Recommendations on Bias: Detect, Mitigate, Repeat

BORIS BELCHEV, University of Twente, The Netherlands

In today's data-driven environment, the adoption of innovative algorithms to improve efficiency and effectiveness has expanded rapidly during the past decade. Despite the benefits they give, they also carry the shortcomings of their creators. The world has already seen these weaknesses in several instances where biased algorithms have sparked public outrage. This can occasionally have severe implications on the lives of individuals affected. A framework has been created to guide the attention of individuals and organizations developing and deploying these algorithms toward their ethical aspects and the sociotechnical system in which they will reside. The framework is intended to stimulate discussion on these ethical challenges, but it does not include recommendations for identifying and mitigating bias. Therefore, the purpose of this study was to identify and synthesize recommendations from the available literature on detecting and mitigating bias. Experiments were conducted with non-expert stakeholders to validate the recommendations for detecting and mitigating bias in algorithms and datasets. A total of 24 recommendations and sub-recommendations for identifying and reducing bias were developed, and the results of the experiments shown that stakeholders with limited expertise in the subject had a reasonable grasp of these recommendations and their applicability.

Additional Key Words and Phrases: bias, mitigating, detecting, algorithms, datasets, recommendations, framework

1 INTRODUCTION

There have been algorithms for thousands of years. In approximately 2500 and 1550 BCE, mathematicians in Babylonia and Egypt employed them [7]. According to the second edition of Introduction to Algorithms, an algorithm is "a sequence of computer steps that transforms the input into the output" [20].

There are two categories of algorithms in the framework for examining the ethical implications of algorithms: static and learning [23]. In the latter category are algorithms, which are commonly included in the notion of artificial intelligence (e.g., machine learning algorithms). These learning algorithms are typically trained with data, from which they "learn" the patterns and apply them to newly encountered cases. Because these data are derived from the real world, they are subject to its imperfections, namely biases. On a worldwide scale, there are several instances of algorithms causing harm due to bias. For example, a 2019 study found that Facebook ads are prejudiced based on gender and race, which is problematic because marginalized groups may be excluded from job adverts and never get the opportunity to apply because it was not visible in their feed [11]. The "Dutch benefits scandal" (Dutch: "Toeslagenaffaire") is an example of a biased algorithm causing harm on the local Dutch stage. In one instance, more than 20,000 parents were incorrectly branded as fraudsters by a system employing a "self-learning" algorithm. Consequently, the government accused and sued those parents and children were unjustly removed from their homes. A parliamentary committee decided that an injustice was committed and that the parents were falsely accused [12]. Therefore, crucial ethical considerations were missed when developing these algorithms. Some studies "blame" it on the fact that people participating in the development of these algorithms are frequently unfamiliar with a variety of ethical considerations [18]. Others think that it is rooted in our language, history, and traditions [3]. Consequently, a framework is required to aid stakeholders in identifying and evaluating ethical defects in the algorithm and the social milieu in which it will be embedded.

One such framework was developed by van Bruxvoort and van Keulen at the University of Twente in 2021 for evaluating the ethical implications of algorithms and their surrounding sociotechnical systems [23].

Five ethical principles comprise the framework: beneficence, nonmaleficence, autonomy, justice, and explicability. The framework is comprised of questions dispersed among the five principles. Its objective is to generate conversation among stakeholders regarding the algorithm and the sociotechnical system that encompasses the algorithm during its design. Therefore, emphasize the most important ethical considerations.

2 PROBLEM STATEMENT

Currently, the framework lacks a level with more specific "recommendations." Particularly needed are advice on techniques and solutions for addressing bias in the justice principle. Due to the fact that different organizations operate in distinct domains, it is challenging to provide recommendations that apply to all or the majority of the domains (e.g., healthcare, education, government etc.). On the other hand, these recommendations typically involve a number of technical ideas that all or at least the majority of stakeholders in the sociotechnical system do not comprehend. In order for these

TScIT 37, July 8, 2022, Enschede, The Netherlands

recommendations to be beneficial to stakeholders who are not as involved in detecting and eliminating bias, their understandability is crucial. This research will seek to deliver comprehensible advice to framework users. It is not intended to give a remedy to bias in algorithms or datasets, but rather to serve as a catalyst for discussion and lead to suitable solutions.

