extreme than expected, based on labour statistics.

- Information (31)

(Examples above.)

- Language / linguistic (2)

There are two examples of language-based biases in our repository; both articles examine sentiment classification on texts. In Rafrafi et al. (2012), the authors offer a means to address document frequency bias. They note that when using linear supervised classifiers to train a polarity detection algorithm, that terms occurring frequently in the input texts end up being overweighted, compared to their actual subjectivities. In contrast, Davidson et al. (2017) address bias in classification algorithms for hate speech detection. This discrimination discovery work demonstrates a tendency for offensive speech to be classified as hate speech inadvertantly. It also shows that sexist language is less likely to be deemed hate speech, as compared to racist and homophobic speech.

- Minority status (1)

Dixon et al. (2018) considers text classification for “toxic” language, which is trained on Wikipedia talk pages. They discover a bias related to words associated with minority status (e.g., “Muslim” or “gay”), in that texts containing these words are more likely to be classified as being toxic, regardless of the ground truth. The problem lies in the training data, which over-represents cases of toxic text with these words.

- National origin (1)

Thelwall and Maflahi (2015) considers the issue of reader exposure on the academic sharing platform, Mendeley. Their focus is on how reader behaviours could lead to articles systematically having more/less exposure. They show that articles are disproportionately read by users in the home country of the author(s).

- Political affiliation / leaning (8)

Given the importance of information access systems in the political process, it is unsurprising that several articles examined algorithmic system biases based on political leaning (either that of the respective user, or that conveyed in the relevant content). In an auditing study on Google search, Robertson et al. (2018) asked participants to install a browser extension collecting their search engine results pages (SERPs). The researchers were looking for evidence of ‘filter bubble’ effects, but instead found that generally, left-leaning results were more likely than right-leaning results to be positioned in lower-ranking positions on the SERPs. Another example is the work on Jiang and colleagues (2019), which considered algorithmic moderation on YouTube comments made on political videos. The research question concerned whether or not there was evidence of political bias; however, while authors found more moderation at right-leaning videos, once the presence of factors such as hate speech were controlled, there was no evidence of political bias.

- Race (1)

Shen et al. (2018) considered the problem of stylistic bias in sentiment analysis algorithms. In particular, they were concerned with the use of African-American English (AAE). They found that a text containing markers of AAE was significantly more likely to be deemed as having negative sentiment, as compared to a semantically equivalent text, containing no markers of AAE. They proposed a translation process for input data, in order to mitigate the algorithmic bias.

- Sensitive attributes (3)

Some articles do not consider a particular social or cultural attribute, but rather more generally, focus on sensitive attributes more generally. For instance, Kliman-Silver et al. (2015) examined the influence of the user's geolocation (a proxy for many sensitive attributes including income, social status, or even race) on the ranked results returned to her during a personalized search for information. A similar example is found in Weber and Castillo's (2010) work, who study web search logs in order to uncover correlations between users' queries and their sensitive demographic attributes.

5.2.3 Summary of IR literature

In summary, we can make the following observations concerning the research trends to date in the IR literature:

- Diversity dimensions: *Information bias* is more of a concern in the core IR venues, as compared to social / cultural biases.
- Research focused on social / cultural biases in IR systems, can be found in more interdisciplinary journals and conferences, such as JASIST or AAAI ICWSM.
- Three key problem areas have been explored in the IR literature: data-, model- and user-based problems.
- Solutions being proposed for data- and model-based problems in IR systems include auditing, discrimination discovery, as well as fairness (sampling and learning). However, in user-focused studies, the emphasis is exclusively on discrimination discovery.

5.3 Recommender Systems

In total, 37 papers from the target RecSys publications were archived, about 10% were found to be irrelevant after they were reviewed. The publication venues were extremely diverse. The largest number of papers, five, came from the FAT conference proceedings. In parallel to the previous sections, Section 5.3.1 presents the problems / solutions explored in these papers, while Section 5.3.2 details the diversity dimensions of interest.

5.3.1 Problems examined / solutions proposed in RecSys

Data

Studies in RecSys showed classical FAT-related problems resulting from imbalanced datasets and correlation between protected attributes and various proxies. Recommender systems are the domain where the issues of FAT appeared to begin with and hence there is no surprise that this characterises data-related problems of RecSys.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Wachs et al. (2017) aimed to show the effect of genderness on success in Dribbble (a social community for user-made artwork).		The researchers found correlation between users skills and network structure on social community and their gender. For revealing user gender they used user names or her photo.	
Chakraborty et al. (2017) aimed to understand who made trends.		The authors showed that under-represented population from certain demographic groups (race/gender/age) does not make trending topics.	
Celis et al. (2019) embedded polarization in the personalization algorithms.			The authors used clusters centers as arms in their bandit algorithm.

Table 10: Data-based problems and solutions in RecSys articles.

Model

Only a few studies in RecSys related publications somehow tried to examine the fairness of the model itself.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Bellogin et al. (2017) studied statistical biases in the evaluation metrics of recommender systems.		The authors referred to popularity and sparsity biases in metrics.	
Eslami et al. (2017) aimed to understand how users perceived and managed biases in reviews	The authors used a cross-platform audit technique that analyzed online ratings. This method is used to analyze the behaviour of the		

	algorithm according to different inputs, where it should not behave differently		
Hannak et al. (2017) found that TaskRabbit and Fiverr are affected by gender and racial biases.	In their study authors performed scraping audit (on crawled data).		

Table 11: Model-based problems and solutions in RecSys articles.

Third Party

Discrimination by third parties seems to be challenging in algorithmic systems, not only but also in RecSys.

Problem	Auditing	Discrimination discovery	Fairness sampling
Speicher et al (2018) claim and show that advertisers in Facebook can discriminate populations		The authors claim that the fact that Facebook disallow the use of attributes such as ethnic affinity by advertisers when targeting ads related to housing or employment or financial services is not enough and that there discrimination measures should be based on the targeted population and not on the attributes used for targeting	
Zhang (2015) examined what kinds of personal data mobile apps share, in addition to geolocation information, and how often.	They propose ways for limiting the sensitive attributes sharing between apps and discuss the lack of formal regulation from the government side.		
Ribeiro et al 2018. On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook - Targeting specific social groups with malicious ads using Facebook ads targeting service			
Edelman (2017) found discrimination in accepting applications in AirB&B		In an experiment on Airbnb, applications from guests with distinctively African American names are 16 percent less likely	

		to be accepted relative to identical guests with distinctively white names. Discrimination occurs among landlords of all sizes, including small landlords sharing the property and larger landlords with multiple properties. It is most pronounced among hosts who have never had an African American guest, suggesting only a subset of hosts discriminate.	

Table 12: Problems and solutions concerning Third Parties in RecSys articles.

Fairness

Fairness is an important issue in RecSys and indeed, about 25% of the papers that were reviewed considered fairness in aspects of RecSys.

Problem	Fairness Certification	Fairness Sampling	Fairness Learning
Karako and Manggala (2018) incorporated fairness in the ranked results to improve precision.		Gender-based fraction of labeled images in the data set	Post-processing step for recommender system
Leonhardt et al. (2018) pointed out on trade-off between recommendation diversity and user fairness			Post-processing for recommendation results
Mehrotra et al. (2018) proposed a personalization recommender system, balancing between user satisfaction, relevance and fairness.			Providing fairness weights according to artist popularity
Chakraborty et al. (Equality of voice, 2018)			Weighting of crowdsourced recommendations to balance out the effects of the silent majority, multiple voters and vote splitting for similar items

Zehlike et al. (FA*IR, 2018)			Adjusting a top-k ranking algorithm to provide fairness for underrepresented groups (race, gender, disability, etc.) where a minimum representation is prescribed by law
Singh and Joachims (2018)			Top-k ranking under fairness constraints
Ekstrand et al. (2018) researched effect of demographics on evaluation metrics in recommender systems.		The authors randomly sampled the same number of male and female users.	
Xiao et al. (2017) Fairness in group recommendation		The authors considered the issue of fairness in group recommendation and suggest an optimization framework for fairness-aware group recommendation.	

Table 13: Problems and solutions concerning Fairness in RecSys articles.

Input/Output

Table 14 analyzes the three articles that concerned biases in the Inputs and/or Outputs of RecSys.

Problem	Recsys	Inputs	Outputs
Sweeney (2013) bias in ad server recommendations caused by positive reinforcement of discriminatory ads accessed when searching for black names	Ad server	Small samples of names with a proven racial bias compared with non-racially identifiable names	Ads for discovery of criminal records disproportionately displayed for black identified names
Ali et al. (2019) detected significantly skewed ad delivery on gender and racial lines in ads for employment and housing.	Ad server	Facebook profiles	Ads for employment and housing
Rosenblat and Stark (2016) investigated Uber's	Uber	Drivers profiles and history	Drivers ratings

drivers perception of the application			

Table 14: Problems and solutions concerning the system's Input/Output in RecSys articles.

Explanation promotion

This aspect seems to be particularly important in the research on RecSys. In this domain the need for explanation and transparency as means to increase trust in what is considered "black box" systems and increase their acceptance (and even their performance) is commonly agreed and indeed a large number of papers (about 40%) dealt with explanation promotion and its importance.

Solution	**Explainability promotion**	**White Box**	**Black Box**
Kouki et al (2019)Personalized Explanations for Hybrid Recommender Systems	The authors propose a system that generates personalized explanation to users of Hybrid (content-based and collaborative) recommender systems		
ZHIYONG et al 2018. The authors present a study about Explainable Recommendation by Leveraging Reviews and Images			
Nunes, I., & Jannach, D. (2017) presented a systematic review on explanations for recommendations in decision support systems.	The authors proposed a taxonomy of concepts that are required for providing explanations.		
Abdollahi, B., & Nasraoui, O. (2018) proposed a taxonomy of explanation styles for various recommendations.	The authors proposed a taxonomy of explanations styles.		Explanations of both model and an outcome (recommendations).
Bountouridis et al. (2019) proposed a simulation framework of news consumption that took into consideration an article content and prominence..	The authors provided a simulation framework for visualization effects of recommender systems to the content providers.		

ter Hoeve et al. (2017) on the need and desire for explanations of content delivered by a news server	The authors discovered that users want explanations, but don't know what type of explanation is preferable, and that the presence or absence of an explanation does not impact the click through rate		
Wang et al. (2018) give an algorithm to generate explanations of recommendations by analyzing the record of user text based opinions			Explanations are not based on analysis of the recommender algorithm (black box), but rather on user opinions and evaluation of previous purchases
Stoica et al. (2018) studies the effect of gender, homophily and growth dynamics under social recommendations.	The authors analyzed Instagram network to understand the effect of gender and homophily, and showed the existence of an algorithmic glass ceiling.		
Eslami et al. (2018) investigated how making algorithmic process more transparent may affect users' perceptions towards ads and platforms.	The authors found that from one side users preferred interpretable and non-creepy explanations, however, from the other side, after getting those they understood that ads are much worse then they used to think.		
Konstan and Riedl (2012) in a review of recommender systems discuss the general issue of keeping the user in control and the need for transparency and explanations as a way to increase trust in recommender systems	A strength of recommender systems is that they reduce the workload on users who are overwhelmed by the choices available to them. However, users are often more satisfied when they are given control over how the recommender functions on their		

	behalf—even, in some cases, when that control increases the effort required of them, and when the resulting recommendations are objectively less accurate. The sweet spot is recommenders that balance serving users effectively, while ensuring that the users have the control they desire.		
Cheng et. al. (2019) aims to show that enhancing the limited transparency of matrix factorization based recommendation with images has a positive impact on the prediction accuracy - hence implicitly suggest that transparency improves the performance of recommenders	The authors coded items' images as part of items representation and demonstrated improved recommendation		
Andreou et al. (2018) evaluated Facebook explanations for ADs recommendation their results show that ad explanations are often incomplete and sometimes misleading while data explanations are often incomplete and vague			

Table 15: Promoting explanation / interpretability in RecSys articles.

