Data-driven AI development: an integrated and iterative bias mitigation approach

Youssef Ennali Msc.[1][0000-0003-0573-4815] and Prof. dr. Tom van Engers[2][0000-0003-3699-8303]

[1] University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands
youssef.ennali@gmail.com
[2] University of Amsterdam - TNO, Science Park 904, 1098 XH Amsterdam, Netherlands
vanEngers@uva.nl

Abstract. This paper presents an explanatory case study aimed at exploring bias leading to discriminatory decisions generated by artificial intelligence decision making systems (AI-DMS). Particularly machine learning-based AI-DMS depends on data concealing bias emerging from society. This bias could be transitioned to AI-DMS models and consequently lead to undesirable biased predictions. Preventing bias is an actual theme both in academia and industry. Academic literature generally seems to be focused on particular bias mitigation methods, while integrating these methods in the development process of AI-DMS models remains underexposed. In this study, the concepts of bias identification and bias mitigation methods are explored to conceive an integrated approach of bias identification and mitigation in the AI-DMS model development process. Reviewing this approach with a case study showed that its application contributes to the development of fair and accurate AI-DMS models. The proposed iterative approach enables the combination of multiple bias mitigation methods. Additionally, its step-by-step design empowers designers to be aware of bias pitfalls in AI, opening doors for an "unbiased by design" model development. From a governance perspective, the proposed approach might serve as an instrument for AI-DMS models' internal auditing purposes.

Keywords: Artificial Intelligence Decision-Making Systems, Bias Mitigation, Bias, Legal Compliance, Explainable Artificial Intelligence, IT-Audit.

1 Introduction

"Unfortunately, we have biases that live in our data, and if we don't acknowledge that and if we don't take specific actions to address it, then we're just going to continue to perpetuate them or even make them worse."
– Kathy Baxter, Ethical AI Practice Architect, Salesforce

This quote reveals a hidden danger in utilizing real-world data in machine learning applications used for decision-making systems (AI-DMS). Several cases revealed that utilization of such technology also comes with a major drawback, referring to bias in prediction and/or decision outcomes. Besides the benefits, these systems might have an undesirable biased outcome. The biased outcomes are derived from data containing either explicit and/or implicit human biases [1], as the data used represents the real-

world. These pre-existing biases manifested in data emerge from society, end up in our technical systems [2, 3], eventually sustaining and even amplifying a discriminative society [1, 3].

A well-known example is the system COMPAS used in the US, determining a risk score for recidivism amongst convicts. Criminal history, among other variables, is used to predict the risk score. African Americans were more likely to score a higher risk value than their actual risk compared to Caucasians [1, 4, 5]. COMPAS was used to support decisions regarding the placement, supervision, and case management of defendants. This COMPAS system is the subject of our case study presented in section 3 of this paper. Another example is SyRI used by the Dutch authorities to track down suspects of social benefits fraud. It was concluded in a court order that SyRI comes with a considerable risk that the system discriminates, stigmatizes, and is an invasion of the citizens' privacy due to the lack of transparency [6]. The presented examples reveal that the implementation of these systems is intertwined with ethical and legal implications.

In this paper, we focus on debiasing a supervised machine learning AI-DMS model. It is argued that debiasing data should contribute to the development of fair, accountable, and transparent AI-DMS [1, 7]. A difficult task since human bias might be hidden in data due to certain proxy variables, resulting in proxy discrimination [3, 8]. An AI-DMS model could incorporate these proxy variables to generate predictions used for decisions, resulting in discriminative decisions, e.g., rejecting an insurance application. There is currently an ongoing concern in the Netherlands regarding a system similar to COMPAS incorporating postal codes of convicts to predict their recidivism score. It is argued that this system, RISC, facilitates ethnical profiling through the proxy variable postal code [9].

Despite that AI-DMS has a large potential with various benefits, the bias drawback is not one to be ignored. Organizations are required to resolve these ethical and legal issues to achieve acceptance of their AI-DMS. In the academic field, a considerable amount of research is focused on debiasing methods to cope with the bias issue. However, these studies focus on specific debiasing methods, while the development process of machine learning models explicit addressing where to apply debiasing methods remains underexposed. What debiasing methods to apply depends on the selected machine learning approach and data used. The novelty of our framework is not the framework itself as there have been different machine learning frameworks suggested before in studies and also in the industry. The novelty of the framework presented in this paper is that it explicitly incorporates debiasing. The framework is reviewed using the COMPAS data set.