2.1 Objectives

There are two sides to the problem mentioned above. One is the social side – e.g., population selection, data collection and the interpretation of the output. The other is the technical side – e.g., the design of the algorithm, the algorithm used for preprocessing of the dataset, the training dataset, and the test dataset. This research focused on the latter.

The main objective of the research is to provide recommendations on dealing with bias from existing literature on the technical side of the problem. There are some sub objectives that need to be fulfilled beforehand:

1. Identify bias detection techniques in algorithms/datasets.
2. Identify bias mitigation techniques in algorithms/datasets
3. Produce the intended recommendations
4. Design an appropriate approach to measure understandability of the recommendations
5. Validate with stakeholders from knowledge background that does not include detecting and mitigating bias methods.

The main methods for detection and mitigation were chosen based on three criterions – firstly "Proof of Concept". Another criterion was if a method/technique is a best practice in the industry. Last criteria was if the method can be applied to multiple domains and not only for the specific problem in the study which produced it. The first criterion is important because it gives justification for the feasibility of the approach by testing it in practice [5]. The second is typically an unofficial norm created by the industry that employs such procedures, which is based on years of experience by qualified professionals. The final criterion is significant because this research seeks general methods that can be applied to datasets regardless of the data or the structure of the data within or without depending on the algorithm's specifics; therefore, only methods that can be applied to most or all domains are considered.

2.2 Research questions

The problem statement produced the following research question:

What recommendations can be made for detecting and mitigating bias that are understandable to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?

To provide the answer to that research question some sub-questions needed to be addressed:

1. Which method for detection and mitigation are relevant for algorithms and their datasets?

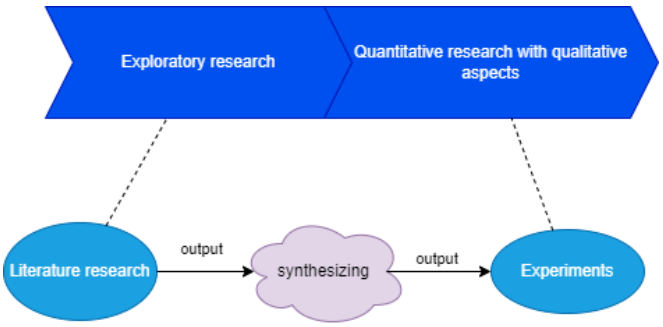

Fig. 1. Diagram representing the methodology.

2. How to make understandable recommendations from the methods in (1) to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?
3. To what extent the proposed recommendations in (2) are understandable to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?

3 METHODOLOGY

The main purpose of the research was to provide synthesized recommendations understandable to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets. To answer the main research question exploratory and quantitative with qualitative aspects research was conducted [15, 19]. To answer research sub-question (1) a secondary research method within the exploratory stage was used – literature research to find existing solutions and synthesize them to answer research sub-question (2). The output of that research was used in the experiments with stakeholders that were conducted in the quantitative stage with qualitative aspects which helped answering research sub-question (3). **Figure 1** represents the main methodology steps. The conducted experiments with the produced recommendations on participants who applied them to specific contexts in order to evaluate their understanding using scoring. The qualitative component consists of observations that were made during the experiment and post-experiment discussions.

3.1 Literature research

For the literature research databases such as Scopus, Science Direct and arXiv were used. The main key terms that were used are: "bias", "detecting", "mitigating", "algorithms", "dataset", "measures". They were used in various combinations such as: "bias" with "detecting" and "algorithms"/"datasets", with "mitigating" and "measures" and "algorithms"/"datasets", with "measures". The literature research included not only scientific articles but also journals, books, and web articles. Whenever relevant sources could not be found Google scholar was used for a broader search. Also, studies that were not relevant for the research were used to find references to literature that is relevant.

3.2 Synthesizing recommendations

The findings were subjected to synthesis, but this was not the case for all of them due to the fact that there are already measures that have been synthesized based on the results of other investigations. Researching the process and ideas that underlie it allowed to translate specific technical details into a format that was more abstract and generic for the purposes of this stage (synthesizing). Aside from that, there were procedures and approaches that were comparable, or the fundamental idea that underpinned them was the same. They were "interpreted" in terms of the overarching principle that underlies them. The results of this step were presented in the form of two tables: one for detecting bias, and another for mitigating its effects. The table for recommendations on mitigating bias can be found in **Appendix A.1** and for detecting in **Appendix A.2**. The recommendations were either labelled with a "D1" or "M1" annotation, where "D" refers for detecting and "M" means for mitigating. Another piece of information that was shown in the tables that is assumed to add value was whether or not the suggestions apply to the dataset (both the training dataset and the dataset), or to the model or algorithm (or to both). In addition, the definitions of a few of the concepts that are utilized, such as "dataset" and "classification," can be found at the beginning of the document. Examples of problematic datasets were included after the experiments.