5.3.2 Diversity dimensions in Recommender Systems (RecSys)

Below the diversity dimensions examined in our collection of RecSys articles are summarized and exemplified.

- Information (7 articles)

Cañamares and Castells (2007) examined popularity biases (i.e., how basing recommendations on other user's tastes might affect the information presented to others). Nunes and Jannach (2017) suggested a taxonomy of explanations in recommender systems. For example, content-tailored explanations suppose to explain recommendations according to users interests, expertise or current context. Karako and Manggala (2018) incorporated fairness in the image ranking. They chose a sample of labeled images, based on gender, though their method is suitable for any information. Leonhardt et al. (2018) argued about diversity of recommender system results (movies recommendations in their case). Abdollahi and Nasraoui (2018) researched a diversity in explanation styles of recommendations. Mehrotra et al. (2018) considered artists popularity as a weighting for fairness. Bellogin et al. (2017) studied the effects of sparsity and popularity bias on evaluation metrics in recommender systems. Bountouridis et al. (2019) proposed a simulation framework of news consumption that took into consideration an article content and prominence. They also used *long-tail diversity* and *unexpectedness diversity*. Stoica et al. (2018) researched a diversity of the network - Instagram users. Rosenblat and Stark (2016) investigated driver perception regarding Uber application, given drivers profiles and their history performance. Eslami et al. (2018) aimed to understand why ads are presented to specific user. They discovered that users preferred to get explanations included specific information that an advertiser used to target an ad, however they should not be “creepy”.

Chakraborty et al. (2018) attempt to find a fair ranking for crowd sourced recommendations taking into account that the vast majority of potential voters are silent, that some people vote multiple times, and that votes for similar topics are split, leading to a bias towards extreme viewpoints. They suggest a system that is widely applicable to k-best rankings and eliminates the biasing due to the issues above.

- Gender (3)

In a study of social community for user-made artworks (Dribble), Wachs et al. (2017) found gender-based biases in user success and popularity on this platform. For a given Dribble user profile, using their username, they revealed the real user name from Twitter. Then they inferred gender from the real name. Authors found that differences in skills and social network structure also affected by gender-based differences.

Ali et al. (2019) detected significantly skewed ad delivery on gender lines in facebook ads for employment and housing. Hannak et al. (2017) found that TaskRabbit and Fiverr are affected by gender and racial biases: “perceived gender and race are significantly correlated with worker evaluations”.

- Race (3)

Sweeney (2013) investigated the advertising recommendations by an ad server when searching for particular names in Google and Reuters search engines. She found 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 (like Latanya). The paper gives informal accounts of techniques for detecting and remedying such biases in the ad server.

Ali et al. (2019) detected significantly skewed ad delivery on racial lines in facebook ads for employment and housing. Hannak et al. (2017) found that TaskRabbit and Fiverr are affected by gender and racial biases: “perceived gender and race are significantly correlated with worker evaluations.” Edelman (2017) showd ethnicity/race based discrimination in Air B&B applications. However, it does not seem that the system is the origin of the discrimination.

- Sensitive attributes (3)

Chakraborty et al. (2017) examined the influence of the user’s demographics (race, age and gender) on the trends promotion on Twitter. Zang et al. (2015) analyzed how frequent mobile apps share user geolocation and whether they share other sensitive attributes. Celis et al. (2019) embedded polarization in the personalization algorithms. As one of the data sets they use online news articles that are conservative or liberal. Zelike et al. (2018) provide a way of ensuring individuals with protected status achieve required levels of representation in a ranking algorithm, without compromising on the utility of the ranking. Singh and Joachims (2018) also provide an algorithm for ranking under fairness constraints. Eslami et al. (2017) aimed to detect biases in the algorithm according to different sensitive attributes.

- Language / linguistic (1)

Hannak et al. (2018) measured linguistic biases in reviews on TaskRabbit and Fiverr to detect abstract and subjective language.

- Minority status (1)

Bountouridis et al. (2019) considered long-tail diversity in their simulation framework of news consumption.

- Age (1)

Ekstrand et al. (2018) found a demographic (age, gender) differences in recommender accuracy. Moreover, they claimed that demographic effects interacted with the popularity bias.

5.3.3 Summary of RecSys literature

As far as diversity dimensions, a large number of dimensions was represented in the RecSys review. However, the RecSys literature seems to be a bit limited; first of all, about 10% of the papers we found using our current methodology were found to be irrelevant. The remaining papers are a diverse set of papers dealing with FAT related aspects in RecSys, but somehow it gives the feeling that this is just the tip of the iceberg in this area. The vast majority of the papers (13 out of 37) investigated and motivated the need for explanations as a way to increase users’ trust in RecSys and also improve RecSys performance. *Explanation promotion* stands out as characterising the RecSys domain, as it is an issue that has been pointed out and discussed for many years in the RecSys literature, as a tool that can increase users’ trust and improve the performance of RecSys.

5.4 Human-Computer Interaction

In total, 25 articles from the target HCI publications were archived. The majority of the articles were published at the ACM Conference on Human Factors in Computing Systems (11 articles) and other related conferences e.g., ACM Intelligent User Interfaces (3 articles) and ACM Conference on Computer-Supported Cooperative Work and Social Media (2 articles).

5.4.1 Problems examined / solutions proposed in HCI

Among the literature we examined, we came across a paper that was proposing a conceptual framework for making transparency clear (Springer & Whittaker, 2019). In their work they are proposing that transparency can be achieved in two ways, through explainability and auditability. In their framework, explainability builds on user experience and enhances trust with the system and auditability can provide third parties the opportunity to test algorithmic outputs and identify biases and possible fairness issues. Furthermore, Shin and Park (2019) stress the importance of helping the user to understand algorithmic affordances in the adoption and use of a system. They have identified that the user experience is affected by the lack of system's transparency and statistically proven that fairness, accountability and transparency in algorithmic systems can help the user to understand how the system takes decision e.g., recommendations and effectively develop trust to the system.

Data

In the HCI research we have found only one paper that looked into experimenting with training data. However, in this study Johnson et al (2017), are initially analysing data collected from Twitter before they manipulated this and use them as training data. The outcome of this study suggests that bias, in some cases, occurs not only in the training data but at the structural elements of the algorithm itself.

Problem	Auditing	Discrimination discovery	Fairness sampling
Demographics Discrimination dimension			
Johnson et al. (2017) looked into the problem of biases based on demographics in social media used algorithms.	They have used twitter API to retrieve geotagged content for this study	The authors identified that demographic (urban-rural) algorithmic biases exist for rural users, where both algorithms (text-based and network based) performed worst compared to urban users. It is interesting that even after balancing the data or oversampling there was a bias towards urban users.	The outcome of this work suggests that bias in some cases occurs not only in the training data but at the structural elements of the algorithm itself.

Table 16: Data-focused studies in the HCI literature.

Model

Four papers looked into the algorithmic model of a system. All four papers report some kind of discrimination discovery that was detected. Three papers report gender and/or age bias and one is looking into race, socioeconomic status, touching also on location and ethnicity. All papers touch on the model being the problem in generating the discrimination, however, for example in Brown et al. (2019), it is difficult to know for sure that the training data did not also play their part.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Gender & Age Diversity dimensions			
Chen et al. (2018) in their work looked into gender-based inequalities in the context of resume search engines.		The authors looked into direct and indirect discrimination by a system towards its users. Direct discrimination happens when the system is explicitly using the inferred gender or other attributes to rank candidates, while indirect discrimination is when the system unintentionally is discriminating over its users. The results show that the system under review is indirectly discriminating against females however, it does not implicitly using gender as a parameter.	
Salminen et al. (2019) 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 was becoming in terms of gender and age representation. Most notably, in low numbers of generated personas biases are increased.	Practitioners who use data generated personas should consider the possibility of unintentional bias in the data they use, that consequently are transferred to the personas they generate.

Keyes (2018) identified the problem of Automatic Gender Recognition in HCI research and how the approaches followed until recently are discriminating upon trans gendered people		The consequences that this perspective will have when incorporated to real world applications.	For systems to be fair towards the trans gender direction, Keyes suggests alternatives to automatic gender recognition and development of more inclusive approaches to evaluate and infer gender.
Race & Socioeconomic Status Diversity dimensions			
A qualitative study was performed by Brown et al (2019) for understanding the public's perspective on algorithmic decision making in public services.		Many participants mentioned discrimination and bias based on race, ethnicity, gender, location, socioeconomic status.	The authors identified that the human in the loop approach for decision making when sensitive attributes are involved is preferred rather than the statistical model approach. In addition, participants requested for access to the information, and explanations on, how the algorithm took some of the decisions and the parameters the decisions were taken upon.

Table 17: Model-focused studies in the HCI literature.

Third Party

Online systems that are relying on human generated data as a way of operating are prone to biases. Two of the papers we reviewed, looked into OpenStreetMap data only to find that the human generated data there is biased. Das et al. (2019) looked into differences between genders when generating data in the system. However, they discovered differences in demographics between urban and rural users. Similarly, Quattrone et al. (2015) identified that geographic information bias exists due to the fact that a lot of content is generated by few people.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Demographics discrimination dimension			
Das et al. (2019) looked into gender differences in OpenStreetMap contributions and potential biases	Analysed OpenStreetMap data to understand the potential differences between male and female contributors, and how these can be a source of self-focus-bias.	No discrimination detected in this dataset in terms of gender, however they authors have found that females tend to focus on more urban rather than rural areas compared to males. Differences do exist in the behaviour of the two genders with male contributors contribute more to feminized spaces and females contributing to more masculinized spaces.	
Information Diversity dimension			
Quattrone et al. (2015) looking at the OpenStreetMap data generated by crowdworkers, they have identified that most of the content was created by a small number of the registered users.		Geographic information bias was detected. The authors have also discovered that culture played a role in this kind of bias.	

Table 18: HCI literature studies concerning the role of the Third Party.

<u>Fairness</u>

Although Fairness is one of the most discussed topics in HCI literature, only one paper provides suggestions on how to incorporate it within the cycle of User/System Interaction. We further discuss how Fairness affects User/system Interaction later.

Problem	Auditing	Discrimination discovery	Fairness sampling
Perceived fairness was examined in Lee and Baykal (2017).		The algorithmic decisions were perceived as not fair by participants for either individuals or groups. This was primarily due to algorithms not account for multiple concepts or fairness or social behaviours.	The authors suggest that to improve the perceived fairness there is a need to provide the opportunity to people to intervene in the process of algorithmic decision making.

Table 19: Fairness in the HCI literature.

Input/output

Three papers looked into the output of algorithmic models for identifying potential discrimination issues. Race, gender and age have been associated with biases within the output with discrimination to be inevitable when these models are incorporated in intelligent systems.

