Artificial intelligence discrimination: how to deal with it?

William Niemiec[1], Rafael F. Borges[1], Dante A. C. Barone[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{williamniemiec,rafaeelfernandes}@hotmail.com, barone@inf.ufrgs.br

Abstract. *The emergence of artificial intelligence has brought many benefits to society through the automation of activities such as driving cars, product delivery, item classification, and predicting trends with a greater degree of accuracy. However, depending on how it is used, it may reflect persistent problems in society, such as discrimination. In this paper, we discuss discrimination by artificial intelligence. We begin by describing this problem and showing that it is a recurring and current problem. Then, we show the origin of this problem and propose a strategy to deal with it in order to prevent it from happening again. Lastly, we discuss future works and how the proposed strategy can be put into practice.*

1. Introduction

Artificial Intelligence (AI) is embedded in the most diverse fields of society, including language, video and voice processing, virtual voice assistants (such as Alexa from Amazon[1]), and even in the control of autonomous cars [Lefevre et al. 2015]. According to Max Tegmark [Discovery 2018], writer and researcher at the Massachusetts Institute of Technology (MIT), AI is the ability of a machine or system to achieve complex goals, i.e., AI is non-biological intelligence. However, despite being a non-biological intelligence, it may contain behavioral problems, which may have a biological origin (such as prejudice) and, consequently, reflect some type of discrimination.

Discrimination is widespread throughout society [Sowell 2019]. It can be interpreted as the activation of some prejudice, being manifested with negative attitudes toward the value of specific social groups. These values can be several things, such as ideas, ideologies, gender, race, among others. Consequently, everything that has the participation of humans is subject to some kind of discriminatory bias, whether intentional or not.

AI algorithms are not neutral, as they can make discriminatory decisions through biases in their systems. Some people believed machine learning systems (a subfield of AI, detailed in Section 3.1) were neutral technologies, where objectivity prevails over subjectivity [Rosa et al. 2020]. However, [Buolamwini and Gebru 2018, Vilarino and Vicente 2021, Silva 2019] showed that machine learning models can present biased behavior and give preference to certain types of groups over others. This was evident in the year 2016 with Google through its web image search service [York 2016]. Searching for ”three white teenagers” showed pictures of white teenagers smiling, but searching for ”three black teenagers” showed sketches of black teenagers who had committed a crime. This sad fact demonstrated that AI can indeed be affected by prejudices from society [Allen 2016].

[1] https://developer.amazon.com/alexa

With the emergence of cases of discrimination committed by AI, many solutions have been proposed in recent years [Suresh and Guttag 2021, Mehrabi et al. 2022]. However, individual solutions are not enough to deal with the problem, given the complexity of it. In this work, we intend to investigate the origin of discrimination by AI as well as propose a strategy to avoid it. We carried out a literature review on occurrences of discrimination by AI - to analyze the cause of discrimination - in addition to analyzing measures that have been adopted to prevent discrimination by AI, in order to summarize such measures and develop an overall strategy.

This paper presents the following contributions:

- We analyze the reasons that lead to discrimination by AI;
- We propose a strategy to deal with discrimination by AI.

The paper is structured as follows: Section 2 discusses the related work. Section 3 presents main concepts used in this work. Sections 4 and 5 describe, respectively, the validation strategy and the results. Finally, Section 6 concludes the paper.

2. Related work

Several techniques have been proposed to prevent AI discrimination in the literature. [Kamiran and Calders 2011] proposes several methods for modifying data, including assigning weights to individuals to balance the data (Reweighting), changing the sample sizes of different subgroups to remove the bias in the data (Sampling), and correcting the labels of some individuals in the data (Massaging). [Dwork et al. 2012] builds a predictive model in which similar individuals should be treated similarly. They achieved this using both individual fairness and statistical parity. In [Luong et al. 2011], individual discrimination is dealt with by putting similar individuals in a cluster. Therewith, discrimination is detected when there is a significantly different decision between the individuals from the protected cluster and the individuals from the non-protected cluster.

All the above works are focused on handling AI discrimination caused by biased data. But as we will see (Section 5.1), AI discrimination can be triggered in model learning, and since we cannot see how a model learned (because it is a black box - see Section 3.2), it is necessary to combine other techniques to address AI discrimination more comprehensively. This work proposes a technique that combines methods to deal with data bias, the black box problem, and what to do when both techniques are not enough to prevent AI discrimination.

3. Background

This section covers the essential concepts used in this work. We first review machine learning concepts and then describe what is discrimination and how it is performed by AI.