2 Preventing Bias

2.1 Bias and Fairness

Bias, a persistent multifaceted societal problem is generally considered to be in favor or against an individual or group with certain properties (e.g.: age, gender, ethnicity, sexual orientation, religious background and so forth) [2], in a way that is unfair. It is a longstanding phenomenon as old as human civilization [3]. Due to its multifaceted character, it is studied in many disciplines including social science, computer science,

psychology, philosophy, law, and so forth. In this study, the coverage of bias is restricted to the following definition: *a prejudice or tendency in predictions made by an AI-DMS leading to decisions against or in favor of one individual or group in a way considered to be unfair.* The last part of this definition covers the distinction between substantive and statistical bias. The first is valuating people based on the group they belong to in absence of a natural link between that value and the group determining factors. Substantive bias should always be avoided but, as we will argue for later, statistical bias may be unlawful as well.

2.2 Explicit and Implicit Bias

Usually, the training data consists of historical events in the real world to predict future outcomes. Since the bias problem originates in society, the problem is obviously transitioned to the data. Bias in data can be present explicitly or implicitly; in other words, direct or indirect bias [7]. Explicit bias is more obvious to identify in data. For instance, data containing variables depicting ethnicity, gender, or other properties could result in discriminative implications. Such variables are also known as sensitive features/attributes.

A more challenging bias to identify is the implicit kind. Bias could be present through proxies [3, 8]. These are variables which indirectly correlate with other features. Some examples are postal codes where the majority of the population is of a certain ethnicity, a first name related to an individuals' gender, a first and/or last name usually used in certain cultures or religions and so forth.

2.3 Other Types of Bias in Machine Learning

Besides bias leading to discriminating or unfair decisions machine learning experts distinguish three other types of (statistical) bias [2, 5].

1. **Covariate shift** is a type of bias where the training set's distribution is different from the test set. E.g., training the model on a younger population while the test set has an older population.
2. **Sample selection bias** is a flaw in the selection process where non-random data is selected, causing a higher or lower sampling. Eventually having a non-representative sample of the population intended.
3. **Imbalance bias** fewer examples for a certain outcome than the other. More examples of convicts that got their early release application rejected than ones accepted.

2.4 Legal Implications

Most countries have anti-discrimination laws included in their constitution to achieve equality between citizens. Social-cultural structure differs per country, and it is likewise considering the legislation regarding anti-discrimination. Considering the Dutch anti-discrimination law, article 1 of the constitution prescribes that all citizens should be treated equally. This article acts as the foundation for all law books in the Netherlands.

Additional laws in these lawbooks prescribe that equality is for all citizens despite their religion, beliefs, political preference, race, gender, nationality, sexual-orientation either hetero- or homosexual, age, physical or mental disability, chronic and psychological diseases, and form of employment (part- or full-time). Citizens, organizations, and authorities are obliged to adhere to these laws, and even prosecutable by law should compliance fail to happen. Making discriminative remarks, remarks to incite hatred, or participate in activities with the aim of discrimination are punishable by law [10].

Taking these laws into account, a biased AI-DMS is an undesired legal implication for all aforementioned stakeholders. However, due to the novelty of the bias problem in AI-DMS, the complexity and lack of transparency of these systems, law enforcement is a difficult task to achieve. Nevertheless, organizations should design AI-DMS free of bias to overcome these legal implications.

It should be noted that AI models explaining a phenomenon is legally permissible. For instance, investigating causal relations between the fatality of disease outbreaks and population properties like gender, age, ethnicity. It becomes illegal once these predictions are used to decide on the door policy of hospital ICU's. In this situation, predictions lead to unfair decisions against individuals or groups with certain traits, thus establishing inequality between citizens.

