
| | |
|---|---|
| **Document Title** | **Paper-based data-side solutions** |
| **Project Title and acronym** | Cyprus Center for Algorithmic Transparency (CyCAT) |
| **H2020-WIDESPREAD-05-2017-Twinning** | Grant Agreement number: 810105 — CyCAT |
| **Deliverable No.** | D4.3 |
| **Work package No.** | WP4 |
| **Work package title** | Promoting algorithmic transparency |
| **Authors (Name and Partner Institution)** | Nandu Chandran Nair (UNITN)<br>Fausto Giunchiglia (UNITN) |
| **Contributors**<br>**(Name and Partner Institution)** | Kalia Orphanou (OUC)<br>Jahna Otterbacher (OUC) |
| **Reviewers** | Frank Hopfgartner (USFD)<br>Michael Rovatsos (UEDIN) |
| **Status**<br>**(D: draft; RD: revised draft; F: final)** | F |
| **File Name** | D4.3_Data_Side_Solutions_M18 |
| **Date** | 31 March 2020 |

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 | 5/10/19 | Jahna Otterbacher | jahna.otterbacher@ouc.ac.cy | Document created |
| V2.0 | 9/12/19 | Nandu Chandran Nair | nandu.chandrannair@unitn.it | Document modified |
| V3.0 | 31/01/20 | Nandu Chandran Nair | nandu.chandrannair@unitn.it | Submitted for feedbacks |
| V4.0 | 3/3/20 | Nandu Chandran Nair | nandu.chandrannair@unitn.it | Document revised with feedbacks |

| **Abstract** | |
|---|---|
| D4.3 elaborates on the data-focused problems identified in D4.1. It provides documentation of various "solutions on paper" that focus on the issue of training data, and how that data is captured/ generated/annotated, in promoting algorithmic transparency. | |
| **Keyword(s):** | Algorithmic bias, Data biases, Solutions for data bias, Discrimination discovery, human-labeled data, training data |

# Contents

# 1. Executive Summary

The goal of D4.3 is to present some "paper-based solutions" for addressing the problem of data biases in the development and use of algorithmic information access (IA) systems. To this end, it uses D4.1 as a point of departure.

Section 2 reviews the different types of biases and gives detailed descriptions for each category. We elaborate on the sources of data biases in this section and provide examples from different scenarios of deployed IA systems. Based on the understanding of sources of data biases we will recommend different solutions in the further sections. Section 3 reviews the key concepts surrounding discrimination discovery, the processes described in the literature for detecting biases in existing sources of data.

In Section 4, we address the problem of building new training datasets for algorithmic systems, with the goal of minimizing their inherent social and cultural biases. Recognizing that Web and social media data sources have become a key means to create training datasets, we first review what is known about their biases, to better understand why these sources can present problems. This section is concluded with considerations for human labeled data in the social data context. Finally, the document concluded in Section 6.

# 2. Sources of data biases

In this section, we present a summary of the deliverable D4.1, which integrates and abstracts the findings of WP3. In WP3, we surveyed the scientific literature in the emerging field of Fairness, Accountability, and Transparency (FAT), characterizing the problem and solution spaces described by FAT researchers within the information and computer science disciplines. Section 2.1 begins by presenting the model of algorithmic systems. In this document, we focus on two types of data biases that may trouble an algorithmic system. The objective of this section is to summarize the types of data issues, identified in D4.1 Section 3.2, which can present risks to fairness in an algorithmic system. We then suggest ways to raise awareness of developers and regulators to such issues and train them how to examine, identify and mitigate such issues.

### 2.1 Biases in algorithmic systems

In order to discuss algorithmic systems, we first present a model that enables us to discuss the potential sources of risks to fairness and transparency and to suggest methods to mitigate these risks. An abstract algorithmic system has five main sources of risk:

**Input (I) -** The particular values input to a specific run of the algorithm.

**Output (O) -** The value(s) produced in response to the input.

**Algorithm (M).** The algorithmic core that, given a particular instance (after being trained), performs computation based on this **Input (I)** and provides an **Output (O)**. In some learning models, the algorithm itself undergoes change due to the addition of the Input (I) to the store of

**Training Data (D)**. For instance, the addition of data may come from third parties, and their interactions with the system (i.e., implicit behaviors and/or constraints, see below).

**Training Data (D).** Data which is used to train the **Algorithmic Model (M)** when some machine learning techniques are applied.

**Third Party Constraints (T).** Implicit and explicit constraints, given by third parties (not necessarily set by developers), that may impact the design and performance of the **Algorithm (M)**. These include operators of the system, regulators, and other bodies that influence the use and outcomes of the system.