3.3 The experiment

The experiments were conducted with individuals from different backgrounds (e.g. teachers, students, administration workers, etc.). The participants were recruited by two requirements – based (studying/working) in the Netherlands because the framework and its recommendations is utilized in entities based in the Netherlands and the lack of expertise or knowledge in the field. Their age ranged from 20-29 years old. They were sent invitations through email with a pdf document that contained the recommendations and a link to a questionnaire. Informed consent was distributed prior to starting the questionnaire that they were intended to complete. The informed consent followed the guidelines of EEMCS Ethics committee.

The questionnaire resembled an examination of their skills to apply the recommendations. It contained cases with context explained and asked for one or two of the recommendations that was the most appropriate to their understanding. One such case can be observed on **Figure 2**. The aim of the questionnaire was to validate how understandable are the synthesized recommendations to stakeholders from different domains with no expertise in the field of detecting and mitigating bias in algorithms and datasets. Another aim was to observe gaps or errors in the recommendation themselves by observing common mistakes in the questionnaire and noting down the comments of the participants. They were also required to give opinion on the recommendations and the questionnaire at the end, which revealed their own observations on the experiment itself.

Scores were assigned for every correct answer (1 or 2 points). However, there were answers that were incorrect but justified by the participants and most of them were accepted as well (0.5 or 1 point).

id	name	surname	profession	gender	age	class
1	John	Smith	lumberjack	male	31	1
2	John	Doe	truck driver	male	26	1
3	Jake	Curting	teacher	male	40	1
4	Vanesa	Doe	secretary	female	35	1
5	Isabel	Sanchez	bank manager	female	32	1
6	Olivia	Smith	accountant	female	24	0
7	Elijah	Johns	architect	male	29	0
8	Benjamin	Williams	artist	male	37	0
9	Liam	Logan	graphical designer	male	27	1
10	Sophia	Addison	logistician	female	40	1
11	Emma	Levine	barrista	female	22	0
12	Ava	Hansley	journalist	female	42	1

Fig. 2. Dataset used in the questionnaire with class imbalance.

3.4 Design flaws of the experiment & changes

The experiment was not perfectly designed, so it required some additional fixes and alterations to ease the participants. Even though in total there were 10 respondents the first one was utilized as the test experiment. Changes were focused only on the questionnaire itself and on the setup of the experiment. The changes are listed as bullet points:

- Questions were changed to further specify the number of possible answers (one or two).
- Additional context information was added to the questions to further clarify them.
- Assistance from the researcher to the participants was done through meetings (online & in-person) or through email and text.
- Setup was changed from initially without the researcher's attendance during the experiment to the researcher attending every experiment for an easier and faster communication.
- Some questions were identified to be vague and they were not changed but the researcher explained them further during the experiments

4 RESULTS

4.1 Literature research results & recommendations

The reader can find the recommendation tables in **Appendix A**. These are the tables of recommendations that resulted after implementing most of the feedback from the participants.

The recommendations for detecting bias (D1 to D8) were found among different studies that show and experiment with metrics and best practices to detect bias in real-world datasets and algorithms. In total 7 recommendations and 8 sub-recommendations were synthesized.

Starting from the first recommendation D1 and its sub-recommendations they were identified in a study that surveys and discusses fairness metrics from existing research[1]. They focus on the use of sensitive attributes in the algorithm and the implicit correlation with them and classification in the training dataset. The core concept behind D2 was synthesized from multiple studies that discuss false positives to detect or evaluate bias[1, 13, 25] The recommendations from D3, D4.1 and from D5.1 to D5.3 including were identified from a study that uses all these metrics (class imbalance, skewness, etc.) in conjunction to

detect bias [22]. D4 was further backed and extended (with D4.2) by other studies that use demographic and sample parity as an approach or part of the approach to identify bias. [1, 8, 13]. In the same manner D5.2 and D5.3 were confirmed by other studies that use Kullback-Leibler divergence and Kolmogorov-Smirnov test to investigate for bias [8, 16]. D5.4 is a recommendation connected in the context of distribution of data of the rest of the sub-recommendations in D5 and it was found in research that investigated popularity bias in systems using recommendations based on popularity of the product (e.g., Netflix) [6]. D6 and D7 were the recommendations that apply solely to the model/algorithm. D7 was synthesized from an approach which swaps the value of biased attributes and compares the classification afterwards [16]. While D6 is based on an approach that uses transparent student model which represents the real black-box model and another model that will predict the actual outcome and then compare them [17].