2.5 **Governance**

The European Commission (EC) prescribes that data governance should be in place when using personal data for privacy purposes [11]. Its core principle revolves around ensuring the quality and integrity of data, data privacy, protection, and data accessibility. Such guidelines of authorities for AI systems are not yet established. Authors emphasize these should be conceived as well and call for Responsible AI [12]. Responsible AI entails a series of principles, governance being one of them, necessary when deploying AI in applications.

There are two perspectives on governance: 1) committees that review and approve AI development, and 2) leaving the responsibility to employees. Both perspectives could co-exist. However, the first perspective is more likely to decrease an organizations' agility in AI development [12].

The preceding governance perspectives focus on internal organizational activities. Though organizations are required to govern their AI-DMS to comply with laws and regulations, another actor is required to oversee compliance, specifically referring to authorities acting as regulators.

With regulations, laws, and regulators transparency of AI-DMS could be achieved, contributing to trust [12, 13], inducing a wider acceptance of these systems. Arranging internal and external audits to assess compliance is a well-known mechanism in the information technology area. The audit reports should be made available to contribute to the trustworthiness of AI-DMS. An external third-party auditor is necessary for this to succeed [12].

2.6 **Intellectual Property**

With the introduction of governance, organizations might be reluctant to cooperate since transparency could mean revealing the AI systems' source code. This puts

organizations at a disadvantage since other competitive organizations could procure and reuse the source code to their own advantage. Barredo Arrieta et al. [12] argue that the assessment of algorithms, data, and design process contribute to the trustworthiness of AI-DMS. In the assessment process, the authors emphasize that the preservation of these AI systems' intellectual property is necessary. Explainable AI (XAI) methods are considered to be a solution for audit purposes. However, in a recent study, it proved to be a rather challenging ordeal [14]. This is due to the fact that confidentiality could be compromised only by giving access to the input and output of these systems [12]. By acquiring input and output of an AI-DMS, the model could be reverse-engineered through XAI methods. In conclusion, further research in the XAI domain is required to assess bias in AI-DMS while preserving its confidentiality.

2.7 **Delegation Issues**

As a result of the emergence of AI-DMS in recent years, organizations are facing a revision of their decision-making structures. Since AI-DMS serve as actors in the decision-making processes, organizations should consider the role of AI-DMS in these structures. Traditionally decision-making is delegated to managers as actors, a full or partial delegation (hybrid) to AI-DMS are to be considered, with both their own implications.

A full human to AI delegation means leaving the decision-making fully to AI-DMS [1]. The benefits of this approach are a high decision-making speed while processing a large amount of data (not restricted by human capacity), and outcomes could be replicated easily. Limitations are low interpretability due to the algorithms' complex nature, and the design should be carried out thoroughly to prevent bias in the models.

A hybrid delegation is to partially incorporate humans and AI-DMS in the decision-making process. Either the humans contribute at the start or the end, and the AI-DMS on the opposite side of the process. Both alternatives have a low decision-speed due to human involvement. Its interpretability depends on whether the AI-DMS is at the start or the end of the process. At the start of the process means a lower interpretability while at the end means the opposite since humans are involved in the final decision. In both cases, the replicability is low since outcomes are vulnerable to human variability.

2.8 **Debiasing Phases in the Development Process**

Various authors argue that bias mitigation (debiasing) should contribute to fair AI-DMS outcomes [1, 3, 5, 7]. To identify bias in data, designers should be aware of bias types in both technical and non-technical sense [12]. A multi-disciplinary background is therefore necessary. Debiasing data could be reached by firstly identifying the types of bias, secondly determining whether to prune or neutralize the bias [5]. Debiasing could be carried out in three different phases in the development process of AI systems [3, 12]. Here follows the list of debiasing techniques:

Pre-processing

- Learning fair representations: by obfuscating information about sensitive features fair representation are achieved

- Optimized preprocessing: a probabilistic transformation approach which edits features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives
- Reweighing: a technique where sensitive attributes are provided with a weight factor to generate predictions. Another measure in this technique could also be the removal (pruning) sensitive features
- Disparate impact remover: the transformation of features to achieve group fairness

In-processing
- Adversarial debiasing: maximizing a model's accuracy while preventing an adversary's ability to incorporate sensitive features into the predictions. Equality constraints are used to achieve this goal
- Prejudice remover: by adding a discrimination-aware regulation to the learning objective, bias is neutralized

Post-processing
- Equalized odds postprocessing: changing the output target by solving a linear program by finding probabilities to optimize equalized odds.
- Calibrated equalized odds postprocessing: similar to the prior method. However, a calibrated classifier is used in the process.
- Reject option classification: a positive/negative discrimination approach, where the privileged group are foreseen with unfavorable outcomes and the unprivileged group with favorable outcomes. This is done within a bandwidth around the decision boundary to neutralize the gap between the two groups.