Next, we highlight the two types of data biases that may trouble an algorithmic system - *input* and *training data* biases, which are the focus of the current document. These are the risks that should be monitored, in order to reduce resulting bias.

Understanding the types of biases that may manifest in an algorithmic system helps to focus on building strategies to minimize bias in a system that shows discrimination towards a certain group of people [22]. Bias in the algorithmic system often occurs in the data or in the algorithmic model. As we work to develop systems we can trust, it's critical to develop and train these systems with data that is as unbiased as possible, and to develop algorithms that can be easily explained.

As outlined in Deliverable D4.1, biases can be classified based on the causal factor, as shown in Figure 1. Figure 1 illustrates the risks to fairness and transparency according to the algorithmic system they are related to – the system components mentioned above.

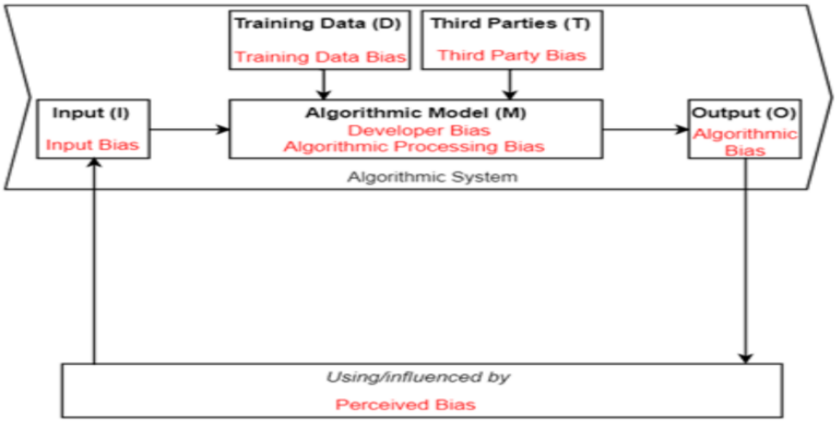


**Figure 1: Sources of biases in the system's pipeline.**

**Data biases** can be divided into two types: biases that occur in the system's **Input (I)** or biases that occur in **training data (D).**

- **Input bias:** The input data may contain information about **sensitive attributes in an implicit or explicit way.** This category also refers to the insertion of incorrect or incomplete information by the user [31].
  For example, in the case of multimedia searching, there are problems in representing image information needs with textual queries, and with representing retrieved images as short textual abstracts as a solution [23] provide data on the image queries which the number of

queries and the number of search terms per user. Users input relatively few terms to specify their image information needs on the Web. Most terms are used infrequently, with the top term occurring in less than 9 percent of queries. There was a high degree of variability across terms, with over half of the terms used only once. Terms indicating sexual content materials appear frequently. They represented a quarter of the 100 most frequently occurring terms but were a small percentage of the total number of terms overall. This is a clear case of input bias where the input information used to judge outcomes. For example, Juhi Kulshrestha [19] presents the first framework to quantify bias of ranked results in a search process, while being able to distinguish between different sources of bias. This framework not only measures the bias in the output ranked list of search results, but is also able to capture how much of this bias is due to the biased set of input data to the ranking system and how much is contributed by the ranking system itself. The analyses revealed the significant effect of both input data and the ranking algorithm in producing a considerable bias in Twitter search results based on various factors such as the topic of a query or how the query is phrased.

- **Training data bias: Information about sensitive attributes of people** may be contained in the training data and such information may be unbalanced and discriminatory with respect to particular groups of people. The training data may also be based on an unrepresentative set of instances and may also suffer from inaccurate or biased classification (i.e., inaccurate "ground truth" / annotation). The sensitive attributes can be confidential information about specific individuals such as race, sex, language, religion, political and more.
  For example, in the datasets for Computer Vision research often prefer particular kinds of images like street scenes, or nature scenes, or images retrieved via internet keyword searches which causes **selection bias [2]**. Another type of bias called **capture bias** is common across all datasets. This is caused by photographers when they tend to take pictures of objects in similar ways. Another type of bias in the dataset caused by the human's preference is **category or label bias**. This comes from the fact that semantic categories are often poorly defined, and different labellers may assign different labels to the same type of the object.
  Many systems will continue to train and update their models using this problematic data and making this an ongoing problem. Based on the understanding of this section on sources of data biases, the next section will cover the brief understanding of discrimination discovery types with some examples.

# 3. Discrimination discovery in datasets

**Discrimination discovery** is the ability to identify discrimination against sensitive groups in the population, caused by biases in an algorithmic system. It can be divided into the following kinds: the first one is **explicit (direct)** discrimination discovery and the second one is **implicit (indirect)** discrimination discovery.