3.1. Machine learning

AI has many sub-fields, including one called Machine Learning. It involves computers learning from data provided to carry out certain tasks, being very useful for handling complex tasks, which can be challenging for a human to manually create the needed algorithms. In short, machine learning creates models to learn patterns, through a database, and make decisions without being programmed to do them.

3.2. Black box

According to [Pasquale 2015], black box means a system whose workings are mysterious, i.e., given an input, an output is generated and we cannot see how it was produced (the means that were used to generate this output).

3.3. Discrimination

Discrimination is the manifestation of some prejudice in which a specific social group ends up being the target of negative attitudes. This negative behavior occurs due to disagreement with the characteristics of these groups, whether in ideas, attitudes, or ideologies, originating from prejudice regarding any of these characteristics [Parker 2012].

Although discrimination is related to prejudice, it is important to distinguish them. Prejudice is related to psychological aspects, coming from baseless opinions based on ignorance and preconceived ideas. On the other hand, discrimination is the activation of prejudice concerning some attitude, i.e., while prejudice is something internal to a person, discrimination is the externalization of this prejudice.

Therefore, every discrimination is originated from some prejudice, some of which are considered crimes and, consequently, may have some legal punishment.

3.4. AI discrimination

AI discrimination occurs when unfair predictions are made. When AI algorithms lower one social group concerning another, this algorithm performs discrimination [Olteanu et al. 2019, Caton and Haas 2020]. Although there are people who see AI as a totally objective technology [Rosa et al. 2020], studies have shown that the existing prejudice in society can affect it. [Bissoto et al. 2019] conducted a study on biases of machine learning algorithms for skin cancer prognosis. In this study, they trained two machine learning models: one of them would only have the images as input and the other model would also have access to clinical information about the lesions. The results showed that, despite both models having a high performance in the training stage, for new data, the second model's predictions were worse. This demonstrated that, if unnecessary information is provided, it can deceive and introduce a bias in it. Also, this study showed that, by removing this unnecessary clinical information, the model did not present bias.

Therefore, unnecessary data can introduce bias in machine learning models, and, by removing this data, model performance does not get worse; on the contrary: improve it.

4. Methodology

Our goal is to do a literature review looking for scientific productions reporting facts about discrimination caused by AI in order to analyze its cause and how it would be possible to prevent such discrimination from happening again. To this end, we evaluate the following research questions (RQs):

RQ1: *What are the causes of discrimination by AI?*

RQ2: *Is it possible to develop a strategy to prevent discrimination by AI?*

To carry out the analysis, a literature review strategy was applied. Our goal was to identify scientific works produced in the last twelve years that address the origins of AI discrimination or works that propose a strategy to deal with this problem. Our inclusion criteria for scientific works were the following:

1. the paper addresses AI discrimination;
2. the paper has been published in the last twelve years.

On the other hand, our exclusion criteria were the following:

1. it was not possible to read the complete version of the paper;
2. the paper addresses AI discrimination, but it does not present a strategy for dealing with it, nor does it addresses the origins of this problem.

With this search, we obtained 73 papers. After reading their title and abstract, we applied the inclusion and exclusion criteria, leaving six works: three related to the causes of AI discrimination and the rest with strategies on how to deal with it. All selected papers were read in full so that the situations of discrimination involving AI presented in these publications could be analyzed.

5. Results

In this section, we present and analyze the results with respect to the established research questions. The analyses are presented in two parts: 1) the results of the analyses conducted to answer RQ1 and 2) the results of the analyses conducted to answer RQ2.

5.1. Answering RQ1

AI is widely used in society, and most of the tools to solve these problems are the so-called ”black box” models (Section 3.2). Consequently, AI algorithms are used without understanding how it came to a certain conclusion [von Eschenbach 2021]. Thus, in cases of AI discrimination, detecting the factor that induced it becomes a hard challenge.

Furthermore, there is another problem that contributes to discrimination in an AI application: human bias. Bias has a big influence on the manifestation of discrimination, and because of this, it is common to find the word bias as a synonym for discrimination. However, it is worth noting that not all bias is negative and will result in discrimination. According to [Ferrer et al. 2021], bias is a deviation from the standard, being necessary in some cases to identify the existence of statistical patterns in the data used. Finally, [Ferrer et al. 2021] explains the three most famous causes where bias can be introduced into systems: modeling bias, training bias, and usage bias.