In the pre-processing methods, the training data is manipulated to achieve fairness. In-processing methods generate classifiers to cope with bias. Lastly, in the post-processing methods, bias is mitigated in the predictions.

2.9 Fairness and Accuracy

The model's accuracy is an important factor to measure the performance of the model. Different statistical measures could be used to determine a model's accuracy. Debiasing is surely to affect a model's accuracy, since debiasing entails the modification of either the variables, the algorithm, or the target labels. To achieve a fair and accurate model, both measures should be considered in the development process.

2.10 eXplainable Artificial Intelligence

eXplainable AI (XAI) is a research area aiming to grasp the unexplainable character of AI systems, thus achieving transparency and interpretability [5, 15]. XAI consists of various methods to achieve the aforementioned goals [16] without limiting the effectiveness of AI-DMS. Therefore, XAI suggests (1) the generation of more explainable models of AI-DMS while maintaining its accuracy and performance. At the same time, (2) providing humans with understandable decision outcomes, eventually reaching a higher level of trust [12]. In this study, XAI methods are used to explain the model by, for instance: calculating and visualizing feature importance, detect bias in models etcetera.

2.11 **AI Model Development Process**

In this study we propose a generic framework for AI model development. This framework takes the developer step by step through a process that enables bias detection and its mitigation. This process is the result of the integration of additional activities into the usual AI model development process. Bias detection, bias mitigation method selection, and applying the mitigation method are the additional activities. The iterative nature of the process is to achieve acceptable fairness while maintaining the model's accuracy. Since the process is set-up generically, it could be applied broadly (supervised and unsupervised), and its utilization is platform independent. The process is illustrated in figure 3.

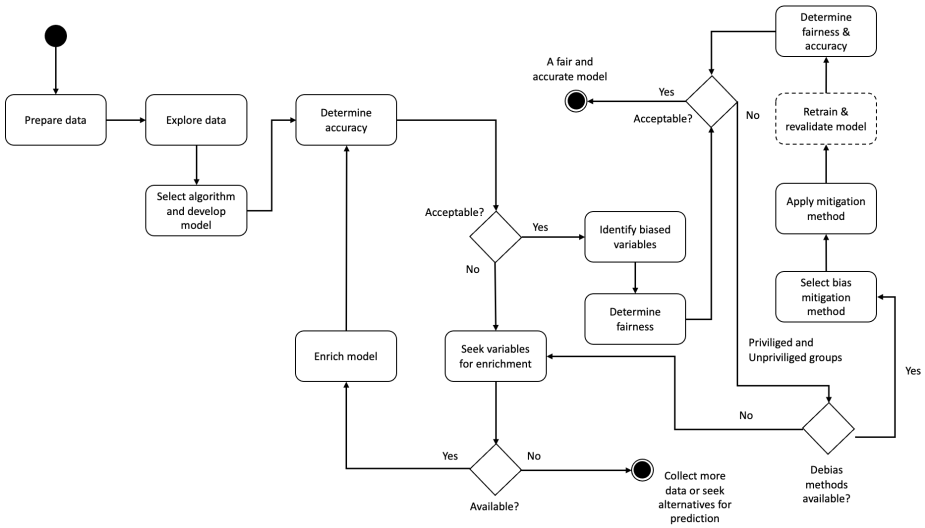

Fig. 1. Machine learning development process with an integration of bias identification and mitigation

1. Data preparation: preparation of data is done in this step. This entails mainly data cleansing and variables transformation. Also, possible biased variables could be uncovered in this step.
2. Data exploration: a data analysis step, where distributions are generated, demographics analyzed and, correlations computed. This step assists in the understanding of the data's emergence in society.
3. Algorithm selection and model development: an algorithm suitable for the data and business case is selected. Number of observations, target variable and whether the data contains solutions (supervised) are criteria for the selection.
4. Accuracy determination: an accuracy computation method is selected depending on the selected algorithm. These methods could be cross-validations, confusion matrices etc.
5. Biased variables identification: using XAI methods to measure feature importance contributing to a prediction, enabling the identification of sensitive features incorporated in the model's prediction. Serves as input for the next step.