- **Explicit (direct) discrimination discovery** is the ability to identify discrimination which is caused by both data biases and inappropriate use of sensitive attributes in algorithms [32].

- **Implicit (indirect) Discrimination Discovery** is the ability to identify discrimination which is caused by algorithmic processing biases and human biases due to the fact that some insensitive attributes are very informative about sensitive attributes [33].

For example, in paper [9], the authors discussed the problem of discovering both **direct and indirect discrimination** from the historical data and removing the discriminatory effects before performing predictive analysis. The proposed approach made use of the causal network to capture the causal structure of the data, and modelled direct and indirect discrimination as different path-specific effects. Based on that, the authors proposed the discovery algorithm PSEDD to discover both direct and indirect discrimination, and the removal algorithm PSE-DR to remove them.

| Measures | Indicate what? | Type of discrimination |
|---|---|---|
| Statistical tests | presence/absence of discrimination | indirect |
| Absolute measures | magnitude of discrimination | indirect |
| Conditional measures | magnitude of discrimination | indirect |
| Structural measures | spread of discrimination | direct or indirect |

**Figure 2: Discrimination measure types**

As discussed in [10], there are four types of discrimination measures (shown in Figure 2),
(1) **Statistical test:** Statistical tests are the earliest measures for indirect discrimination discovery in data. Statistical tests are formal procedures to accept or reject statistical hypotheses, which checks how likely the result is to have occurred by chance. In discrimination analysis typically the null hypothesis, or the default position, is that there is no difference between the treatment of the general group and the protected group. The test checks how likely the observed difference between groups has occurred by chance. If the chance is unlikely then the null hypothesis is rejected, and discrimination is declared.
(2) **Absolute measures:** Absolute measures are designed to capture the magnitude of the differences between (typically two) groups of people. The groups are determined by the protected characteristic (e.g. one group is males, another group is females). If more than one protected group is analysed (e.g. different nationalities), typically each group is compared separately to the most favored group. Absolute measures take into account only the target variable y and the protected variable s. Absolute measures consider all the differences in treatment between the protected group and the regular group to be discriminatory.
(3) **Conditional measures:** Conditional measure try to capture how much of the difference between the groups is explainable by other characteristics of individuals, recorded in X, and only the remaining differences are deemed to be discriminatory
(4) **Structural measures:** Structural measures are targeted at quantifying direct discrimination. The main idea behind structural measures is for each individual in the dataset to identify whether s/he is discriminated, and then analyse how many individuals in the dataset are affected.

The paper [21] is an example of a study that investigated the discrimination discovery problem on the basis of the situation testing methodology. Authors have defined a distance function on the direct causes of the decision, which takes into consideration the value difference as well as the causal effect of each attribute on the decision. Empirical assessments of the proposed approach using real data have been conducted. Another article [4] proposes a discriminative framework that directly exploits dataset bias during training in object recognition models. In particular, this

framework learns two sets of weights: (1) bias vectors associated with each individual dataset, and (2) visual world weights that are common to all datasets, which are learned by undoing the associated bias from each dataset. For more details please refer to the D4.1 section 3.3.1.

# 4. Building new datasets

A data set is a collection of data. In IA systems, we need a training data set. It is the actual data set used to train the model for performing various actions. This section will cover the possible biases found while building the datasets for different types of algorithmic information access systems. Any remedy for bias must start with the awareness that bias exists; for example, most mature societies raise awareness of bias through affirmative-action programs, and, while awareness alone does not completely alleviate the problem, it helps guide us toward a solution. The purpose of this section is to give awareness to the stakeholders and take the necessary steps to prevent them. Section 4.1 highlights the focus of this document which is recommendations while creating a dataset of "social data" (i.e., data collected from social platforms). We consider here the social data as a key source of training data for algorithmic systems. We conclude section 4.1 with a list of recommendations to mitigate the bias in social data. Bias on the web will be discussed in section 4.2. Section 4.2 concluded with proposed solutions from literature to mitigate the web bias. Finally, in Section 4.3 we discuss suggestions to mitigate the bias in datasets in any case.

### 4.1 Bias in social data

**Social data** is information that social media users publicly share, which includes metadata such as the user's location, language(s) spoken, biographical data, and/or shared links (e.g., "friends" or "contacts"). Social data is valuable to marketers looking for customer insights that may increase sales or, in the case of a political campaign, win votes. Search engines are one common scenario of an information access system where social data is important. To make good customized search results, algorithms use information specific to the person. This collected personal information is the key source of the algorithms. There are many types of social data, including tweets from Twitter, posts on Facebook, pins on Pinterest, posts on Tumblr, and check-ins on Foursquare and Yelp. Facebook for Business and Twitter Ads are two programs that help advertisers use social data to market targeted users who are likely to be interested in their ads. Social data is the training data in our algorithmic model mentioned in section 2. In the case of social data, the biggest challenge is that there is no agreement on the vocabulary of biases [5].