For mitigation of bias 6 recommendations were synthesized and 4 sub-recommendations.

Starting from M1 this recommendation was produced as a follow-up of recommendation D1 and its sub-recommendations [1]. M2 and M3 was identified in a study that investigates mitigating bias approaches from literature and builds on them [2]. M3 was divided in 3 sub-recommendations because the methods differ but are all applicable to the same problems (parity/parity combined with class imbalance). M2 & M3.2 were found a study that solves unbalanced and noisy data by using the Snowball technique and duplication of instances[24]. M3.1 was implicitly synthesized from M2.1. M3.3 was identified in a study that makes a literature review on existing biasing techniques and one of the studies included was using a method that constraints the predictions of the model [10, 21]. M4 carries the core concept behind the solution using geometric deep learning to improve the detection hateful speech [25]. M5 is a conjunction of multiple works that use augmented datasets to debias their data – one approach is to create new artificial training datasets which is unbiased and the other proposes to take the union between the original dataset and one with swapped values for the protected attribute [9, 14]. Continuing the notion of using adversarial learning M6 was found and synthesized which proposes to use a model (discriminator) that predicts the protected variables just from the classifications outputted by the model which is to be debiased [4].

4.2 Results from the experiments

There was a total of 10 participants in the experiments. The first participant's score is excluded from the final result since it was utilized as a so-called test experiment to identify flaws and mistakes in the setup and questionnaire. Consequently, the results are based on the nine participants.

The average (mean) and the median are similar, with the former having a value of 9.22 and the latter having a value of 9. The minimum score on the questionnaire is 0 and the maximum is 16. The minimum score in these studies is 3.5 and the greatest is 14. Seven participants score greater than or equal to 50 percent (8 p.), while the remaining two score below 50 percent (**Table 1**).

Table 1. Quantitative statistics of the experiment

Average(mean)	Median	# =>50%	# <50%
9.22	9	7	2
# - number of participants			

When it comes to the separate recommendations the most understood ones from the detection of bias were D3, D1.2, D4, D5.1, D5.2, D5.3 with 5-7 out of 9 respondents applying them appropriately. While for mitigation of bias were M1 with record score of 8/9 respondents using it appropriately. And the rest M4, M5, M6 with 5-6 respondents out of 9.

Besides the quantitative results of the trials, there were also qualitative findings. During the experiment, notes were obtained from participant conversations. The observations of the researchers are also recorded.

It was discovered that sub-recommendations D1.1 and D1.2 are understood similarly, and the distinction is difficult for participants to grasp. The first is primarily concerned with the model/algorithm process of prediction (what happens in the black box), whereas the second is concerned with the algorithm's output categorization as well. Although it was previously believed that D3 is straightforward and simple to comprehend, this proved not to be the case. Participants agreed that the recommendation is not sufficiently self-explanatory. Regarding sub-recommendations D5.1 to D5.4, three participants stated that they saw no distinction between them.

Participants noted, based on the recommendations for reducing bias, that M3.2 and M3.3 had the same underlying concept, causing confusion.

After the debriefing that followed the submission of the questionnaire, participants saw the correct responses and their mistakes, and the majority of them stated that if the context was made clearer and there were more instances of biased and unbiased datasets and cases, they would get better results.

From the side observations of the experiments conducted by the researchers, it was determined that D2 is also somewhat problematic due to the "false positive" idea. In addition, D6 and D7 were the most popular proposals for detecting bias whenever the context of the questionnaire demanded their application to a model or algorithm. However, they were frequently mistaken with D1 (D1.1 & D1.2) due to their applicability to models and algorithms. D2 was frequently mistaken for D7. Given the context of incorrectly categorized minorities, M4 and M3.2 appeared to have the same meaning. Side observations also confirmed that the comment of participants that mentioned the lack of context is justified.