Problem	Auditing	Discrimination discovery	Fairness sampling
Race & Gender Diversity dimensions			
Green and Chen (2019) run a crowdsourcing study to examine the influence of algorithmic risk assessment to human decision making.		Race was influencing the human risk assessment. Specifically, participants were more likely to increase their risk prediction when black defendants were involved and more likely to deviate from their initial predictions towards higher risk levels.	Need to examine whether real judges express the same behaviour on duty.

Barlas et al. (2019) 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.	For algorithmic generated descriptions Carifai's image analysis algorithm was used.	The authors identified that human generated descriptions were perceived as more fair except when the depicted person was white and attractive where Clarifai's descriptions were perceived as more fair. Thus, fair treatment is not expected across social groups. In addition, there were evidences that men compared to women are more likely to discuss fairness related to physical characteristics of the person depicted in the image.	Fairness was judged by crowdworkers based on the dimensions of accuracy of the descriptions, physical characteristics of the depicted person, and based on objectivity/subjectivity of the descriptions.
In Matsangidou and Otterbacher (2019) the authors are looking into inferences on attractiveness made by image tagging algorithms following the evolutionary biases theory.	The authors audited four image recognition APIs for their inferences on attractiveness using the Chicago Face Database as input.	Discrimination discovered between male and female depicting images. A negative association was observed between masculine and attractive tags, as well as youthfulness positively correlated to attractiveness. Furthermore, more attractive people were associated to positive emotions, while images of white people were associated to more positive tags in comparison to other racial groups.	Make developers aware of the dangers for discrimination and bias when using Image Tagging APIs in their applications, so they consider dimensions of making those more fair to sensitive and under-represented groups.

Table 20: HCI studies focusing on the system's input/output.

User System Interaction

The most frequently appearing category in HCI literature is the User System Interaction. Nine papers have been reviewed with the most popular concept being Fairness, however, there are more papers that are discussing fairness and relate to other categories as well, e.g. Barlas et al. (2019)). The Fairness concept is usually discussed along with trust to the system and/or its output (e.g. Woodruff et al. (2018)). Fairness is not a simple concept according to the literature. We came across a lot of papers discussing the different dimensions of Fairness Hlaca et al. (2018) and trying to understand how the users perceive Fairness in a system (Bins et al. (2018)). Other studies looked into transparency and how much is too much. For example, Eslami et al (2019) discovered that the level of transparency that a system provides to the user affects the behaviour of the user with the system. Transparency has been found to positively affect the trust of the user to the system and to correlate positively with the user's engagement with the system. Different explanation approaches have been examined at Rader et al (2018), that proved to improve transparency and benefit the user interaction with the system.

Problem	**Auditing**	**Discrimination discovery**	**Fairness sampling**
Eslami et al. (2019) Biased and opaque algorithms and how people perceive and interact with these algorithms	The paper describes a qualitative study of online discussions about Yelp on the algorithm existence and opacity. The authors further enhanced the results with conducting 15 interviews with Yelp users. In this work the users acted as auditors of the Yelp system in an attempt to understand how the reviews filtering algorithm works.		Transparency should begin with acknowledging the very existence of the algorithms. Furthermore, they suggest that while a level of a system's transparency is required, full transparency is neither necessary nor desirable since this affects the user behaviour with the system
Chen and Sundar (2018) looked into perceived control and how it relates to overt personalization and information transparency.			Overt personalization affects perceived control of the user during system interaction. Information transparency was found to affect positively trust and negatively user information privacy, while positively correlate to user engagement and product involvement.
Eiband et al. (2018) proposed a participatory design methodology for incorporating transparency			Through the transparency that will be a core part of the design process, the authors aim at making

in the design of intelligent user interfaces.			intelligent systems more fair, transparent and explainable.
Lee (2018) examined how people perceived algorithmic decisions when tasks required human versus mechanical skills.			With tasks that require mechanical skills participants in their study trusted algorithmic and human decisions equally and though they were fair. With tasks that involved human skills, participants thought that algorithmic decisions were less fair and trusted the decision less.
Rader et al (2018) examined four different explanation methods for materializing algorithmic transparency. They argue that transparency can empower users of decision support systems in making informed choices.			Help the participants identify when the system was biased, helped them to make informed decisions on their follow-up actions and how they could control what they were seeing. All participants appreciated the explanations and all explanations had an impact on user awareness. However, the explanations were not as effective for evaluating the correctness of the system's output and the consistency of its behaviour.
News consumption is a very common activity for a wider population. Horne et al. (2019) went to investigate whether AI assistance can improve the user perception for bias and how different types of users are affected in this end.		The findings of this study suggest that users who read and share news often are worse in identifying reliability and bias issues in news articles than those who do not. However, those who are familiar with politics and are frequent	

		news readers performed better.	
Binns et al. (2018) run three experimental studies in understanding how people perceive justice in automatic algorithmic decision making.			They have found that for some the whole idea of an algorithm deciding is perceived as unfair while others thought that the algorithm does what it is supposed to do as long as the information upon the decision is made is accurate. In addition, the results presented here show that further work is needed to understand when, and what type of, explanations should be provided along with system's decision.
Race & Socioeconomic status diversity dimensions			
Woodruff et al. (2018) explore in a qualitative study the perception of algorithmic fairness by populations that have been marginalized due to their race and socioeconomic status.			Most participants were not aware of algorithmic unfairness even thought they have experience with discrimination in their daily lives. Particularly stereotyping and discrimination due to their race e.g. with law enforcement, disadvantageous targeted advertising. Through the interviews and workshops carried out, the participants expressed concerns about their trust to the companies and the government services.

Hlaca et al. (2018) are looking into how humans perceive fairness in algorithmic decision-making systems. They have developed a framework for understanding why people perceive certain features as fair or unfair.			The authors found that people's unfairness perception involves several dimensions that do not only concern with discrimination. In addition, people answers show disagreement in their fairness judgments.

Table 21: HCI studies focusing on system/user interaction and user perceptions.

5.4.2 Diversity dimensions in the HCI literature

Within the broader literature in HCI we have found articles that are describing both quantitative and qualitative studies, for exploring and understanding issues with system's Fairness, Transparency, Biases and Accountability. However, there are articles that explore users' perception of fairness or user evaluations of different transparency methods. For example, Veale et al. (2018) conducted interviews with public sector servants who deal with either the development or the use of public sector decision support systems. In their work they have identified several issues including the absence of discrimination aware machine learning approaches and lack of transparency in the decisions provided. They finally discuss the ethical implications of these systems and stress the need for changes in their design to become more fair and accountable. In their work Woodruff et al. (2018) run a series of workshops and interviews with participants coming from several populations, with the aim to identify fairness or (un)fairness related to racial and social status in the society. Indifferent from the above, Chen et al. (2018) run a statistical analysis on a large dataset from identifying direct and indirect discrimination related to their gender.

Diversity dimensions are taken into consideration in most of the articles we reviewed. In this section we will summarise this work and the diversity dimensions that are taken into consideration in both quantitative and qualitative approaches.

- Gender (5)

Chen et al. (2018) in their work looked into gender-based inequalities in the context of resume search engines. The results show that the system under review is indirectly discriminating against females however, it does not implicitly using gender as a parameter. Gender was also identified by Brown et al (2019) as a discrimination dimension in decision making at public welfare services.

- Race & Socio-economic status (3)

Woodruff et al. (2018) explore in a qualitative study the perception of algorithmic fairness by populations that have been marginalized due to their race and socioeconomic status. In particular they looked into 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 thought they have experienced with discrimination in their daily lives. Brown et al (2019) run also 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, socioeconomic status. Race was influencing the risk assessment judgment of participants in the crowdsourcing study performed by Green and Chen (2019).

- Demographic (2)

Johnson et al. (2017) focused on identifying discrimination and biases in social networks algorithms between urban-rural populations. Their results suggest that even when they overcorrect the training data algorithms still behave in a discriminatory manner. Similarly, Das et al. (2019) looked into gender differences in OpenStreetMap contributions and potential biases. No discrimination detected in this dataset in terms of gender, however the authors discovered that females tend to focus on more urban rather than rural areas compared to males.

- Information (1)

Quattronne et al. (2015) analysed data collected from OpenStreetMap users' contributions to find that only very few of the 1.2M registered contributors have actually contributed to the system. Although, they found no content bias, they have discovered significant geographic bias varying also by culture.

5.4.3 Summary of HCI literature

After the review of the HCI articles in our collection, we made the following observations of the trends in this domain:

- Bias in some cases occurs not only in the training data but at the structural elements of the algorithm itself. Consequently, the model itself can be the problem for discrimination generation.
- User-generated data (Third Party) proved to be biased in different dimensions; thus, this type of data should be used with caution either as input data or as training data to algorithmic models.
- Conceptual frameworks on building fairness in algorithmic systems have been proposed but further work is required on implementing those in real systems.

- Race, gender and age have been associated with biases within the output of algorithmic systems, so when these models are incorporated in intelligent systems, discrimination is inevitable, for some groups or individuals.
- Fairness appears to have different dimensions and perceived differently by different users. These dimensions can be relevant or irrelevant according to the system's application and context. Future work aiming at building fairness parameters into algorithmic systems need to take into account the results of the qualitative studies discussed in HCI.
- Transparency proved to improve user engagement and interaction with the system, however, different levels of transparency are needed to be explored.

5.5 Other Areas

Finally, we analyze the set of 43 articles collected in the repository that did not align with the above four research communities, and that we have labelled as "Other." In particular, we are interested in discovering whether these articles might represent research communities that are only recently addressing problems related to FAT and also whether the diversity dimensions / problems / solutions discussed in this literature differs significantly from the communities on which we have focused.

Non-systems articles

Although our aim was to review articles that describe particular algorithmic systems, occasionally we identified pertinent articles that were more broad in nature. Several articles collected aim to raise awareness of the issue of algorithmic biases within a particular professional or scientific community. For instance, Ayre and Craner (2018) writing in *Library Quarterly*, provide examples that illustrate how biases in library information systems can challenge the work of professional librarians. Similarly, the article by Diakopoulos and Koliska (2017), published in *Digital Journalism*, aims to highlight the issue of algorithmic transparency - as well as the human role - in mitigating bias in news media systems. The CACM article by Baeza-Yates (2018) provides a general, broad introduction to the various sources of bias that users may encounter while using the Internet and social media. Finally, Pope and colleagues (2018) aim to raise awareness in the Management Science community, concerning racial bias in algorithmic social media systems.

Five papers can be described as being conceptual works directed at information systems researchers and practitioners, in a broad sense. Friedman and Nissenbaum (1996), writing before the rise of Big Data and the widespread use of Internet technologies, defined computer bias as occurring when: i) there is a systematic slant in the manner in which information is presented to users, ii) that slant could result in discrimination againt certain people or groups. They also provided several examples from business information systems. Noble's article (2013) brings specific attention to the problem of racial bias in Google search, with a particular emphasis on detailing the consequences that this might have on young, Black girls.

Jenna Burrell (2016) provides a conceptual overview of opacity in algorithmic systems, describing three main reasons that systems lack transparency. These include: i) that systems are often proprietary (i.e,. constitute trade secrets), ii) that systems are technically complex (i.e., based on methods that are not interpretable or explainable), iii) that users lack the technical literacy necessary to understand how the systems work. Olteanu et al. (2019) present a survey focusing on the potential biases of social data, which are used for analytics as well as for training algorithmic systems. They particularly focus on the digital traces left behind in social systems (i.e., user-generated content), and they outline a framework for detecting problems with such data, for researchers and practitioners. Finally, in a Big Data & Society article, Veale and Binns (2017) provide a discussion, from an organizational perspective, on issues concerning the implementation of fairer machine learning "in the real world." In particular, three approaches are presented in order to enable organizations that lack specific knowledge/capacity on fairness issues, to identify and manage them in their everyday contexts.