Figure 3 shows the overview of the possible biases in social data. We can classify them mainly in two categories: based on how biases manifest and based on where the biases come from.

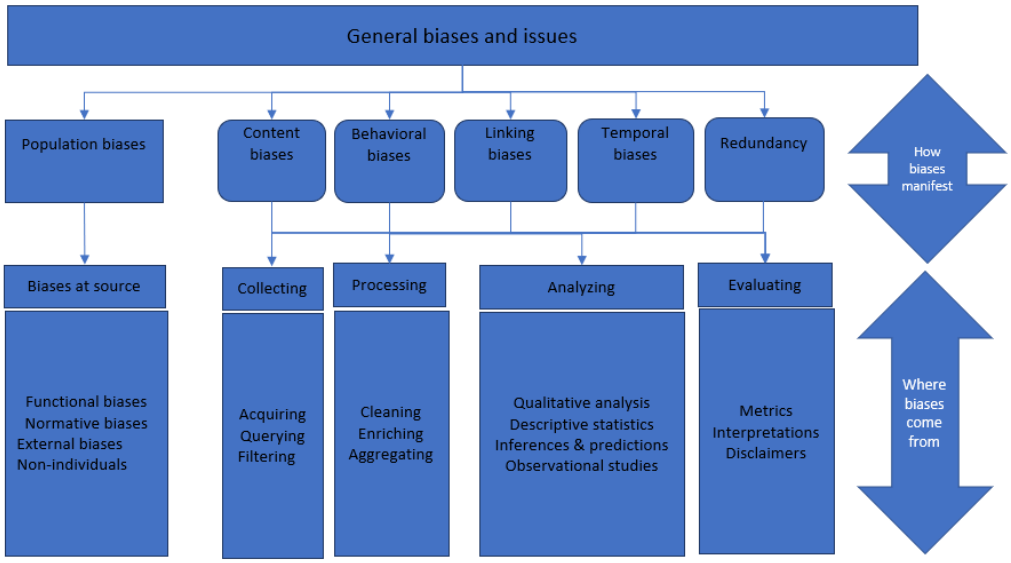


**Figure 3: Biases and pitfalls when using social data [5]**

**Based on how biases manifest, there are 6 types of biases. They are population biases, behavioral biases, content biases, linking biases, temporal variations, and redundancy.**

A. Population biases: A systematic distortion in demographics or other user characteristics between a population of users represented in a dataset or on a platform and some target population.

B. Behavioral biases: Systematic distortions in user behavior across platforms or contexts or across users represented in different datasets.

C. Content production bias: Behavioral biases that are expressed as lexical, syntactic, semantic and structural differences in the content generated by users. It is similar to data bias.

D. Linking biases: Behavioral biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity.

E. Temporal biases: Systematic distortions across user populations or behaviors over time. Data collected at different points in time may differ along diverse criteria, including who is using the system, how the system is used, and in the platform affordances. Further, these differences may exhibit a variety of patterns over time, including with respect to granularity and periodicity.

F. Redundancy: Single data items that appear in the data in multiple copies, which can be identical (duplicates), or almost identical (near-duplicates). Lexical and semantic redundancy often accounts for a significant-fraction of content and may occur both within and across social datasets. Other sources of content redundancy often include non-human accounts such as the same entity posting from multiple accounts or platforms, multiple users posting from the same account, or multiple entities posting or reposting the same content. This can sometimes distort results, yet, redundancy can be a signal by itself, for instance, reposting may be a signal of importance.

**Population biases are classified into functional biases, normative biases, external biases, and non-individual biases. Classification is due to the platform design and affordances and due to behavioral norms that exist or emerge on each platform.**

A. **Functional biases**: Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment. The functional peculiarities of social platforms may introduce population and behavioral biases by influencing which user demographics are drawn to each platform and the kind of actions they are more likely to perform.

B. **Normative biases**: Biases that are a result of written or unwritten norms and expectations of acceptable patterns of behavior on a given online platform or medium. These norms are shaped by factors including the specific value proposition of each platform, and the composition of their user base.

C. **External sources of bias**: Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, personality, culture, social pressure, privacy concerns, and external events.

D. **Non-individual accounts**: Interactions on social platforms that are produced by organizations or automated agents.

**Biases introduced due to the selection of data sources, or in which data from the sources are acquired and prepared are called data collection biases.**

A. **Data acquisition:** Acquisition of social data is often regulated by social platforms, and hinges on the data they capture and make available, on the limits they may set to access, and on the way in which access is provided. Adversarial nature of data collection leads to several challenges:

- Many social platforms discourage data collection by third parties
- Programmatic access often comes with limitations
- The platform may not capture all relevant data
- Platforms may not give access to all the data they capture.
- Sampling strategies are often opaque.