4.3 Resulting recommendations after the experiments

During the experiment participants were presented with tables of recommendations that were changed by implementing their feedback and researcher observations. **Table 2** shows what and how the changes were done. Some recommendations changed their numbering because of others that were combined (e.g., D3 became D2 because D1 & D2 were combined).

Table 2. Changes to the recommendations and their structure

Before the experiments	After the experiments
D1 & D2	Became D1 with D1.1(D1) & D1.2(D2)
D3->D2 was concerned with the dataset and model/algorithm	Now only concerned with the model/algorithm. Example of false positive was added
D4->D3	Added a simplified example of class imbalance and balance
D5,D7,D8	Became one D5 with D5.1(D5), D5.2(D7), D5.3(D7.1), D5.4(D8)
M1	Added explicit mention that to be done only if
M2 was concerned with both class imbalance and parity	Now it is concerned only with class imbalance
M2.2, M2.3, M2.4	Now in M3 (parity) as M3.1, 3.2 and 3.3

Apart from the changes in **Table 2** recommendations were improved by rewording, explicitly mentioning "training dataset" where beforehand was omitted and the column "Concerned with" was limited or made more specific to what part the recommendation should be applied.

5 DISCUSSION

The literature review conducted to find techniques of detecting and mitigating bias to address the first research sub-question – "*Which methods for detection and mitigation are relevant for algorithms and their datasets?*", found that, despite the thousands of studies that detect and mitigate bias from specific (e.g., photos, text, etc.) to more generic examples, it always returns to the same idea and concept. Some of the discovered strategies were anticipated, such as exploiting class imbalance to detect prejudice, but others were quite unexpected, such as using adversarial networks to counteract bias. The discovered methods may not exhaust all options for detecting and mitigating bias in literature and industry, but they do give a solid foundation for future research in this area. The second research sub-question – *"How to make understandable recommendations to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?"*, is addressed by the synthetization part of the methodology (**Section 3.2**). The recommendations synthesized from the discovered methods do not necessarily represent the method explicitly. The majority of them were retrieved using one or many techniques (e.g., D2). In certain instances, they are a simplification of the procedures (e.g., D7). They do not represent the raw techniques or solutions to problems including bias in a model/algorithm or dataset. Their objective is to fulfil the justice principle of the framework so that stakeholders participating in the debate of the to-be-integrated algorithm will be better able to combat bias and achieve better results.

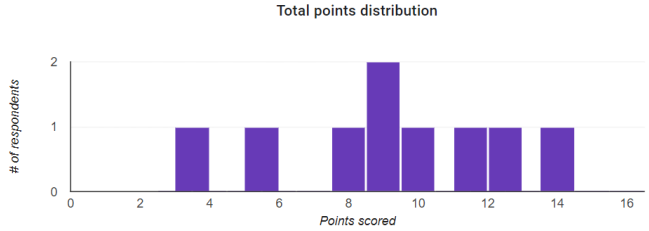

Fig. 3. Distribution of number of respondents to scores

The experiments in quantitative research with qualitative aspects have yielded encouraging results on the recommendations comprehensibility and applicability which addresses the third research sub-question – *"To what extent the proposed recommendations are understandable to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?"*. The average (mean) and median are nearly identical with a difference of 0.22, indicating that the distribution is symmetric, with the majority of responders scoring around 9 points. This suggests that they can effectively implement at least half of the recommendations. This is also evident if we divide the respondents into those who scored more than fifty percent and those who scored less than fifty percent (**Table 1**). Visually examining the bar chart (**Figure 3**) that displays the number of respondents/participants for each score, we can see that it is slanted to the right (to the maximum). All of the foregoing indicate quantitatively that the majority can comprehend and implement the recommendations. Even though the participants were assisted during the experiment that was needed due to the lack of context. This assistance was given only if participants asked, and it was limited to extra explanations and examples to better understand the questions.

The qualitative results indicate that the comments participants made about the recommendations are frequently justified and that some of them can be restructured in a more effective manner, such as the changes described in **Section 4.3**. Given that the experiment lasted little more than 45 minutes, it is understandable that the topic's context was insufficient, as indicated by the side observations confirmed after the experiment's debriefing phase. The two lowest results were the ones that spend the least time on acquainting with the recommendations, so correlation is assumed between time invested in them and scoring afterwards. In actual situations where the framework is applied, the framework's users and the recommendations have a greater understanding of the context of their algorithm and datasets.