Seven papers from the domain of law appear in our repository. The article by Barocas and Selbst (2016) provides a general introduction into the ethical problems associated with Big Data, emphasizing big data practices' tendency to harm minorities and those of a lower socio-economic status. In a similar vein, (Zarsky, 2014) explores broadly the issue of discrimination in the digital society, where not only things, but also people, are "scored." Two articles focus on the issue of the accountability of algorithmic processes, and the legal aspects surrounding their governance (Kross et al., 2017; Schubert & Hutt, 2019), while another two are particularly focused on the legal aspects of automated prediction processes (Zarsky, 2013; Zarsky, 2017). Finally, the article by Goldman (2008) is a commentary on commercial search engines such as Google, and argues against their central regulation.

Three more articles fall into the domain of ethics, without focusing on particular aspects of algorithmic systems. While one (Holzapfel et al., 2018) addresses a specific system - information retrieval (i.e., search engines) for music - it explores the ethical issues associated with storing representations of music, as well as retrieving results for users with varied musical tastes. The article by Sazena et al. (2019) examines people's perception of fairness as a concept, highlighting the diversity of perspectives amongst potential system users. Finally, (Raji & Buolamwini, 2019) describe their experiences in auditing commercial computer vision algorithms for racial biases. Their account is not so much about the process of auditing the algorithms, but rather, the positive impact that they were able to have on industry stakeholders.

Systems articles

The articles collected that describe particular problems within particular algorithmic systems fall into the following types:

- Computer vision (14 articles)
- Medical and healthcare applications (4 articles)
- Crowdsourcing and human computation platforms (3 articles)
- Social media (2)

5.5.1 Problems examined / solutions proposed

Data

Article	Summary	Solution(s)
(Misra et al., 2016)	The authors take the perspective that all human-generate image labels suffer from reporting bias (information dimension).	They propose a method that corrects for "human" noise and maps human data to the ground truth. (fairness learning)
(Tommasi et al., 2017)	The authors examine 12 training datasets for image classification (information dimension).	They carry out analyses to compare the datasets, detecting biases. (discrimination discovery)
(Hendricks et al., 2018)	The authors recognize the disproportionate use of gender-words in image captioning datasets.	They propose a method to equalize the distributions of gender-words used in training. (fairness sampling)

Table 22: Data-based problems in Computer Vision articles.

Article(s)	Summary	Solution(s)
(Otterbacher, 2015)	The author studied linguistic biases in crowd-sourced biographies of Black/White actors/actresses at IMDb.com. (gender, race dimensions)	(discrimination discovery)
(Otterbacher, 2018)	The author conducted experiments at Mechanical Turk, in which workers described images of diverse individuals depicted in images of professions (police, firefighter, bartender). Linguistic biases were detected as a function of the race/gender of the depicted person. (gender, race dimensions)	(discrimination discovery)
(Otterbacher, 2019)	At FigureEight, workers were asked to label highly uniform	(discrimination discovery)

	headsets of women/men across four racial groups. Significant linguistic differences (including ethnicity marking) were discovered in the resulting descriptions. (gender, race dimensions)	

Table 23: Data-based problems in Human Computation & Crowdsourcing articles.

Model

Article(s)	**Summary**	**Solution(s)**
(Bojarski et al., 2016) (Selvaraju et al., 2017) (Simonyan et al., 2013) (Zintgraf et al., 2017)	The problem concerns the lack of transparency of the classification model (information dimension)	Authors proposed a visualization technique for studying the CNN's behaviours. (black box explainability)
(Xu et al., 2016)	The problem concerns the lack of transparency of the caption-generation model (information dimension)	Authors proposed a visualization technique for studying the CNN's behaviours. (black box explainability)
(Zhou et al., 2016)	The problem concerns the lack of transparency of the classification model (information dimension)	Authors propose a method to learn a set of explainable features from the set of deep feature representation. (black box explainability)
(Montavon et al., 2002)	The problem concerns the lack of transparency of the classification model (information dimension)	Authors propose a decomposition method for promoting explainability. (black box explainability)
(Fong & Bedaldi, 2017)	The problem concerns the lack of transparency of the classification model (information dimension)	Authors propose a general, model-agnostic framework for learning explanations. (black box explainability)
(Wang et al., 2018)	The authors detected gender stereotyped descriptions of images (gender dimension)	They proposed the adversarial removal of gender from deep image representations (fairness learning)

(Buolamwini & Gebru, 2018)	The researchers detected significant differences in accuracy on tasks like gender classification from images, across racial / gender groups. (race & gender dimensions)	(discrimination discovery)
(Kyriakou et al., 2019)	In proprietary image tagging algorithms, the authors found systematic differences in descriptions across racial / gender groups. (race & gender dimensions)	(discrimination discovery)

Table 24: Model-based problems in Computer Vision articles.

Article	**Summary**	**Solution(s)**
Gjoka et al. (2010)	The authors examine the problem of how to identify a balanced, unbiased set of users on Facebook, in the context of research studies. (information dimension)	They propose a fair sampling technique (fairness sampling).
Chen and colleagues (2016)	In a study on fairness in pricing, Chen and colleagues (2016) detect the presence and study the behaviours of dynamic pricing algorithms at Amazon.com. (information dimension)	Discrimination discovery

Table 25: Model-based problems in Social Media articles.

Article	**Summary**	**Solution(s)**
(Gibbons et al., 2013)	The authors studied a DSS for diagnosing depression, based on symptoms and medical history, in a multivariate model. (information dimension)	A method for making the resulting decisions more interpretable for the clinician was proposed. (White box explainability)
(Haufe et al., 2014)	The authors studied a high-dimension model used in neuroimaging. (information	A general method to enable easier interpretation of multivariate models was

	dimension)	proposed. (White box explainability)
(Mac Namee et al., 2002)	A regression model learned via ANNs was studied, which was used to predict medical outcomes. Significant training data biases were found. (information dimension)	(Discrimination discovery)
(Obermeyer & Mullainanthan, 2019)	The researchers studied a DSS used in the US for determining whether or not a given patient should be enrolled in a healthcare management program with extra benefits. Significant racial disparities were discovered. (race dimension)	(Discrimination discovery)

Table 26: Model-based problems in Medical / Health-related DSS articles.

5.5.2 Diversity dimensions in the literature of other communities

In the “other” articles, we observe only three diversity dimensions discussed: gender, race, and information (with gender and race often explored together).

Gender / Race

Amongst the computer vision articles, four explore gender and/or race. Hendricks et al. (2018) and Wang et al. (2018) find gender-based biases in image captioning and classification, respectively. Buolamwini and Gebru (2018) and Kyriakou et al. (2019) explore gender- and race-based biases in the output of computer vision algorithms when processing people images.

The three articles related to human computation and crowdsourcing address gender and race-based biases. Otterbacher (2015) considers the language used in biographies of White/Black men and women actors at the IMDb (crowd-generated data). In contrast, the other works by Otterbacher (2018, 2019) consider the generation of such biases within crowdsourced descriptions of people images.

One article related to health-related decisions examines race as a diversity dimension. Obermeyer and Mullainathan (2019) consider racial biases in a decision support system for referring patients to a specialized health management program.

Information

The remaining articles describe information as the diversity dimension of interest. For instance, Gjoka et al. (2010) and Chen et al. (2016) consider information available to researchers and end users, at Facebook and Amazon, respectively. In the *health-related domain*, three articles

(Gibbons et al., 2013; Haufe et al., 2014; Mac Namee et al. 2002) consider the interpretability of information provided to the user of a DSS. Similarly, several articles collected on *computer vision* (Bojarski et al., 2016; Fong & Bedaldi, 2017; Montavon et al., 2002; Selvaraju et al., 2017; Simonyan et al., 2013; Xu et al., 2015, Zhou et al., 2016; Zintgraf et al., 2017) consider the quality (interpretability) of information generated by the respective models. Finally, two other papers (Misra et al. 2016; Tommasi et al., 2017) consider the quality of information provided in datasets for training computer vision algorithms.

5.5.3 Summary - other communities

While we should be careful about making conclusions concerning the state-of-the-art in research communities beyond the ones that we have systematically reviewed, we can note a few trends in the "other" communities. For instance, most of the "other" articles we examined focused on problems with the system's algorithmic model. To this end, researchers in computer vision and health-related DSS are looking into solutions including how to detect problematic behaviours (i.e., discrimination discveroy) as well as how to make a model's behaviours more understandable and interpretable by the user. Most (but not all) of the "other" articles collected examine information as a diversity dimension, and are thus not central to the heart of CyCAT's focus on social and cultural biases in algorithmic systems.

Finally, another interesting observation / comment is with respect to the Human Computation and Crowdsourcing community. Three articles in our collection consider racial and gender biases in datasets collected from user-generated content, as well as via paid micro-tasking crowdwork platforms. We anticipate that work in this area grow, as there are many new initiatives taking place recently (e.g., HCOMP's 2019 call for a FAT* track, the 1st Symposium on Biases in Human Computation and Crowdwork, etc.)

6. Conclusion

We reviewed 245+ articles published in key publication venues across five domains within the larger scientific area of the information and computer sciences, with the aim of understanding the ways in which researchers are addressing problems of algorithmic system bias. We learned that the state-of-the-art research reports a number of different approaches. Therefore, we have described our current work in D3.1 as a being a literature survey on promoting Fairness, Accountability and Transparency ("FAT") in algorithmic systems. In the near future, we aim to produce a comprehensive survey article (D3.4) that will incorporate all of these approaches into one integrated framework. Furthermore, the current deliverable will guide the work undertaken in WPs 4 and 5. In WP4, our integrated framework, based on the insights produced in WP3 (and in particular D3.1 and D3.3), shall be "translated" into guides for particular stakeholders, including end users, developers and teachers. We shall also sketch out "solutions on paper" for promoting FAT in algorithmic systems for information access. In WP5, we aim to implement one such solution (i.e., technical intervention) and to evaluate it.

In addition to examining the problem and solution spaces of FAT in the current deliverable, we have argued that there is a need to consider the *diversity dimensions* addressed in the research. As previously explained, diversity dimensions are the social, cultural and information dimensions upon which a given system's behaviours may vary - often in problematic and/or discriminatory ways. Given the lack of an explicit discussion on the issue of diversity (e.g., through the data and information that informs the development of systems, but also in terms of our interactions with one another in a global, highly-networked world) in the FAT literature to date and its relationship to algorithmic system biases, we argue that a diversity lens brings a fresh perspective to FAT research.[7]

In this final section, we summarize the trends identified through the current literature review (i.e., the problems addressed and the solutions proposed), as well as the diversity dimensions that are being addressed across domains. Finally, we articulate directions for future work and in particular, describe our goals in moving toward publishing a comprehensive survey paper in M18 (D3.4).

6.1 FAT approaches for preventing / mitigating algorithmic system bias across domains

Table 27 correlates the problematic components of algorithmic systems, as reported in our collection of articles, with their proposed solutions, while Table 28 details the solutions that have been proposed and applied by researchers across the five domains we examined.

[7] Again, please see D3.3 for a full description of our conceptual framework based on the concept of diversity and perspective taking.