B. **Data querying**: Data access through APIs usually involves a query specifying a set of criteria for selecting, ranking, and returning the data being requested. Different APIs may support different types of queries. The challenges related to the formulation of these queries are the following:

- APIs have limited expressiveness regarding information needs.
- Information needs may be operationalized (formulated) in different ways.
- The choice of keywords in keyword-based queries shapes the resulting datasets.

C. **Data filtering**: Data filtering entails the removal of irrelevant portions of the data, which sometimes cannot be done during data acquisition due to the limited expressiveness of an API or query language. The data filtering step at the end of a data collection pipeline is often called post-filtering, as it is done after the data has been acquired or obtained by querying (hence the prefix “post-”). Typically, the choice to remove certain data items implies an assumption that they are not

relevant for a study. This is helpful when the assumption holds, and harmful when it does not. For example,

- Outliers are sometimes relevant for data analysis: Outlier removal is a typical filtering step. A common example is to filter out inactive and/or unnaturally active accounts or users from a dataset. In the case of inactive accounts, a significant fraction of users, though interested in a given topic, choose to remain silent. Depending on the analysis task, there are implications for ignoring such users. Similarly, non-human accounts often have anomalous content production behavior, but despite not being "normal" accounts, they can influence the behavior of "normal" users and filtering them out may hide important signals.
- Text filtering operations may bound certain analyses: A typical filtering step for text, including that extracted from social media, is the removal of functional words and stopwords. Even if such words might not be useful for certain analyses, for other applications they may embed useful signals about e.g., authorship and/or emotional states, threatening, as a result, the research validity.

Biases introduced by data processing operations, analyzing and evaluating are not focused on this document and for more details please refer to the article "Social data: Biases, methodological pitfalls, and ethical boundaries".

### 4.1.1 Proposed Recommendations to mitigate social data bias from Literature

- Many works propose the **use of registries** to audit the social data and provide transparency to the system. The registries are used for documenting the whole processing of the data by the system and to examine in depth the data and the system in order to detect any possible bias. Gebru et al. (2018) suggest maintaining a registry including any possible issues in data such as why and how the data was collected and pre-processed, what are the policies for its re-distribution and maintenance, and outlining possible legal/ethical concerns. Similarly, Mitchell et al., (2019) suggest using registries in the form of model cards that document the process of creating the pre-trained models. Hind et al. (2018) suggest using registries in the form of supplier's declarations of conformity (SDoCs) to describe the processing of data along with the safety and performance testing it has undergone.
- Another suggestion is the auditing of social software systems and the evaluation of social data bias and its source (Sandvig et al. (2014), Kulshrestha et al. (2017), Olteanu et. al. 2019).
- There are growing efforts to address social data limits, by encouraging the adoption of guidelines, standards, and new methodological approaches in social software
    - Employing techniques from the causal inference literature that can lead to more robust research results (Landeiro and Culotta, 2016; Proserpio et al., 2016)
    - Calibrating non-representative social data samples (Zagheni and Weber, 2015).
- Fairness sampling is the main solution provided for balancing the training social data, considering the imbalanced protected attributes. An example is the hate speech detection (Davidson et al. 2017, Badjatiya et al. 2019).

- Kusner et al. (2017) use an alternative fairness approach, the counterfactual fairness, that captures the social biases that may arise towards individuals based on sensitive attributes. They provide optimization of fairness and prediction accuracy of the classifier using a causal model.

### 4.2 Bias on the web

In this section we discuss the detection, understanding and mitigation of bias on the Web data. According to Baeza-Yates [8], the Web is today's most prominent communication channel, as well as a place where our biases converge. As social media are increasingly central to daily life, they expose us to influencers we might not have encountered previously. This makes understanding and recognizing bias on the Web more essential than ever. Any remedy for bias must start with the awareness that bias exists; for example, most mature societies raise awareness of social bias through affirmative-action programs, and, while awareness alone does not completely alleviate the problem, it helps guide us toward a solution. Bias on the Web reflects both societal and internal biases within ourselves, emerging in subtler ways. In addition, **cultural biases** can be found in our inclinations to our shared personal beliefs, while **cognitive biases** affect our behavior and the ways we make decisions. Figure 4 shows how bias influences both the growth of the web and its use.