Modeling bias is most often introduced in the selection and manipulation of the dataset to be used in AI training [Ferrer et al. 2021]. It can be introduced by compensating, smoothing or regularization of parameters (algorithmic processing bias), as well as manipulating objective categories to make them subjective (algorithmic focus bias). This bias is more detailed in [Mujtaba and Mahapatra 2019], which states that the bias of the analyst (the human who manipulates the data in data modeling) is transferred to the model with the selection of features used in the model. This happens because some features may not be relevant in the application of the model or may arise because of erroneous data or low reliability, resulting in lower accuracy of predictions for a specific group. Even hiding an unnecessary attribute, the algorithm can still, in some cases, infer it from the other attributes and thus become biased. The example cited in [Mujtaba and Mahapatra 2019], regarding the Amazon hiring application, makes this type of bias very clear. Despite removing the gender attribute from the model, the system gave lower grades to women's curricula when it discovered this characteristic through the list of educational institutions in the curriculum (all-female or all-male college).

Training bias is introduced into the training step when a dataset has some bias. If a dataset is biased, the algorithm will eventually learn to apply this bias [Mujtaba and Mahapatra 2019]. A dataset can have two types of distortion [Ferrer et al. 2021]: the lack of representation of the characteristics of a population (representing an inequality situation) or being distorted through the training labels. The origin of label bias can come from a human who has somehow transferred their biased view (modeling bias) or when an item label contains a vague description of the outcome that can result in unfair predictions. In [Mujtaba and Mahapatra 2019], label bias is explained through an exemplification: a classifier model is chosen to label a candidate as a ”good” or ”bad” hire. In this way, many factors can be obscured by the prediction, and it is suggested to model different ways in which a candidate is considered a ”good” hire.

Finally, usage bias can be seen in two different contexts [Ferrer et al. 2021], interpretation bias and context transfer bias. When we apply an algorithm designed to predict characteristics of a specific population to others than the trained one, we have context transfer bias. Interpretation bias is when the wrong interpretation of the algorithm result leads to discriminatory actions.

Thus, we verified that the bias inserted in the applications and the difficulty of finding and correcting the origin of the bias in the ”black box” models are the main factors (Figure 1) that contribute to most of the discrimination committed by AI.

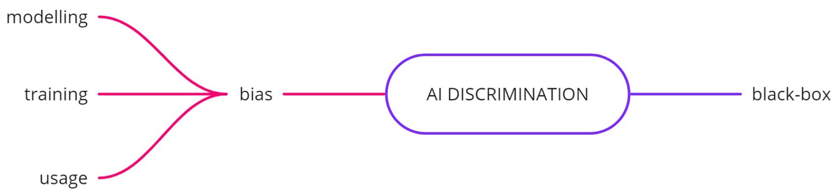

Figure 1. Main causes of AI discrimination

5.2. Answering RQ2

Cleaning training dataset [Hajian et al. 2011] is a method that has shown effectiveness in preventing AI discrimination. In the study, algorithms are used to make training datasets as discrimination-free as possible. The results obtained in this work showed that, when performing cleaning training dataset correctly, there is very little loss of information, and it has a high degree of discrimination prevention. In the worst case of the experiments, the DPD metric (Discrimination Prevention Degree - measure to assess success in preventing discrimination) was 90.90, with the maximum being 100, demonstrating excellent performance in preventing discrimination by AI when performing cleaning training dataset.

Visual analytics system [Sperrle et al. 2019] consists of encapsulating the AI algorithm to bring humans closer to the algorithms. This allows humans to check if there is any discriminatory factor and to take action if necessary. This work proposes the use of the ’human-trust-modeling’ model, which consists of the user’s interaction with AI models gradually so that it is calibrated with the help of the user and avoids being biased. To prevent the user from introducing bias during this process, the work proposes to compare the user’s interaction with the interactions of other agents, whether real or virtual. In this way, it would be possible to detect and avoid possible discrimination by AI that may not have been avoided in the previous step (cleaning training dataset).

An analysis made by [Borgesius 2020] has verified what has been done in the legal field to prevent AI discrimination. In this work, the main data protection laws are analyzed, the largest and most important being the General Data Protection Regulation (GDPR). This is a very strict security and privacy law, and it imposes obligations that companies anywhere in the world have to comply with when handling data from people in the European Union. The GDPR has an article (article 22) that combats discrimination by fully automated AI, providing greater protection to people against possible discrimination that AI can commit against them.

Based on the three strategies mentioned above, we proposed a strategy called DHJ (Data-Human-Juridic). This strategy (Figure 2) encompasses all three intending to further reduce cases of discrimination by AI applying three steps. The ”Data” step proposes applying dataset training cleaning on input data. The next step (”Human”) brings humans to the model learning process to detect possible discrimination committed during the learning process. But these methods do not prevent AI discrimination completely. So, if the two previous steps were not enough to prevent AI discrimination, then discrimination happened. To deal with it, we created the ”Juridic” whose goal is to amortize its effects on people who can suffer from this. This can be done with specific laws that protect people from AI discrimination, such as article 22 of GDPR, applying punishments to responsible organizations and compensating victims.