	Problem(s) addressed: Algorithmic System Component(s)					
	Input	**Data**	**Third Party**	**Model**	**Fairness**	**Output**
Auditing	+			+		+
Explainability Management						
-Black-box	+			+		+
-White-box	+	+	+	+		+
-Model explanation	+			+		+
-Outcome explanation	+			+		+
Discrimination Discovery						
-Explicit	+	+		+		+
-Implicit	+	+		+		+
Fairness Management						
-Fairness sampling		+			+	
-Fairness learning		+	+	+	+	
-Fairness certification	+				+	+

Table 27: FAT tools by their respective algorithmic system components.

Several conclusions can be drawn by synthesizing the correlations between the problems and solutions across the domains we examined:

- *Discrimination Discovery*
  Discrimination Discovery approaches are used across all domains, and can be applied to the study of particular tasks and/or algorithms (e.g., a top-k ranking algorithm) as well as to deployed systems, which may consist of a whole collection of algorithmic processes (e.g., a proprietary search engine, which uses not only relevance ranking, but also personalization / localization algorithms, among others). Obviously, the former case is more commonly addressed in the ML literature, while the latter case is more often discussed in domains such as IR and RecSys. In sum, Discrimination Discovery consists of *tools and practices* for detecting unfair treatment by data / algorithms / systems. Furthermore, these tools can involve the input, data, model, and output, of a system.

	Research Domains				
	ML	**IR**	**RecSys**	**HCI**	**Other**
Auditing	+	+	+	+	+
Explainability Management					
-Black-box	+		+		+
-White-box	+				+
-Model explanation	+				+
-Outcome explanation	+		+		+
Discrimination Discovery					
-Explicit	+	+	+	+	+
-Implicit	+	+	+	+	+
Fairness Management					
-Fairness sampling	+	+	+	+	+
-Fairness learning	+	+	+		+
-Fairness certification	+				

Table 28: FAT tools used across the five research domains examined in the literature review.

- *Auditing*
  Like Discrimination Discovery, references to "Auditing" appear in the literature across all five domains. However, the term is used in different ways and the review provides evidence that *clarification is needed* surrounding auditing - e.g., which actor(s) perform auditing and through which means. In most of the relevant articles in our repository, auditing processes concern the model, as well as the system's inputs and outputs. As previously described, auditing can involve making cross-system or within-system comparisons, and is typically done by an analyst / observer who does not have access to the inner-workings of the system (Sandvig et al., 2014). However, it should be pointed out that auditing uses the *tools* of Discrimination Discovery (or at least, those available to the particular analyst). In this sense, auditing as a term seems to refer to who is doing the discrimination discovery and why; it does not necessarily refer to a different set of tools and techniques. Finally, it should be noted that within Machine Learning, beyond

involving the model, inputs and outputs, auditing can also involve pre-processing (i.e., data-focused) techniques, such as the generation of biased datasets for conducting a black-box audit, e.g., (Cardoso et al., 2019).

- *Fairness Management*
  The issue of ensuring that people and/or groups of people are treated fairly by an algorithm or algorithmic system was found to be of interest to researchers across all domains we considered. However, the tools they have at their disposal vary. For instance, those working "inside the box" (i.e., those involved in the development of a system or algorithm) may take pre-processing measures (i.e., ensure fair sampling when building training datasets) or during-processing measures (i.e., introduce fairness constraints within the learning process). Developers also have methods to "certify" that their algorithms are fair, using internal processes. On the other hand, the HCI literature, typically describing system observers from the outside (i.e., those who study, but who are not involved in the system's development) presents a challenge to our initial taxonomy of solutions (Figure 2). This is because HCI studies often concern the user's *perceptions* of a system's behaviours and/or decisions (e.g., Grgic-Hlaca et al., 2018; Lee 2018), which can be difficult to measure and for which there are no established standards or techniques.

- *Explainability Management*
  Research articles focused on explainability management were primarily found in the domains of Machine Learning and RecSys, as well as in the "Other" category. While producing models and/or outcomes that are easily interpretable to the user is, in and of itself, viewed as a positive characteristic (Gunning, 2017), it is important to emphasize the particular role of *explainability management* within FAT. Specifically, in the FAT 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.

6.2 Diversity dimensions and algorithmic system bias across domains

Figure 3 analyzes the frequency with which the diversity dimensions were examined in the literature across domains in our repository of articles (as of September 2019). The information dimension has clearly been the most studied dimension in the FAT literature thus far. As mentioned, information is the primary dimension addressed in the ML literature (in particular, with respect to *explainability*). Likewise, IR articles often consider information as the diversity 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 do not often address information as a diversity dimension. In these fields, FAT-related articles more often consider social and cultural dimensions.

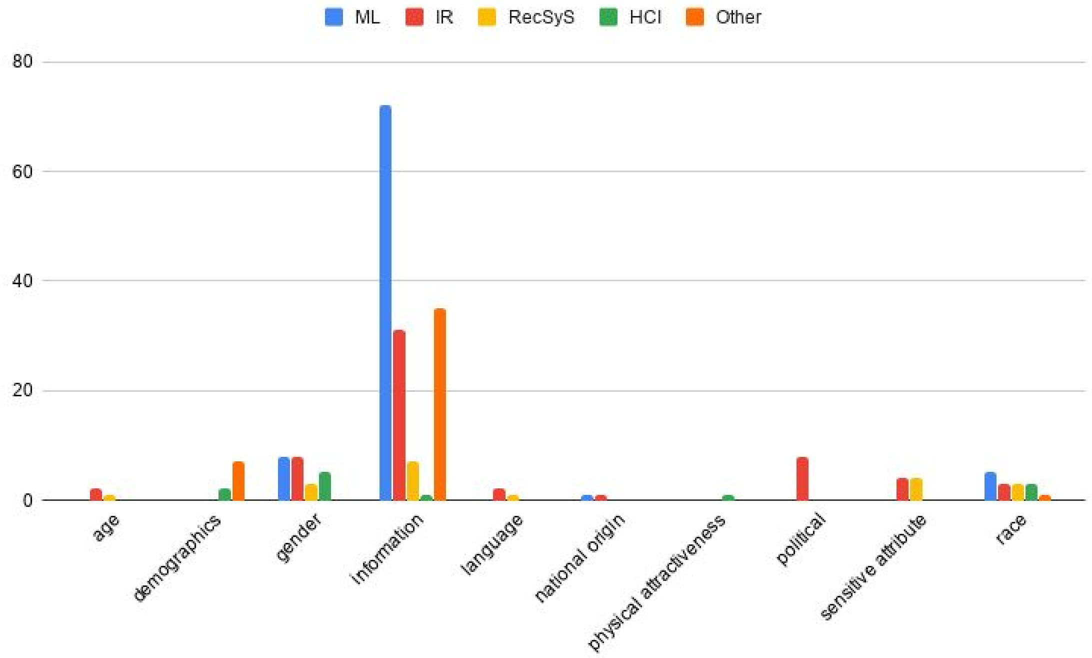

Figure 3: Diversity dimensions explored in research across domains.

6.3 Future work and goals for the survey paper (M18)

The final goal for WP3 is to produce a comprehensive survey article, which shall provide a holistic framework for promoting Fairness, Accountability and Transparency (FAT) in algorithmic systems. From the analysis thus far, it has become clear that there is a need to refine our notions of the problem and solution spaces, as well as to define the various roles of individual stakeholders. As described in Section 6.1 and as depicted in Figure 4 below, 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 promoting and assuring FAT in algorithmic systems. Thus, these roles must be outlined in our framework, and the relationships between them must be specified.

A second consideration to be explored, which was alluded to in Section 6.1, is that while many of the FAT processes described in the literature have been formalized (e.g., discrimination detection methods, internal certification procedures), there are many other issues surrounding *perceived fairness*. The perceived fairness of the user is somewhat subjective and it is not yet clear how formal FAT processes relate to users' perceptions of the systems and their value judgements.

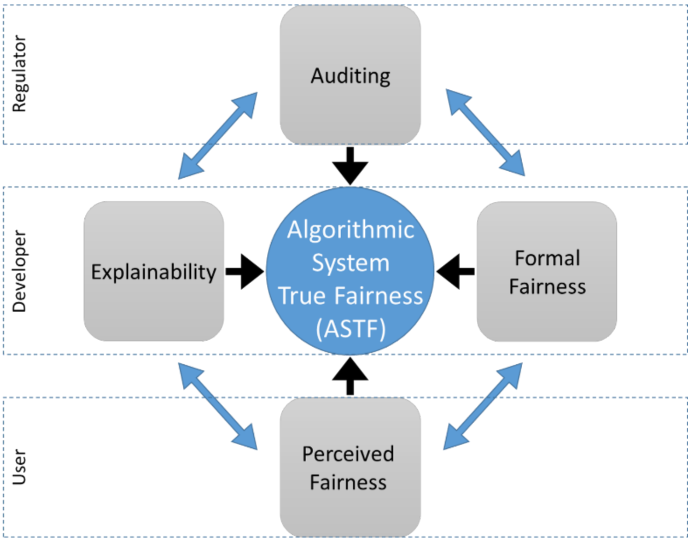

Figure 4: Defining processes and stakeholder roles in promoting FAT algorithmic systems.

Finally, having defined the stakeholders and their roles, and having untangled the relationships between the formal and informal notions of fairness, our framework will need to specify the flow (ordering) and processes between all of the solutions in the FAT toolbox. Furthermore, within the deliverables of WP4, we shall outline the core concepts for target user groups of algorithmic systems (D4.1, D4.2), enabling them to be educated and involved stakeholders in promoting FAT.

7. References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems. In *Human and Machine Learning* (p. pp 21-35). Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-319-90403-0_2

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *ArXiv:1904.02095 [Cs]*. Retrieved from http://arxiv.org/abs/1904.02095

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *Proceedings 2018 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society. https://doi.org/10.14722/ndss.2018.23191

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters*, *35*(2), 131–150. https://doi.org/10.1007/s11063-011-9207-8

Ayre, L., & Craner, J. (2018). Algorithms: avoiding the implementation of institutional biases. *Public Library Quarterly*, *37*(3), 341–347. https://doi.org/10.1080/01616846.2018.1512811

Badjatiya, P., Gupta, M., & Varma, V. (2019). Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-based Generalizations. In *The World Wide Web Conference* (pp. 49–59). New York, NY, USA: ACM. https://doi.org/10.1145/3308558.3313504

Baeza-Yates, R. (2018, June). Bias on the Web. *Communications of the ACM*, *61*(6), pp 54-61. Retrieved from 10.1145/3209581

Bar‐Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (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), 135–149. https://doi.org/10.1002/asi.20941

Barlas, P., Kleanthous, S., Kyriakou, K., & Otterbacher, J. (2019). What Makes an Image Tagger Fair? In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 95–103). New York, NY, USA: ACM. https://doi.org/10.1145/3320435.3320442

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact, *104*(3), 671–732. http://dx.doi.org/10.15779/Z38BG31

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

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., … Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *ArXiv:1810.01943 [Cs]*. Retrieved from http://arxiv.org/abs/1810.01943

Bellogín, A., Castells, P., & Cantador, I. (2017). Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Inf. Retr.*, *20*(6), 606–634. https://doi.org/10.1007/s10791-017-9312-z

Benthall, S., & Haynes, B. D. (2019). Racial categories in machine learning. In *FAT* '19 Proceedings of the Conference on Fairness, Accountability, and Transparency* (p. pp 289-298). Atlanta, GA, USA: ACM New York, NY, USA. https://doi.org/10.1145/3287560.3287575

Binnis, R., Van Kleek, M., Michael, V., Lyngs, U., Zhao, J., & Shadbolt, N. (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

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., & Zieba, K. (2016). Visualbackprop: visualizing cnns for autonomous driving. *ArXiv:1611.05418 [Cs]*. Retrieved from http://arxiv.org/abs/1611.05418

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *In Advances in neural information processing systems* (pp. 4349–4357). Barcelona, Spain: Curran Associates Inc. , USA ©2016.