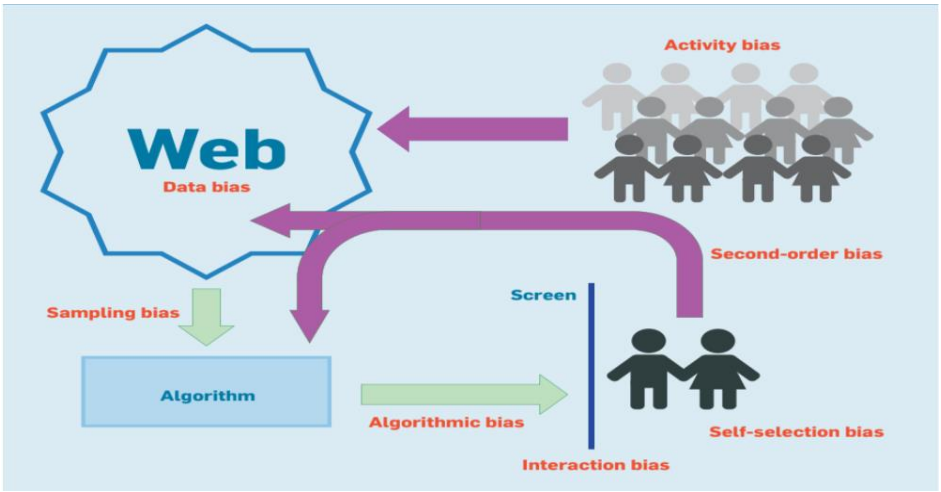


**Figure 4: The vicious cycle of bias on the web [8]**

Algorithmic bias is added by the algorithm itself and not present in the input data. In a 2016 research effort that used a corpus of U.S. news to learn she-he analogies through word embeddings, most of the results were reported as biased, as in nurse-surgeon and diva-superstar instead of queen-king [42]. A quick Web search showed that approximately 70% of influential journalists in the U.S. were men, even though at U.S. journalism schools, the gender proportions are reversed. Algorithms learning from news articles are thus learning from texts with demonstrable and systemic gender bias. Yet other research has identified the presence of other cultural and cognitive biases [43][44].

On the other hand, some Web developers have been able to limit bias. For instance, "De-biasing" the gender-bias issue has been addressed by factoring in the gender subspace automatically [42]. Regarding geographical bias in news recommendations, large cities and centers of political power surely generate more news. If standard recommendation algorithms are used, the general public

likely reads news from a capital city, not from the place where they live. Considering diversity and user location, Web designers create websites that give a less centralized view that also shows local news [45].

Social bias defines how content coming from other people affects our judgment. Consider an example involving collaborative ratings: Assume we want to rate an item with a low score and see that most people have already given it a high score. We may increase our score just thinking that perhaps we are being too harsh. Such bias has been explored in the context of Amazon reviews data [46] and is often referred to as "social conformity," or "the herding effect."[47]

Imagine we are a blogger planning our next blog post. We first search for pages about the topic we wish to cover. We then select a few sources that seem relevant to us. We select several quotes from these sources. We write new content, putting the quotes in the right places, citing the sources. And, finally, we publish the new entry on the Web. Based on the content-creation process for blogs, content used in reviews, comments, social network posts, and more, the problem occurs when the subset of content that is selected is based on the search engine that is being used. The ranking algorithm of the search engine biases a portion of a given topic's organic growth on the Web. For example, Baeza-Yates [48] found that approximately 35% of the content on the Web in Chile was duplicated, and we could trace the genealogy of the partial (semantic) duplication of those pages. Today, the semantic-duplication effect might be even more widespread and misleading.

#### 4.2.1 Proposed Solutions to mitigate web data bias from Literature

The Proposed recommendations that address the concerns of mitigating bias in web data from different articles are listed below:

- In the paper [49], authors introduced a new algorithm to gather unbiased data with reasonable user engagement metrics. They discussed how it differs from traditional exploitation and exploration work and why the proposed framework would gather unbiased data. Also, they demonstrated the effectiveness of the proposed approach through a live bucket test and showed that the method significantly improved the user engagement metrics and the skewness of a number of distributions of items.
- Joachims introduced [50] a principled approach for learning-to-rank under biased feedback data. Drawing on counterfactual modeling techniques from causal inference, the author presented a theoretically sound Empirical Risk Minimization framework for LTR. They instantiated this framework with a Propensity Weighted Ranking SVM and provided extensive empirical evidence that the resulting learning method is robust to selection biases, noise, and model misspecification. Furthermore, their real-world experiments on a live search engine show that the approach leads to substantial retrieval improvements, without any heuristic or manual interventions in the learning process.