6. Fairness determination: fairness between groups using sensitive features are measured using XAI methods.
7. Bias mitigation method selection: a bias mitigation method is chosen to neutralize bias between groups.
8. Bias mitigation method application: the selected mitigation method is applied to the data, model or prediction.
8a. Model retraining and revalidation: an optional activity in case of a bias mitigation method that manipulates the data set.
9. Fairness and accuracy determination: fairness and accuracy are recalculated. In case of an unacceptable outcome step 8 is repeated until there are no mitigation methods available. Otherwise other variables should be sought to enrich the data. Provided that enrichment is impossible, more data should be collected or alternatives for the prediction should be considered.

3 Case Study

Although we initially wanted to test our framework on a variety of data sets, the sensitivity of the topic made that the organizations we approached for their data sets did not want to cooperate. Therefore, we decided to use the COMPAS dataset [17] to illustrate and test our debiasing framework. Since the focus of our efforts was not on the machine learning itself, but rather on detecting and mitigating bias, we chose to use a not overly complex machine learning method, i.e. we developed a logistic regression model using Python and the sci-kit learn package [18]. For model explainability and bias detection, the FairML [19] and AIF360 packages were used, which derive from XAI concepts. With the AIF360 package, bias mitigation is applied. The interested reader can find our source code, the dataset used for this research and data visualization at GitHub [20].

4 Findings

4.1 Data Preparation

The dataset contains data of over 10,000 criminal defendants with 18,316 observations. Enclosed are the defendants' criminal history, COMPAS outcomes and demographics.

Particularly the demographics data consists of sensitive features e.g. sex, ethnicity, and age. First and last names are also included, which could be proxy variables (related to ethnicity and/or religious background).

The dataset was cleansed in this step from data quality issues e.g.: duplicate variables (could cause multicollinearity issues), incomplete records, records with incorrect dates etcetera. After cleansing 14,241 observations remained. Additionally, some transformation between two field dates has been carried out (duration of jail time – days_in_jail). Lastly, the recidivism risk level was transformed into a binary variable, in which the low level is indicated by 0 and the medium and high levels by 1. This variable will be the target variable in the model.

4.2 Data Exploration

To measure the statistical relationship between the variables, Pearson's correlation coefficient was calculated. The strongest positive relations with the target variable are violence score, priors count, the African American race, and whether the defendant is a recidivist.

On the negative correlation side, it is mainly age. There are some weaker relations which could also contribute to the prediction's accuracy in the model. Based on these correlation coefficient values, it seems that the model could be biased should these features be incorporated.

Considering the demographics of the observations in the data the largest group is male. From an ethnicity perspective, African Americans are the largest group, followed by Caucasians and Hispanics. Furthermore, the mean age is 34, the median is 30, and the mode is 20. Most observations are in the age between 19 and 33. The higher the age, the less observations. The distribution for both sexes follows the same trend.

It seems that priors show no visible trend. However, the violence and recidivism score show a similar trend. Both appear to be high in young adulthood and slowly decreasing while the age increases. For the priors, this seems to increase over age, a logical development since a criminal track record is more likely to be higher at an older age.

4.3 Model Development and Accuracy Determination

The logistic regression model was trained by a couple of independent variables, which indicated a correlation with the target variable in the data exploration step. To prevent multicollinearity of the model, a variance inflation factor (VIF) calculation was performed and one of the variables was dropped after dummy transformations. Violence score, priors count, juvenile incident counts, days in jail, an event during imprisonment, age, sex and race were used as independent variables to train the model for recidivism risk predictions.