The approach employed to find and synthesize the recommendations on bias and the total number of recommendations produced and validated through experiments allows us to answer the primary research question. - ***"What recommendations can be made for detecting and mitigating bias that are understandable to stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and datasets?",*** addressed by the findings of this

study. As more than the majority of the participants scoring above 50% on these same recommendations. The answer is further expanded by the observations of the different type of recommendations – mitigation is more understandable and easier to apply than the detection recommendations. And lastly the answer becomes even more fine grained when the results for individual recommendations are considered (**Section 4.2, 3rd paragraph**)

5.1 Limitations

During the research, some limitations were apparent. One of them was the restarting of research on bias mitigation strategies. The majority of the identified techniques required to be replaced with ones that are more generalizable and appropriate for the intended users of the framework and recommendations. A large number of strategies and methodologies could not be synthesized due to their complexity and notions that required to be explained beforehand. The duration of the studies could not exceed 45 minutes in order to maintain the participants' concentration and prevent fatigue. The context of the questions has been reduced to a minimum to ensure that the questionnaire covers all recommendations.

5.2 Future work

This research did not identify and generalize all of the known strategies for detecting and reducing bias in its findings. To adequately generalize the majority of them would require additional assistance from experts in the field. It would be necessary to obtain validation from these very same experts. In the future, research should also look into ways to expand the current recommendations and the issues that they address (such as class imbalance) with additional sub-recommendations that are more domain specific (e.g., healthcare systems). The currently produced recommendations have to be verified with instances from the real world and in real practice.

6 CONCLUSIONS

This research aimed to generate and evaluate recommendations on detecting and mitigating bias, which would be added to the justice principle of the framework for assessing the ethical elements of algorithms and their socio-technical system. It sought to demonstrate that it is possible to develop such recommendations that can be understood and utilized in conversations by stakeholders with no expertise in the field of detecting and mitigating bias in algorithms and their datasets. It turns out that this is also a study into bridging the gap between the domains that produce algorithms and their datasets and those who are affected by them. Which corresponds to the concept behind the framework they will inhabit. This endeavor showed that it is possible to develop such suggestions with simply their key concepts expressed in a language that the majority of stakeholders can comprehend. Understanding is essential if we seek to bring value to the framework and the discussions that the framework and recommendations will now spark. And this is of utmost relevance for the future development of algorithms that do not harm or negatively disrupt the lives of people.

ACKNOWLEDGMENTS

First, I would like to express my gratitude for the research topic and help throughout this research to my supervisors Maurice van Keulen and Xadya van Bruxfoort without who this endeavor would not be possible. Next, I want to thank the participants in my experiments for their time and patience. Lastly, I want to thank University of Twente and especially the EEMCS faculty its members for the hard work to provide us with education even in times of crisis during my bachelor at Technical Computer Science.

REFERENCES

[1] Agathe Balayn, Christoph Lofi, and Geert -J. Houben. 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal* 30, 5 (Sep. 2021),739–768. doi: 10.1007/s00778-021-00671-8.

[2] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, (Jan. 2019), 289–295. doi: 10.1145/3306618.3314243.

[3] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science (1979)* 356, 6334 (Apr. 2017), 183–186. doi: 10.1126/science.aal4230.

[4] Brian H. Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (Dec. 2018), 335–340. doi: 10.1145/3278721.3278779.

[5] Catherine E. Kendig. 2016. What is proof of concept research and how does it generate epistemic and ethical categories for future scientific practice? *Science and Engineering Ethics* 22, 3 (Jun. 2016), 735-753. doi.org/10.1007/s11948-015-9654-0

[6] Emre Yalcin and Alper Bilge. 2021. Investigating and counteracting popularity bias in group recommendations. *Information Processing & Management* 58, 5 (Sep. 2021), 102608. doi: 10.1016/j.ipm.2021.102608.

[7] Évelyne. Barbin, Jacques Borowczyk, Michel Guillemot, and Anne Michel-Pajus. 1999. *A History of Algorithms*, vol. 23. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-18192-4.

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

[9] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-W. Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. (Apr. 2018). arXiv:1804.06876. https://arxiv.org/abs/1804.06876

[10] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-W. Chang. 2017. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". (Jul. 2017). arXiv:1707.09457. https://arxiv.org/abs/1707.09457

[11] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–30. doi: 10.1145/3359301.