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., & Hauff, C. (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* (pp. 150–159). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287583

Boz, O. (2002). Extracting Decision Trees from Trained Neural Networks. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 456–461). New York, NY, USA: ACM. https://doi.org/10.1145/775047.775113

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Springer Netherlands*, *15*(3), pp 209-227. https://doi.org/10.1007/s10676-013-9321-6

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (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* (pp. 41:1–41:12). New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300271

Buckley, C. E., Dimmick, D. L., Soboroff, I. M., & Voorhees, E. M. (2007). Bias and the Limits of Pooling for Large Collections | NIST. *Information Retrieval*. Retrieved from https://www.nist.gov/publications/bias-and-limits-pooling-large-collections

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77–91). Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

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

Cañamares, R., & Castells, P. (2017). A Probabilistic Reformulation of Memory-Based Collaborative Filtering: Implications on Popularity Biases. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 215–224). New York, NY, USA: ACM. https://doi.org/10.1145/3077136.3080836

Card, D., Zhang, M., & Smith, N. A. (2019). Deep Weighted Averaging Classifiers (pp. 369–378). Presented at the FAT*Fairness, Accountability, and Transparency, Atlanta, GA, USA: ACM New York, NY, USA ©2019. https://doi.org/10.1145/3287560.3287595

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (p. Pp. 319-328). Presented at the FAT*Fairness, Accountability, and Transparency, Atlanta, GA, USA: ACM New York, NY, USA ©2019. https://doi.org/10.1145/3287560.3287586

Celis, L. E., Kapoor, S., Salehi, F., & Vishnoi, N. (2019). Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 160–169). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287601

Chakraborty, A., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2019). Optimizing the recency-relevance-diversity trade-offs in non-personalized news recommendations. *Information Retrieval Journal*. https://doi.org/10.1007/s10791-019-09351-2

Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In *Eleventh International AAAI Conference on Web and Social Media* (pp. 22–31). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15680

Chakraborty, A., Patro, G. K., Ganguly, N., Gummadi, K. P., & Loiseau, P. (2019). Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 129–138). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287570

Chen, L., Ma, R., Hannák, A., & Wilson, C. (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* (pp. 651:1–651:14). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3174225

Chen, L., Mislove, A., & Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide*

Web (pp. 1339–1349). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883089

Chen, T.-W., & Sundar, S. S. (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* (pp. 537:1–537:13). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3174111

Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R. C., & Kankanhalli, M. (2019). MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Trans. Inf. Syst.*, *37*(2), 16:1–16:28. https://doi.org/10.1145/3291060

Chipman, A., George, E. I., E.F, R., & McCullochDepartment. (2007). Making sense of a forest of treesH.

Cho, J., & Roy, S. (2004). Impact of Search Engines on Page Popularity. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 20–29). New York, NY, USA: ACM. https://doi.org/10.1145/988672.988676

Cho, J., Roy, S., & Adams, R. E. (2005). Page Quality: In Search of an Unbiased Web Ranking. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 551–562). New York, NY, USA: ACM. https://doi.org/10.1145/1066157.1066220

Cowgill, B., & Tucker, C. (2017). *Algorithmic bias: A counterfactual perspective* (p. 3). Working Paper: NSF Trustworthy Algorithms.

Craven, M., & Shavlik, J. W. (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* (pp. 37–45). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=3091574.3091580

Craven, M., & Shavlik, J. W. (1996). Extracting Tree-Structured Representations of Trained Networks. In *Advances in Neural Information Processing Systems 8* (pp. 24–30). MIT Press. Retrieved from http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

Das, M., Hecht, B., & Gergle, D. (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* (pp. 563:1–563:14). New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300793

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

Datta, Amit, Datta, A., Makagon, J., Mulligan, D. K., & Tschantz, M. C. (2018). Discrimination in Online Personalization: A Multidisciplinary Inquiry. In *Proceedings of Machine Learning Research* (p. 16).

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Eleventh International AAAI Conference on Web and Social Media* (pp. 512–515). Montreal, Canada. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

Deng, H. (2014). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, *7*(4), 277–287. https://doi.org/10.1007/s41060-018-0144-8

Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 412:1–412:14). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3173986

Dimitrakakis, C., Liu, Y., Parkes, D., & Radanovic, G. (2018). Bayesian Fairness. Presented at the Thirty-Third AAAI Conference on Artificial Intelligence. Retrieved from https://hal.inria.fr/hal-01953311

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67–73). New York, NY, USA: ACM. https://doi.org/10.1145/3278721.3278729

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

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22. https://doi.org/10.1257/app.20160213

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces* (pp. 211–223). New York, NY, USA: ACM. https://doi.org/10.1145/3172944.3172961

Eickhoff, C. (2018). Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 162–170). New York, NY, USA: ACM. https://doi.org/10.1145/3159652.3159654

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., & Pera, M. S. (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* (pp. 172–186). Retrieved from http://proceedings.mlr.press/v81/ekstrand18b.html

Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017). Suppressing the Search Engine Manipulation Effect (SEME). *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 42:1–42:22. https://doi.org/10.1145/3134677

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 432:1–432:13). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3174006

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (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* (pp. 62–71). AAAI Press. Retrieved from https://experts.illinois.edu/en/publications/be-careful-things-can-be-worse-than-they-appear-understanding-bia

Eslami, M., Vaccaro, K., Lee, M. K., Elazari Bar On, A., Gilbert, E., & Karahalios, K. (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* (pp. 494:1–494:14). New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300724

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). New York, NY, USA: ACM. https://doi.org/10.1145/2783258.2783311

Fong, R., & Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 3449–3457). https://doi.org/10.1109/ICCV.2017.371

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A Comparative Study of Fairness-enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287589

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Trans. Inf. Syst.*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Fung, G., Sandilya, S., & Rao, R. B. (2005). Rule extraction from linear support vector machines. In *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 32–40).

Germano, F., Gómez, V., & Mens, G. L. (2019). *The few-get-richer: a surprising consequence of popularity-based rankings* (Economics Working Paper). Department of Economics and Business, Universitat Pompeu Fabra. Retrieved from https://econpapers.repec.org/paper/upfupfgen/1636.htm

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., … Kupfer, D. J. (2013). The CAD-MDD: A Computerized Adaptive Diagnostic Screening Tool for Depression. *The Journal of Clinical Psychiatry*, *74*(7), 669–674. https://doi.org/10.4088/JCP.12m08338

Gjoka, M., Kurant, M., T. Butts, C., & Markopoulou, A. (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs (pp. 1–9). Presented at the INFOCOM, 2010 Proceedings IEEE. https://doi.org/10.1109/INFCOM.2010.5462078

Goh, G., Cotter, A., Gupta, M., & Friedlander, M. P. (2016). Satisfying Real-world Goals with Dataset Constraints. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 2415–2423). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/6316-satisfying-real-world-goals-with-dataset-constraints.pdf

Goldman, E. (2008). Search Engine Bias and the Demise of Search Engine Utopianism. In A. Spink & M. Zimmer (Eds.), *Web Search: Multidisciplinary Perspectives* (pp. 121–133). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75829-7_8

Green, B., & Chen, Y. (2019). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). New York, NY, USA: ACM. https://doi.org/10.1145/3287560.3287563

Grgic-Hlacˇa, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (n.d.). The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS 2016 - MACHINE LEARNING AND THE LAW SYMPOSIUM* (Vol. 1, p. 11). Barcelona, Spain.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (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* (pp. 903–912). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186138

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *In Thirty-Second AAAI Conference on Artificial Intelligence.*

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5), 93:1–93:42. https://doi.org/10.1145/3236009

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). *Local Rule-Based Explanations of Black Box Decision Systems*. Retrieved from http://arxiv.org/abs/1805.10820

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (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* (pp. 1914–1933). New York, NY, USA: ACM. https://doi.org/10.1145/2998181.2998327

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*. Retrieved from http://arxiv.org/abs/1610.02413

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Heindorf, S., Scholten, Y., Engels, G., & Potthast, M. (2019). Debiasing Vandalism Detection Models at Wikidata. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. https://doi.org/10.1145/3308558.3313507

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018* (pp. 793–811). Springer International Publishing.

Henelius, A., Puolamaki, K., Bostrom, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery*, *28*(5–6), 1503–1529. https://doi.org/10.1007/s10618-014-0368-8

Herdağdelen, A., & Baroni, M. (2011). Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, *62*(9), 1741–1749. https://doi.org/10.1002/asi.21579

Hofmann, K., Mitra, B., Radlinski, F., & Shokouhi, M. (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* (pp. 549–558). New York, NY, USA: ACM. https://doi.org/10.1145/2661829.2661922

Holstein, K., Wortman Vaughan, J., Daumé, H., III, Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 600:1–600:16). New York, NY, USA: ACM. https://doi.org/10.1145/3290605.3300830

Holzapfel, A., Sturm, B. L., & Coeckelbergh, M. (2018). Ethical Dimensions of Music Information Retrieval Technology. *Transactions of the International Society for Music Information Retrieval*, *1*(1), 44–55. https://doi.org/10.5334/tismir.13

Hjørland, B. (2002). Domain analysis in information science: eleven approaches–traditional as well as innovative. Journal of documentation, 58(4), 422-462.

Horne, B. D., Nevo, D., O'Donovan, J., Cho, J.-H., & Adali, S. (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* (Vol. 13, pp. 247–256).

Hu, D., Jiang, S., E. Robertson, R., & Wilson, C. (2019). Auditing the Partisanship of Google Search Snippets. In *The World Wide Web Conference* (pp. 693–704). New York, NY, USA: ACM. https://doi.org/10.1145/3308558.3313654

Hu, Z., Wang, Y., Peng, Q., & Li, H. (2018). Unbiased LambdaMART: An Unbiased Pairwise Learning-to-Rank Algorithm. In *In The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313447

Jahna Otterbacher. (2018). Social Cues, Social Biases: Stereotypes in Annotations on People Images. In *The Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018)*. 2275 East Bayshore Road, Suite 160 Palo Alto, California 94303. Retrieved from https://zenodo.org/record/2670019#.XRsgMOgzZPY

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

Jiang, S., Robertson, R. E., & Wilson, C. (2019). Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 278–289). Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/3229

Jiang, X., Sun, X., Yang, Z., Zhuge, H., & Yao, J. (2016). Exploiting heterogeneous scientific literature networks to combat ranking bias: Evidence from the computational linguistics area. *Journal of the Association for Information Science and Technology*, *67*(7), 1679–1702. https://doi.org/10.1002/asi.23463

Johansson, U., & Niklasson, L. (2009). Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 238–244). Nashville, TN, USA: IEEE. https://doi.org/10.1109/CIDM.2009.4938655

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

Johnson, I., McMahon, C., Schöning, J., & Hecht, B. (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* (pp. 1167–1178). New York, NY, USA: ACM. https://doi.org/10.1145/3025453.3026015

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). Springer Berlin Heidelberg.