For accuracy measures, the k-Fold Cross-Validation ($k = 5$) was carried out. Secondly, the accuracy was calculated for the training and test set using the mean accuracy. The cross-validation resulted in an average accuracy of 0.83. A nearly equal outcome (0.82) for the accuracy of the training and test set were computed, which in all cases is an acceptable accuracy for the model. Lastly, a confusion matrix was generated which confirms the prior outcomes. 1,629 (true positives) and 1934 (true negatives) were predicted against 339 (false positives) and 371 (false negatives). From which can be concluded that 83% of the total predictions were correctly predicted.

4.4 Bias Identification and Fairness Determination

In the data exploration step, sensitive features in the demographics were identified. These attributes (age and race) were used as input to develop the model. Subsequently, with the application of an XAI package FairML the feature importance was computed. Fairness was calculated with the AIF360 package of IBM. This method computes the mean difference of the prediction outcome between sensitive features.

Features contributing positively to the model's predictions are violence score, age, African American ethnicity, and priors count. Meaning that a higher age, violence score, and priors count results in higher recidivism risk level. African Americans are more likely to score higher than other ethnicities. On the negative side, it is mainly the male gender and Caucasian and Hispanic ethnicity, meaning that this group is more likely to score lower. Based on this outcome, gender and ethnicity are identified as bias, where females and African Americans are the unprivileged group and males and Caucasians the privileged groups. A minor contributing feature is the Hispanic ethnicity. Indicating that more ethnicities might be privileged compared to the African Americans. The fairness calculation for gender resulted in a mean difference of -0.06 between males and females. For the ethnicity, the fairness gap was larger with a mean difference of -0.26 between African Americans and other ethnicities. Another identified bias is the defendant's age, although this could be defensible in the context of the justice system. Clearly, the model generates biased outcomes that should be eliminated. For bias mitigation, the focus will be on gender and ethnicity.

4.5 Bias Mitigation Selection and Application

Considering the gender bias, a reweighing method was chosen since the fairness gap is small between the groups. Since ethnicity consists of more groups, this attribute was transformed into a binary variable, which indicates whether the ethnicity is African American. Other ethnicities are grouped in one binary value, which results in a partial anonymization bias mitigation method. This also enables unfairness detection between African Americans and other ethnicity groups besides Caucasians and Hispanics, simultaneously eliminating possible bias between non-African American ethnicities.

Reweighing the gender feature resulted in a fairness mean difference of 0.00 between males and females. Indicating that the model is debiased from gender inequality. The accuracy computation resulted in 0.82 in all accuracy calculation methods.

By anonymizing other ethnicities, the unfairness gap for African Americans seems to decrease. Still, this small bias should be eliminated to achieve a fair model. This was done by using the reweighing method again.

After reweighing the race attribute, the model's accuracy appeared to decrease insignificantly to a score of 0.81, with a mean difference of 0.00 between ethnicities. Meaning that the racial bias was eliminated from the model. The reweighing method neutralizes the gap between groups by generating the same weight for feature importance of African Americans as other ethnicity groups. Thus, arriving at an equal opportunity for both groups in the model's predictions.

5 Conclusion and future research

Exploring academic literature resulted in a conceptual framework and integration of an iterative bias identification and mitigation in the AI-DMS model development process. Subsequently, the proposed approach was reviewed by developing an AI-DMS model. The iterative nature of this approach enabled the combination of multiple debiasing mitigation methods in the model. An accurate and fair model was the result of the proposed approach. Bias mitigation methods prove to be powerful in eliminating

unfairness between groups while restricting the effect on the model's accuracy. Avoiding bias and as a consequence sacrificing a model's accuracy however might be demanded by law.

Although it can be argued that it would have been best to not use data sets containing sensitive features, one cannot always prevent them to be included, as the user of these data sets may be unintentionally using proxies that may lead to undesirable bias. So even when features that can be expected to introduce bias are not incorporated in the data sets, one could benefit from our proposed integrated and iterative approach as it provides a clear step-by-step walkthrough to assess whether bias is incorporated in a model and helps the developer to build an accurate and fair model. Making it an instrument for internal audit purposes would provide organizations that develop and exploit machine learning a first line of defense against legal liability claims.