[12] Patricia Huisman, Hoe de toeslagenaffaire kon gebeuren. *Management Kinderopvang* 26, 2 (Mar. 2020), 36–37. doi: 10.1007/s41190-020-0260-2.

[13] Pedro Saleiro Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfield, Kit T. Rodolfa and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. (Nov. 2018). arXiv: 1811.05577. https://arxiv.org/abs/1811.05577

[14] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan and Kush R. Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63, 4/5 (Jul. 2019), 3:1-3:9. doi: 10.1147/JRD.2019.2945519.

[15] Richard Swedberg. Exploratory research. 2020. In C. Elman, J. Gerring, and J. Mahoney eds. *The production of knowledge: Enhancing progress in social science,* Cambridge University Press, 17–41. doi: 10.1017/9781108762519.

[16] Salem Alelyani. 2021. Detection and Evaluation of Machine Learning Bias. *Applied Sciences* 11, 14 (Jul. 2021), 6271. doi: 10.3390/app11146271.

[17] Sarah Tan, Rich Caruana, Giles Hooker and Yin Lou. 2018. Distill-and-Compare. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (Dec. 2018), 303–310. doi: 10.1145/3278721.3278725.

[18] Solon Barocas, Elizabeth Bradley, Vasant Honavar and Foster Provost. 2017. Big Data, Data Science, and Civil Rights. (Jun. 2017.). arXiv:1706.03102. https://arxiv.org/abs/1706.03102

[19] Steven J. Taylor, Robert Bogdan and Marjorie DeVault, *Introduction to qualitative research methods: A guidebook and resource*. 2015. John Wiley & Sons.

[20] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2001. Algorithms. In *Introduction to algorithms* (2nd ed.), MIT Press, Ed. Cambridge, Mass. : MIT Press, ©2001, 5–6.

[21] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang and William Y. Wang. 2019 Mitigating Gender Bias in Natural Language Processing: Literature Review. (Jun. 2019). arXiv:1906.08976. https://arxiv.org/abs/1906.08976

[22] Venkata N. Mandhala, Debnath Bhattacharyya, Diviya Midhunchakkaravarthy and Hye-jin Kim. 2022. Detecting and Mitigating Bias in Data Using Machine Learning with Pre-Training Metrics. *Ingénierie des systèmes d information* 27, 1 (Feb. 2022), 119–125. doi: 10.18280/isi.270114.

[23] Xadya van Bruxvoort and Maurice van Keulen. 2021. Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Applied Sciences* 11, 23 (Nov. 2021), 11187. doi: 10.3390/app112311187.

[24] Yi Lu, Hong Guo, and Lee Feldkamp. 1998. Robust neural learning from unbalanced data samples. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)* 3, 1816–1821. doi: 10.1109/IJCNN.1998.687133.

[25] Zo Ahmed, Bertie Vidgen, and Scott A. Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science* 11, 1 (Dec. 2022), 8. doi: 10.1140/epjds/s13688-022-00319-9.

A APPENDIX A

A.1 Appendix A.1

Legend: TD – training dataset, M/A – model/algorithm, D - dataset

Recommendations #	Recommendation	Concerned with
M1	Remove all protected variables (e.g., race, gender, age, etc.) from your training dataset or from your algorithm's decision-making/prediction process if doing so does not degrade the algorithm's performance.	M/A, TD
M2	Dealing with class imbalance: One solution is to duplicate some of the records from the underrepresented class in the training dataset. In this manner, you would balance the number of instances from the below-mentioned class.	TD
M3	Dealing with parity: **M3.1** In a dataset, the same concept as in M2 can be applied to demographic disparity or sample size disparity. **M3.2** Another approach is to train your model solely with examples from the minority that are positively classified (e.g., they get the loan) and then add an increasing number of examples from the majority that are negatively classified (e.g., they don't get the loan) in the next iteration of training. As a result, the model is first taught to be favorable to positive examples, and thus prioritizes them. **M3.3** Constrain the predictions of your model: If you are unable to add/change your dataset in which the proportion of groups (male proportion greater than female proportion) is not equal, you can restrict the ratio of one group to another predicted to be of a certain class.	TD, D M/A M/A
M4	If your algorithm produces a high number of false positives, you can include more contextual information in the training process. If your algorithm, for example, deals with a social network of people, you can include the connections between them.	M/A
M5	Generating debiased data: There are methods for creating a new "fake" dataset that is fairer than the original dataset in terms of the attributes that cause the bias. **M5.1** Data augmentation: Create a second dataset using the original one by just swapping the values of the protected variable (e.g. male becomes female and the other way around). Merge both datasets into one and use the newly created dataset for training your model.	TD
M6	Another approach to reducing bias is to use another model that has been specifically trained to predict the protected variable (e.g. gender, race, age, etc.). Then, by adjusting itself, your model will attempt to "fool" the aforementioned one (retraining). As a result, the model that attempts to predict the protected variable will eventually fail.	M/A