Karako, C., & Manggala, P. (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* (pp. 23–28). New York, NY, USA: ACM. https://doi.org/10.1145/3213586.3226206

Kay, M., Matuszek, C., & Munson, S. A. (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* (pp. 3819–3828). New York, NY, USA: ACM. https://doi.org/10.1145/2702123.2702520

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

Khademi, A., Lee, S., Foley, D., & Honavar, V. (2019). Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. *The World Wide Web Conference on - WWW ’19*, 2907–2914. https://doi.org/10.1145/3308558.3313559

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., & Weller, A. (2018). Blind Justice: Fairness with Encrypted Sensitive Attributes. *ArXiv:1806.03281 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1806.03281

Kim, B., Rudin, C., & Shah, J. A. (2014). The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In *Advances in Neural Information Processing Systems 27* (pp. 1952–1960). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5313-the-bayesian-case-model-a-generative-approach-for-case-based-reasoning-and-prototype-classification.pdf

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017*. Retrieved from http://arxiv.org/abs/1609.05807

Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference* (pp. 121–127). New York, NY, USA: ACM. https://doi.org/10.1145/2815675.2815714

Kodama, C., Jean, B. S., Subramaniam, M., & Taylor, N. G. (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), 403–432. https://doi.org/10.1007/s10791-017-9306-x

Kokkodis, M. (2019). Reputation Deflation Through Dynamic Expertise Assessment in Online Labor Markets. In *The World Wide Web Conference* (pp. 896–905). New York, NY, USA: ACM. https://doi.org/10.1145/3308558.3313479

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1), 101–123. https://doi.org/10.1007/s11257-011-9112-x

Koolen, C., & van Cranenburgh, A. (2017). These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 12–22). Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1602

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 379–390). New York, NY, USA: ACM. https://doi.org/10.1145/3301275.3302306

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

Krishnan, S., & Wu, E. (2017). PALM: Machine Learning Explanations For Iterative Debugging. In *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics* (pp. 4:1–4:6). New York, NY, USA: ACM. https://doi.org/10.1145/3077257.3077271

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*, 633–706. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/pnlr165&i=648

Kuhlman, C., VanValkenburg, M., & Rundensteiner, E. (2019). FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (pp. 2936–2942). New York, NY, USA: ACM. https://doi.org/10.1145/3308558.3313443

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (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* (pp. 417–432). New York, NY, USA: ACM. https://doi.org/10.1145/2998181.2998321

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30* (pp. 4066–4076). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf

Kyriakou, K., Barlas, P., Kleanthous, S., & Otterbacher, J. (2019). Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 313–322). Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/3232

L. Cardoso, R., Meira Jr., W., Almeida, V., & J. Zaki, M. (2019). A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 437–444). New York, NY, USA: ACM. https://doi.org/10.1145/3306618.3314262

Landecker, W., Thomure, M. D., Bettencourt, L. M. A., Mitchell, M., Kenyon, G. T., & Brumby, S. P. (2013). Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 32–38). Singapore, Singapore: IEEE. https://doi.org/10.1109/CIDM.2013.6597214

Le, H., Maragh, R., Ekdale, B., High, A., Havens, T., & Shafiq, Z. (2019). Measuring Political Personalization of Google News Search. In *The World Wide Web Conference* (pp. 2957–2963). New York, NY, USA: ACM. https://doi.org/10.1145/3308558.3313682

Leavy, S. (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* (pp. 14–16). New York, NY, USA: ACM. https://doi.org/10.1145/3195570.3195580

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Baykal, S. (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 (pp. 1035–1048). New York, NY, USA: ACM. https://doi.org/10.1145/2998181.2998230

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 107–117). Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1011

Leonhardt, J., Anand, A., & Khosla, M. (2018). User Fairness in Recommender Systems. In *Companion Proceedings of the The Web Conference 2018* (pp. 101–102). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3186949

Lin, Y. L., Trattner, C., Brusilovsky, P., & He, D. (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*, 1785–1798. Retrieved from http://d-scholarship.pitt.edu/25927/

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate Intelligible Models with Pairwise Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623–631). New York, NY, USA: ACM. https://doi.org/10.1145/2487575.2487579

Lu, J., Tokinaga, S., & Ikeda, Y. (2006). Explanatory rule extraction based on the trained neural network and the genetic programming. https://doi.org/10.15807/jorsj.49.66

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*(1), 51–70.

Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., & Saxena, M. (2018). Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Conference on Fairness, Accountability and Transparency* (pp. 92–105). Retrieved from http://proceedings.mlr.press/v81/madaan18a.html

Magno, G., Araújo, C. S., Meira Jr., W., & Almeida, V. (2016). Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. *ArXiv:1609.05413 [Cs]*. Retrieved from http://arxiv.org/abs/1609.05413

Matsangidou, M., & Otterbacher, J. (2019). What Is Beautiful Continues to Be Good. In *Human-Computer Interaction – INTERACT 2019* (pp. 243–264). Springer International Publishing.

Maxwell, D., Azzopardi, L., & Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*, 1–25. https://doi.org/10.1007/s10791-019-09353-0

Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., & Yilmaz, E. (2017). Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 626–633). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054197

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (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* (pp. 2243–2251). New York, NY, USA: ACM. https://doi.org/10.1145/3269206.3272027

Misra, I., Lawrence Zitnick, C., Mitchell, M., & Girshick, R. (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, pp. 2930–2939). https://doi.org/10.1109/CVPR.2016.320

Mitra, B., Shokouhi, M., Radlinski, F., & Hofmann, K. (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* (pp. 1055–1058). Gold Coast, Queensland, Australia: ACM Press. https://doi.org/10.1145/2600428.2609508

Montavon, G., Bach, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring Search Engine Bias. *Inf. Process. Manage.*, *41*(5), 1193–1205. https://doi.org/10.1016/j.ipm.2004.05.005

Nikolov, D., Lalmas, M., Flammini, A., & Menczer, F. (2018). Quantifying Biases in Online Information Exposure. *ArXiv:1807.06958 [Cs]*, *70*(3), 218–229. Retrieved from http://arxiv.org/abs/1807.06958

Niu, S., Lan, Y., Guo, J., Wan, S., & Cheng, X. (2015). Which noise affects algorithm robustness for learning to rank. *Information Retrieval Journal*, *18*(3), 215–245. https://doi.org/10.1007/s10791-015-9253-3

Noble, S. (2013). Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. *InVisible Culture*, *19*. Retrieved from https://urresearch.rochester.edu/institutionalPublicationPublicView.action?institutionalItemId=27584

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

Núñez, H., Angulo, C., & Català, A. (2002). Rule extraction from support vector machines. In *ESANN* (pp. 107–112). Belgium.

Obermeyer, Z., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (pp. 89–89). Atlanta, GA, USA: ACM Press. https://doi.org/10.1145/3287560.3287593

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*(3). https://doi.org/10.3389/fdata.2019.00013

Ortega, J. L., & Aguillo, I. F. (2014). Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, *65*(6), 1149–1156. https://doi.org/10.1002/asi.23036

Otterbacher, J., Bates, J., & Clough, P. D. (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*. Retrieved from https://doi.org/10.1145/3025453.3025727

Otterbacher, Jahna. (2015). Linguistic Bias in Collaboratively Produced Biographies: Crowdsourcing Social Stereotypes? In *ICWSM*.

Otterbacher, Jahna, Checco, A., Demartini, G., & Clough, P. (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* (pp. 933–936). New York, NY, USA: ACM. https://doi.org/10.1145/3209978.3210094

Otterbacher, J., Barlas, P., Kleanthous, S., & Kyriakou, K. (2019). How Do We Talk About Other People? Group (Un)Fairness in Natural Language Image Descriptions. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*. Stevenson, WA USA: AAAI.

Pal, A., Harper, F. M., & Konstan, J. A. (2012). A Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. *ACM Transactions on Information Systems (TOIS)*, *30*(2), 10. https://doi.org/10.1145/0000000.0000000

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Integrating Induction and Deduction for Finding Evidence of Discrimination. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law* (pp. 157–166). New York, NY, USA: ACM. https://doi.org/10.1145/1568234.1568252

Pedreschi, D., Ruggieri, S., & Turini, F. (n.d.). Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics* (pp. 581–592).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30* (pp. 5680–5689). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf

Plumb, G., Molitor, D., & Talwalkar, A. S. (2018). Model Agnostic Supervised Local Explanations. In *Advances in Neural Information Processing Systems 32* (pp. 2515–2524). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7518-model-agnostic-supervised-local-explanations.pdf

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness Reduces Racial Bias. *Management Science*, *64*(11), 4988–4995. https://doi.org/10.1287/mnsc.2017.2901

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., … Anvik, J. (2006). Visual Explanation of Evidence in Additive Classifiers. In *In Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, p. 8).

Quattrone, G., Capra, L., & De Meo, P. (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* (pp. 1021–1032). New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675235

Rader, E., Cotter, K., & Cho, J. (2018). Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 103:1–103:13). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3173677

Rafrafi, A., Guigue, V., & Gallinari, P. (2012). Coping with the Document Frequency Bias in Sentiment Classification. In *Sixth International AAAI Conference on Weblogs and Social Media* (pp. 314–321). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4582

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *AAAI/ACM Conf. on AI Ethics and Society.* (p. 7).

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., … Redmiles, E. M. (2019). On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* ’19*, 140–149. https://doi.org/10.1145/3287560.3287580

Ribeiro, M. T., Singh, S., & Guestrin, C. (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* (pp. 1135–1144). San Francisco, CA, USA: ACM. http://dx.doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High Precision Model-Agnostic Explanations. In *In Thirty-Second AAAI Conference on Artificial Intelligence.* (p. 9).

Robertson, R. E., Friedland, L., JOSEPH, K., Lazer, D., Wilson, C., & Jiang, S. (2018). Auditing Partisan Audience Bias within Google Search. In *Proceedings of the ACM on Human-Computer Interaction* (Vol. 2, pp. 1–22). https://doi.org/10.1145/3274417

Robertson, R. E., Jiang, S., Lazer, D., & Wilson, C. (2019). Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 235–244). New York, NY, USA: ACM. https://doi.org/10.1145/3292522.3326047

Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference* (pp. 955–965). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186143

Rokicki, M., Herder, E., & Trattner, C. (2017). How Editorial, Temporal and Social Biases Affect Online Food Popularity and Appreciation. In *Eleventh International AAAI Conference on Web and Social Media* (pp. 192–200). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15707

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638. https://doi.org/10.1017/S0269888913000039

Rosenblat, A., & Stark, L. (2016). *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers* (SSRN Scholarly Paper No. ID 2686227). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2686227

Rudinger, R., May, C., & Van Durme, B. (2017). Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 74–79). Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1609

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *ArXiv:1811.05577 [Cs]*. Retrieved from http://arxiv.org/abs/1811.05577

Salminen, J., Jung, S.-G., & Jansen, B. J. (2019). Detecting Demographic Bias in Automatically Generated Personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (p. LBW0122:1–LBW0122:6). New York, NY, USA: ACM. https://doi.org/10.1145/3290607.3313034

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (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.

Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., & Liu, Y. (2018). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *AI Ethics and Society (AIES) 2019*. Retrieved from http://arxiv.org/abs/1811.03654

Schetinin, V., Fieldsend, J. E., Partridge, D., Coats, T. J., Krzanowski, W. J., Everson, R. M., … Hernandez, A. (2007). Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications. *IEEE Transactions on Information Technology in Biomedicine*, *11*(3), 312–319. https://doi.org/10.1109/TITB.2006.880553

Schubert, C., & Hütt, M.-T. (2019). Economy-on-demand and the fairness of algorithms. *European Labour Law Journal*, *10*(1), 3–16. https://doi.org/10.1177/2031952519829082

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). Venice: IEEE. https://doi.org/10.1109/ICCV.2017.74

Shandilya, A., Ghosh, K., & Ghosh, S. (2018). Fairness of Extractive Text Summarization. In *Companion Proceedings of the The Web Conference 2018* (pp. 97–98). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3186947

Shen, J. H., Fratamico, L., Rahwan, I., & Rush, A. M. (2018). Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Stockholm, Sweden.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. In *International Conference on Machine Learning* (pp. 3145–3153). Retrieved from http://proceedings.mlr.press/v70/shrikumar17a.html

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv:1312.6034 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6034

Singh, A., & Joachims, T. (2018). Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2219–2228). New York, NY, USA: ACM. https://doi.org/10.1145/3219819.3220088

Singh, S., Ribeiro, M. T., & Guestrin, C. (2016). Programs as Black-Box Explanations. In *arXiv:1611.07579 [cs, stat]*. Barcelona, Spain. Retrieved from http://arxiv.org/abs/1611.07579

Solomon, J. (2014). Customization Bias in Decision Support Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3065–3074). New York, NY, USA: ACM. https://doi.org/10.1145/2556288.2557211

Sonnenburg, S., Zien, A., Philips, P., & Rätsch, G. (2008). POIMs: positional oligomer importance matrices—understanding support vector machine-based signal detectors. *Bioinformatics*, *24*(13), i6–i14. https://doi.org/10.1093/bioinformatics/btn170

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F., Arvanitakis, G., Benevenuto, F., … Mislove, A. (2018). Potential for Discrimination in Online Targeted Advertising. In *FAT 2018 - Conference on Fairness, Accountability, and Transparency* (Vol. 81, pp. 1–15). New-York, United States. Retrieved from https://hal.archives-ouvertes.fr/hal-01955343

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18* (pp. 2239–2248). London, United Kingdom: ACM Press. https://doi.org/10.1145/3219819.3220046

Springer, A. (2019). Making Transparency Clear. In *In Joint Proceedings of the ACM IUI 2019 Workshops* (p. 5). Los Angeles, USA.

Springer, A., & Whittaker, S. (2019). Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 107–120). New York, NY, USA: ACM. https://doi.org/10.1145/3301275.3302322

Stoica, A.-A., Riederer, C., & Chaintreau, A. (2018). Algorithmic Glass Ceiling in Social Networks: The Effects of Social Recommendations on Network Diversity. In *Proceedings of the 2018 World Wide Web Conference* (pp. 923–932). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186140

Strumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, *11*, 1–18. https://doi.org/10.1145/1756006.1756007

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

Tan, H. F., Hooker, G., & Wells, M. T. (2016). Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. *ArXiv:1611.07115 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1611.07115

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2017). Detecting Bias in Black-Box Models Using Transparent Model Distillation. *ArXiv*, *abs/1710.06169*.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 303–310). New Orleans, LA, USA: ACM New York, NY, USA ©2018. https://doi.org/10.1145/3278721.3278725

ter Hoeve, M., Heruer, M., Odijk, D., Schuth, A., Spitters, M., & de Rijke, M. (n.d.). Do News Consumers Want Explanations for Personalized News Rankings? In *In FATREC Workshop on Responsible Recommendation Proceedings.*. Como, Italy. https://doi.org/10.18122/B24D7N

Thelisson, E., Padh, K., & Celis, L. E. (2017). Regulatory Mechanisms and Algorithms towards Trust in AI/ML. In *In Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)* (p. 5). Melbourne, Australia.

Thelwall, M., & Maflahi, N. (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), 1124–1135. https://doi.org/10.1002/asi.23252

Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 465–474). New York, NY, USA: ACM. https://doi.org/10.1145/3097983.3098039

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2017). A Deeper Look at Dataset Bias. In *Domain Adaptation in Computer Vision Applications* (pp. 37–55). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58347-1_2

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

Urbano. (2016). Test Collection Reliability: A Study of Bias and Robustness to Statistical Assumptions via Stochastic Simulation. *Information Retrieval Journal*, *19*(3), 313–350. https://doi.org/10.1007/s10791-015-9274-y

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 205395171774353. https://doi.org/10.1177/2053951717743530

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 440:1–440:14). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3174014

Vidovic, M. M.-C., Görnitz, N., Müller, K.-R., & Kloft, M. (2016). Feature Importance Measure for Non-linear Learning Algorithms. In *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Barcelona, Spain. Retrieved from http://arxiv.org/abs/1611.07567

Vilenchik, D., Yichye, B., & Abutbul, M. (2019). To Interpret or Not to Interpret PCA? This Is Our Question. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 655–658. Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/3265

Vincent, N., Johnson, I., Sheehan, P., & Hecht, B. (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, pp. 505–5016).

Wachs, J., Hannak, A., Vörös, A., & Daróczy, B. (2017). Why Do Men Get More Attention? Exploring Factors Behind Success in An Online Design Community. In *Eleventh International AAAI Conference on Web and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15651

Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018). Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 165–174). New York, NY, USA: ACM. https://doi.org/10.1145/3209978.3210010

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2018). Adversarial Removal of Gender from Deep Image Representations. *ArXiv:1811.08489 [Cs]*. Retrieved from http://arxiv.org/abs/1811.08489

Weber, I., & Castillo, C. (2010). The demographics of web search. In *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 523–530). New York, NY, USA. https://doi.org/10.1145/1835449.1835537

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

White, R. W., & Horvitz, E. (2015). Belief Dynamics and Biases in Web Search. *ACM Transactions on Information Systems*, *33*(4), 1–46. https://doi.org/10.1145/2746229

Wilkie, C., & Azzopardi, L. (2014a). 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 (pp. 81–90). Shanghai, China. https://doi.org/10.1145/2661829.2661948

Wilkie, C., & Azzopardi, L. (2014b). Best and Fairest: An Empirical Analysis of Retrieval System Bias. In *In European Conference on Information Retrieval* (Vol. 8416, pp. 13–25). Springer, Cham. https://doi.org/10.1007/978-3-319-06028-6_2

Wilkie, C., & Azzopardi, L. (2017). Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable? In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2375–2378). Singapore. https://doi.org/10.1145/3132847.3133135

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 656:1–656:14). New York, NY, USA: ACM. https://doi.org/10.1145/3173574.3174230

Wu, Y., Zhang, L., & Wu, X. (2019). On Convexity and Bounds of Fairness-aware Classification. In *The World Wide Web Conference* (pp. 3356–3362). San Francisco, CA, USA: ACM New York, NY, USA ©2019. https://doi.org/10.1145/3308558.3313723

Xiao, L., Min, Z., Yongfeng, Z., Zhaoquan, G., Yiqun, L., & Shaoping, M. (2017). Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 107–115). New York, NY, USA: ACM. https://doi.org/10.1145/3109859.3109887

Xu, K., Ba Lei, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., … Bengio, Y. (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, pp. 2048–2057). Lille, France. Retrieved from https://dl.acm.org/citation.cfm?id=3045336

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

Yu, H.-T., Jatowt, A., Blanco, R., Joho, H., & Jose, J. M. (2017). Decoding multi-click search behavior based on marginal utility. *Kluwer Academic Publishers Hingham, MA, USA*, *20*(1), 25–52. https://doi.org/10.1007/s10791-016-9289-z

Zang, J., Dummit, K., Graves, J., Lisker, P., & Sweeney, and L. (2015). Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps. *Technology Science*. Retrieved from https://techscience.org/a/2015103001/

Zarsky, T. Z. (2013). Transparent Predictions. *University of Illinois Law Review*, *2013*(4), 1503–1570. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2324240

Zarsky, T. Z. (2014). Understanding discrimination in the scored society. *Washington Law Review*, *89*(4), 1375–1412.

Zarsky, T. Z. (2017). An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics. *I/S: A Journal of Law and Policy for the Information Society*, *14*(1), 11–36.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1569–1578). New York, NY, USA: ACM. https://doi.org/10.1145/3132847.3132938

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. In *International Conference on Machine Learning* (pp. 325–333). Retrieved from http://proceedings.mlr.press/v28/zemel13.html

Zhang, J., & Bareinboim, E. (2018). Fairness in Decision-Making — The Causal Explanation Formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949

Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (n.d.). Machine Learning Testing: Survey, Landscapes and Horizons, 35. Retrieved from https://arxiv.org/abs/1906.10742

Zhang, L., & Wu, X. (2017). Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, *4*(1), 1–16. https://doi.org/10.1007/s41060-017-0058-x

Zhang, L., Wu, Y., & Wu, X. (2016). Situation Testing-based Discrimination Discovery: A Causal Inference Approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 2718–2724). New York, New York, USA: AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=3060832.3061001

Zhang, L., Wu, Y., & Wu, X. (2017). Achieving Non-Discrimination in Data Release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1335–1344). New York, NY, USA: ACM. https://doi.org/10.1145/3097983.3098167

Zhang, L., Wu, Y., & Wu, X. (n.d.). A causal framework for discovering and removing direct and indirect discrimination. Retrieved from https://arxiv.org/abs/1611.07509

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *ArXiv:1707.09457 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1707.09457

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (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). Retrieved from http://arxiv.org/abs/1804.06876

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning Gender-Neutral Word Embeddings. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4847–4853). Retrieved from http://arxiv.org/abs/1809.01496

Zhou, B., Khosla, A., Lapedriza, A., Olivia, A., & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. In *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929). Retrieved from https://arxiv.org/abs/1512.04150

Zhou, Z.-H., Jiang, Y., & Chen, S.-F. (2003). Extracting symbolic rules from trained neural network ensembles. *AI Communications - Artificial Intelligence Advances in China*, *16*(1), 3–15. Retrieved from https://dl.acm.org/citation.cfm?id=1218644

Zien, A., Kramer, N., Sonnenburg, S., & Ratsch, G. (2009). The Feature Importance Ranking Measure. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases* (Vol. 2, pp. 694–709). Bled, Slovenia. Retrieved from https://dl.acm.org/citation.cfm?id=3121646.3121692

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis (p. 12). Presented at the ICLR 2017. Retrieved from https://arxiv.org/abs/1702.04595

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *ArXiv:1511.00148 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1511.00148

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

Annex: Metadata used in Zotero CyCAT Survey Collection.

DOMAIN (each paper belongs to one of five categories)

Machine_Learning
Rec_Sys
HCI
IR
Other

PROBLEM (each paper has 0+ “problem” tags)

Input
Data
Output
Model
Human
Fairness

DIVERSITY DIMENSION (each paper has 0+ “diversity” tags)
Note: More tags may be added as time goes on. To date, we have the following:

Age
Altruistic
Attractiveness
Cognitive
Country
Consumers
Cultural
Demographics
Emotion
Ethnicity
Gender
Information
Labor
Linguistic
National_origin
Opinion
Personality
Political
Pricing
Protected_minority (attributes and classes of people)
Sensitive attributes
Sensitive features
Sensitive Groups
Socioeconomic
Race
Risk
Utility

SOLUTIONS (each paper has 0+ “solutions” tags)

Auditability
Discrimination_discovery_direct
Discrimination_discovery_indirect
Explainability
Explainability_blackbox
Explainability_whitebox
Explainability_UX
Fairness (includes fairness perception)
Fairness_certification
Fairness_learning
Fairness_sampling
Other