Additionally, this approach enables bias and bias mitigation awareness, which should enable organizations to arrive at an acceptable solution for all stakeholders regarding the bias issue. It is an approach that is platform-independent, meaning that other machine learning tools could be used. Certainly, an understanding of the mentioned concepts is the criterion for the utilization of the approach. The additional framework provides an explanation of relevant concepts that contributes to meet this criterion. Utilization of the proposed framework and approach provides a roadmap for an "unbiased by design" development of AI-DMS models. For our case study, we combined multiple in-processing methods. For future research, other mitigation methods could be combined. How XAI could contribute to model auditing while protecting intellectual property is an interesting direction for future research. Since law enforcement is only achievable if AI-DMS are auditable by an independent party.

We intend to explore whether the proposed framework and process could contribute to an automated bias identification by implementing this approach in machine learning pipelines.

References

1. Shrestha YR, Ben-Menahem SM, von Krogh G (2019) Organizational Decision-Making Structures in the Age of Artificial Intelligence. Calif Manage Rev 66–83. https://doi.org/10.1177/0008125619862257
2. Gu J, Oelke D (2019) Understanding Bias in Machine Learning. 1–12
3. Ntoutsi E, Fafalios P, Gadiraju U, et al (2020) Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdiscip Rev Data Min Knowl Discov 10:1–14. https://doi.org/10.1002/widm.1356
4. Angwin J, Larson J (2016) Machine Bias. In: ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 1 Aug 2020
5. Sileno G, Boer A, van Engers T (2019) The role of normware in trustworthy and explainable AI. CEUR Workshop Proc 2381:
6. Steen M (2020) 'Discussie over de transparantie van algoritmen blijft nodig' | Het Parool. In: Het Parool. https://www.parool.nl/columns-opinie/discussie-over-de-transparantie-van-algoritmen-blijft-nodig~b882ed5f/. Accessed 1 Aug 2020
7. Bolukbasi T, Chang KW, Zou J, et al (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst 4356–4364
8. Datta A, Fredrikson M, Ko G, et al (2017) Proxy Non-Discrimination in Data-Driven Systems

9. Schuilenburg M (2020) Ook politiedata kunnen gekleurd zijn of vervuild - NRC. In: NRC.nl. https://www.nrc.nl/nieuws/2020/07/06/ook-politiedata-kunnen-gekleurd-zijn-of-vervuild-a4005092. Accessed 1 Aug 2020
10. Rijksoverheid (1954) Wettelijk verbod op discriminatie | Discriminatie | Rijksoverheid.nl. In: Rijksoverheid. https://www.rijksoverheid.nl/onderwerpen/discriminatie/verbod-op-discriminatie. Accessed 1 Aug 2020
11. Otto M (European U (2018) Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR). Int Eur Labour Law 2014:958–981. https://doi.org/10.5771/9783845266190-974
12. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al (2020) Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
13. Kroll JA (2018) Data Science Data Governance. IEEE Secur Priv 16:61–70
14. Oh SJ, Schiele B, Fritz M (2019) Towards Reverse-Engineering Black-Box Neural Networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11700 LNCS:121–144. https://doi.org/10.1007/978-3-030-28954-6_7
15. Páez A (2019) The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds Mach 29:441–459. https://doi.org/10.1007/s11023-019-09502-w
16. Nassar M, Salah K, ur Rehman MH, Svetinovic D (2020) Blockchain for explainable and trustworthy artificial intelligence. Wiley Interdiscip Rev Data Min Knowl Discov 10:1–13. https://doi.org/10.1002/widm.1340
17. Ofer D (2017) COMPAS Recidivism Racial Bias | Kaggle. In: Kaggle. https://www.kaggle.com/danofer/compass?select=cox-violent-parsed.csv. Accessed 17 Aug 2020
18. Scikit Learn Developers (2020) sklearn.linear_model.LogisticRegression — scikit-learn 0.23.2 documentation. In: Scikitlearn.org. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Accessed 20 Oct 2020
19. Adebayo J (2017) fairml · PyPI. In: pypi.org. https://pypi.org/project/fairml/. Accessed 13 Oct 2020
20. Ennali Y (2020) yousnali/AI_Bias_Mitigtaion: For this study an integrated and iterative approach for bias mitigation is proposed. In: GitHub. https://github.com/yousnali/AI_Bias_Mitigtaion. Accessed 3 Nov 2020