### 4.3 Considerations of human-labelled data

Trained with poor and biased data, an algorithm is unable to deliver an accurate outcome. So possible suggestions to remove or minimize bias in datasets in the first place are:

- Make sure the datasets are representative: Making training datasets representative and balanced is key to a viable ML model that would not yield unintended or even offensive results. By analysing your training dataset from the perspective of an end user, you may be surprised to find some gaps that will require collecting additional data [51].
- Keep Only Relevant Variables: Sensitive personal attributes like gender and race are known to introduce bias and discrimination into ML algorithms. While controlling for specific input parameters like gender, race, or age is a necessary first step, it is not enough. Predictive ML algorithms can still learn these biases from other variables since they are interrelated. Zip codes, for example, can be related to income and race, profession to gender. Stripping your training dataset down to only relevant components will help reduce potential disparities and result in a fairer prediction.
- Engage External Experts: Machine Learning algorithms can easily pick up the biases of their creators. An ML model that uses historical data to predict outcomes will inadvertently reinforce any bias found in past decisions, metrics, or parameters. It should be noted that the smaller the group of people responsible for decisions is, the higher the risk of bias will be. One of the ways to combat these past injustices is to diversify our data scientist team. People with different backgrounds and life experiences will provide a fresh and even unexpected perspective to the problem at hand, helping to balance out the training dataset and make it more neutral.
- Keep Humans in the Loop: It’s erroneous to think that once an ML model is trained and put in the wild, it does not need human supervision any longer. An algorithm predicting house prices, for example, will require regular re-training with fresh, up-to-date data since the prices tend to change all the time, and predictions will become inaccurate before you know it. To ensure our Machine Learning algorithm continues to deliver accurate, unbiased outcomes, you need to remain vigilant and continue monitoring your Machine Learning model even after the launch. By frequently checking your algorithm performance against a set of indicators that reflect non-discrimination, you will be able to detect bias early on and correct the ML model by isolating and removing a problematic variable from the training dataset.

# 5. Conclusion

The goal of D4.3 is to present some “paper-based solutions” for addressing the problem of data biases in the development and use of algorithmic information access (IA) systems from the social data perspective. We started with explaining sources of data biases using the model explained in D4.1 and concepts of discrimination discovery. The major contribution of this document is to give awareness to the stakeholders about the biases and hence suggest solutions or recommendations to fix the bias in the data. We also recommend the presence of humans in the loop which we highlighted in our document.