A.2 Appendix A.2

Recommendation #	Recommendation	Concerned with
D1	Use of sensitive attributes: **D1.1** Check if your algorithm uses sensitive attributes (e.g. race, gender, age) in the prediction/classification process. If that is the case, then bias exist. **D1.2** Bias exists when individual data records with the same attribute values except for the sensitive attributes are classified with different outcomes.	M/A (TD)
D2	Number of false positives – you can take the dataset and classify the records by other means apart from your model/algorithm (e.g. manually "by hand") this would be the "true" and expected predictions of the model. Then you give as an input that same dataset to your model/algorithm and compare the output from that to the "true" predictions. As false positives are considered records that are predicted as positive (e.g. deserves the loan, marked as hate speech) even though they are not positive in the "true" predictions.	M/A
D3	Class imbalance – Algorithms are mostly used to classify data records (e.g., positive or negative, deserves a loan or does not deserve a loan). One method for detecting bias is to count the number of instances/records labeled as belonging to a specific class. If this class has more instances than another, you can expect to have a biased algorithm after training it.	TD, D
D4	Parity: **D4.1** Sample size – A bias can be observed if the majority of your data comes from one group (e.g., White Americans, male, etc.) and the rest from another (e.g., African Americans, female, etc.). **D4.2** Demographic – When classified by the algorithm, all groups (e.g., male and female) should receive equal positive outcomes. For example, if 20% of the male population receives a loan, the female population should follow suit. Of course, if male and female populations are proportionately equal to the total population (50/50).	TD, D
D5	Data distribution: **D5.1** Skewness – If you disperse or distribute unbiased data based on intervals of values of a specific attribute (e.g., age), it will usually take the shape of a bell (Figure 1). In reality, data is skewed from that shape. Bias can be detected by measuring the level of skewness from that shape. Of course, this is based on some threshold that should be determined ahead of time for the data that will be measured. **D5.2** Probability for data to be generated – The majority of the time, data is generated at random. In reality, this is frequently not the case because gathering data from a population is difficult and has many limitations (the demography of the population that is close to the data center and etc.). Distribution refers to how different data records appear in the dataset (Figure 2). Every record has a chance of occurring. Probability distribution refers to the measurement of the likelihood of data records occurring in a specific sequence. This distribution can also be used to detect bias. For example, if you know the "ideal" probability distribution of unbiased data. And then you can compare the probability distribution of your data to the "ideal" distribution, and if the divergence is greater than a certain threshold, you can conclude that there is bias. **D5.3** You can use the same steps to compare the protected group/attribute to the rest of the data in the dataset. If their probability distributions are close or equal, the protected group/attribute is not biased. Otherwise, it may be considered biased. **D5.4** Rich gets richer (popularity bias) – This is commonly observed in recommendation systems, where movies, products, and so on that are liked by the majority of users will be the most recommended in the future and will receive even more approval from users. Of course, this is not limited to recommendations; it can be applied to any algorithm that selects from the most highly "rated" data. If you have rating systems and product ratings, you should compare the distribution of your rated items to the long-tail distribution (Figure 3). If you see a long-tail trend in your data, your system may have a popularity bias.	TD, D
D6	Simulate the model – You can train a transparent model that approximates your model. It's known as "teacher-student" model distillation. Like in the real world, the teacher model teaches the student model by training it with the results of its own predictions. Aside from that, you'd need a model that has been trained on real-world data. Examine the student and true models' classifications/predictions. If the differences are significant, you can conclude that the model under scrutiny for bias (the teacher model) is biased.	M/A
D7	Swap the values of the potentially biased attributes – With the original dataset, train the model and predict the class label for each record. The model is then trained to predict the class label for each attribute by swapping (alternating) the values of those attributes (e.g., "male" becomes "female"). If the results of both steps differ, the model may be considered biased.	M/A