# References

[1] Eickhoff, Carsten. "Cognitive biases in crowdsourcing." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.
[2] Torralba, Antonio, and Alexei A. Efros. "Unbiased look at dataset bias." CVPR. Vol. 1. No. 2. 2011.
[3] Ferraro, Francis, et al. "A survey of current datasets for vision and language research." arXiv preprint arXiv:1506.06833 (2015).
[4] Khosla, Aditya, et al. "Undoing the damage of dataset bias." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012.
[5] Olteanu, Alexandra, et al. "Social data: Biases, methodological pitfalls, and ethical boundaries." Frontiers in Big Data 2 (2019): 13.
[6]Faltings, Boi, et al. "Incentives to counter bias in human computation." Second AAAI conference on human computation and crowdsourcing. 2014.
[7]Aroyo, Lora, and Chris Welty. "Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard." WebSci2013. ACM 2013.2013 (2013).
[8]Baeza-Yates, Ricardo. "Bias on the web." Communications of the ACM 61.6 (2018): 54-61.
[9]Zhang, Lu, Yongkai Wu, and Xintao Wu. "A causal framework for discovering and removing direct and indirect discrimination." arXiv preprint arXiv:1611.07509 (2016).
[10]Zliobaite, Indre. "A survey on measuring indirect discrimination in machine learning." arXiv preprint arXiv:1511.00148 (2015).
[11]Wang, Tianlu, et al. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations." arXiv preprint arXiv:1811.08489 (2018).
[12] Madaan, Nishtha, et al. "Analyze, detect and remove gender stereotyping from bollywood movies." Conference on Fairness, Accountability and Transparency. 2018.
[13] Friedman, Batya, and Helen Nissenbaum. "Bias in computer systems." ACM Transactions on Information Systems (TOIS) 14.3 (1996): 330-347.
[14]Callahan, Ewa S., and Susan C. Herring. "Cultural bias in Wikipedia content on famous persons." Journal of the American society for information science and technology 62.10 (2011): 1899-1915.
[15] Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods." arXiv preprint arXiv:1804.06876 (2018).
[16] Rokicki, Markus, Eelco Herder, and Christoph Trattner. "How editorial, temporal and social biases affect online food popularity and appreciation." Eleventh International AAAI Conference on Web and Social Media. 2017.
[17] Zhao, Jieyu, et al. "Learning gender-neutral word embeddings." arXiv preprint arXiv:1809.01496 (2018).
[18] Otterbacher, Jahna. "Linguistic bias in collaboratively produced biographies: crowdsourcing social stereotypes?." Ninth international AAAI conference on web and social media. 2015.
[19] Kulshrestha, Juhi, et al. "Quantifying search bias: Investigating sources of bias for political searches in social media." Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017.
[20] Misra, Ishan, et al. "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[21] Zhang, Lu, Yongkai Wu, and Xintao Wu. "Situation Testing-Based Discrimination Discovery: A Causal Inference Approach." IJCAI. Vol. 16. 2016.
[22] Tommasi T., Patricia N., Caputo B., Tuytelaars T. (2017) A Deeper Look at Dataset Bias. In: Csurka G. (eds) Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition. Springer, Cham
[23] Goodrum, Abby, and Amanda Spink. "Image searching on the Excite Web search engine." Information Processing & Management 37.2 (2001): 295-311.
[24] https://opendatahandbook.org/glossary/en/terms/crowdsourcing/
[25]Yan, Rui, et al. "Are two heads better than one? crowdsourced translation via a two-step collaboration of non-professional translators and editors." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.
[26] Kunchukuttan, Anoop, et al. "TransDoop: A Map-Reduce based Crowdsourced Translation for Complex Domain." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013.
[27] Silvertown, Jonathan. "A new dawn for citizen science." Trends in ecology & evolution 24.9 (2009): 467-471.
[28] Sen, Shilad, et al. "Turkers, Scholars," Arafat" and" Peace" Cultural Communities and Algorithmic Gold Standards." Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. 2015.
[29] Jiang, Heinrich, and Ofir Nachum. "Identifying and correcting label bias in machine learning." arXiv preprint arXiv:1901.04966 (2019).
[30] Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased learning-to-rank with biased feedback." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017.
[31] Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems." IJCAI. 2017.
[32] Hannák, Anikó, et al. "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr." Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017.
[33] Speicher, Till, et al. "Potential for discrimination in online targeted advertising." 2018.
[34] Gebru, Timnit, et al. "Datasheets for datasets." *arXiv preprint arXiv:1803.09010* (2018).
[35] Mitchell, Alex L., et al. "InterPro in 2019: improving coverage, classification and access to protein sequence annotations." *Nucleic acids research* 47.D1 (2019): D351-D360.
[36] Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." *arXiv preprint arXiv:1810.01943* (2018).
[37] Sandvig, Christian, et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).
[38] Kulshrestha, Juhi, et al. "Quantifying search bias: Investigating sources of bias for political searches in social media." *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017.
[39] Landeiro, Virgile, and Aron Culotta. "Reducing confounding bias in observational studies that use text classification." (2016).
[40] Proserpio, Davide, Scott Counts, and Apurv Jain. "The psychology of job loss: using social media data to characterize and predict unemployment." *Proceedings of the 8th ACM Conference on Web Science*. 2016.

[41] Weber, Ingmar. "Demographic research with non-representative internet data." *International Journal of Manpower* 36.1 (2015): 13-25.
[42] Bolukbasi, R., Chang, K.W., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? De-biasing word embeddings. In *Proceedings of the 30th Conference on Neural Information Processing Systems* (Barcelona, Spain, Dec. 5–10). Curran Associates, Inc., Red Hook, NY, 2016, 4349–4357.
[43] Caliskan, A., Bryson, J.J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (Apr. 2017), 183–186.
[44] Saez-Trumper, D., Castillo, C., and Lalmas, M. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the ACM International Conference on Information and Knowledge Management* (San Francisco, CA, Oct. 27-Nov. 1). ACM Press, New York, 2013, 1679–1684.
[45] Graells-Garrido, E. and Lalmas, M. Balancing diversity to countermeasure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago, Chile, Sept. 1–4). ACM Press, New York, 2014, 231–236.
[46] Wang, T. and Wang, D. Why Amazon's ratings might mislead you: The story of herding effects. *Big Data 2*, 4 (Dec. 2014), 196–204.
[47] Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*, SSRN, Rochester, NY, Dec. 20, 2016; https://ssrn.com/abstract=2886526
[48] Baeza-Yates, R., Pereira, Á., and Ziviani, N. Genealogical trees on the Web: A search engine user perspective. In *Proceedings of the 17th International Conference on the World Wide Web* (Beijing, China, Apr 21–25). ACM Press, New York, 2008, 367–376.
[49] Hong, Liangjie, and Adnan Boz. "An Unbiased Data Collection and Content Exploitation/Exploration Strategy for Personalization." *arXiv preprint arXiv:1604.03506* (2016).
[50] Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased learning-to-rank with biased feedback." *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017.
[51] https://www.dataversity.net/making-machine-learning-datasets-unbiased/