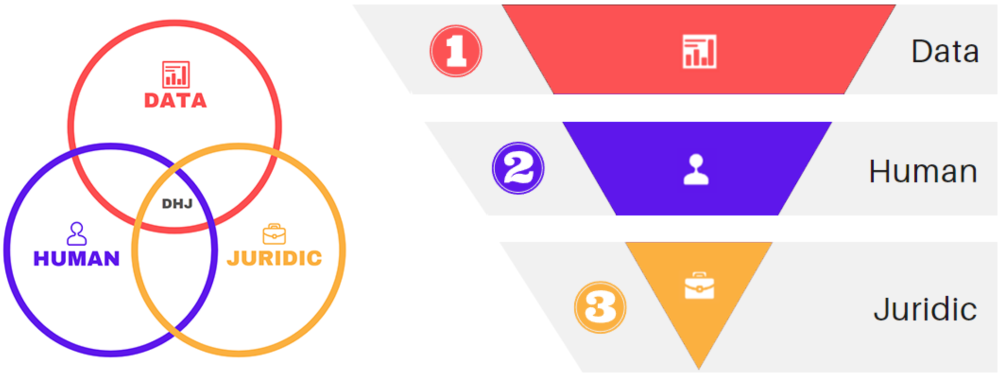

Figure 2. Data-human-juridic (DHJ) strategy

6. Conclusion and future works

Discrimination is in everything that has human participation, including AI. Several works have been developed to prevent AI discrimination, but they could be better performance if used together with other techniques. In this paper, we analyzed the origins of AI discrimination and proposed a strategy to prevent new cases of AI discrimination from occuring. For this, we carried out a literature review to verify the causes of AI discrimination as well as what has been done to deal with it. We verified that the causes are related to the input data, which reflects the discrimination that exists in reality as well as the difficulty in understanding how the machine learning models learned from the data, given that they are seen as a black box by users.

To deal with AI discrimination, we proposed a strategy called DHJ (Data-Human-Juridic), with tries to prevent discrimination by applying three steps: dataset training cleaning (the ”Data” step), bringing humans to the model learning process to deal with the black box problem (the ”Human” step) and, if these methods were not enough to prevent AI discrimination, then discrimination happened. To deal with it, we created the ”Juridic” step whose goal is to amortize discrimination effects on people who can suffer from this applying punishments to responsible organizations and compensating victims. In future works, we hope to apply the proposed strategy and evaluate its performance when put into practice.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This work is also partially supported by Propesq-UFRGS, FAPERGS and CNPq, Brazil.

References

Allen, A. (2016). The “three black teenagers” search shows it is society, not google, that is racist. `https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet`. Accessed: 2022-02-06.

Bissoto, A., Fornaciali, M., Valle, E., and Avila, S. (2019). (de)constructing bias on skin lesion datasets.

Borgesius, F. J. Z. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, 24(10):1572–1593.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey.

Discovery (2018). Discovery brasil — inteligência artificial - ibm. `https://www.youtube.com/watch?v=W95YlM5-iPk`. Accessed: 2022-02-11.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS ’12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Ferrer, X., Nuenen, T. v., Such, J. M., Cote, M., and Criado, N. (2021). Bias and discrimination in ai: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80.

Hajian, S., Domingo-Ferrer, J., and Martínez-Ballesté, A. (2011). Rule protection for indirect discrimination prevention in data mining. In Torra, V., Narakawa, Y., Yin, J., and Long, J., editors, *Modeling Decision for Artificial Intelligence*, pages 211–222, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kamiran, F. and Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33.

Lefevre, S., Carvalho, A., and Borrelli, F. (2015). Autonomous car following: A learning-based approach. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 920–926.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). K-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 502–510, New York, NY, USA. Association for Computing Machinery.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A survey on bias and fairness in machine learning.

Mujtaba, D. F. and Mahapatra, N. R. (2019). Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Parker, R. (2012). Stigma, prejudice and discrimination in global public health. *Cadernos de Saúde Pública [online]*, 28(1):164–169.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Rosa, A., Pessoa, S. A., and Lima, F. S. (2020). Neutralidade tecnologica: reconhecimento facial e racismo. *REVISTA V! RUS*, 21.

Silva, T. (2019). Visao computacional e vieses racializados: branquitude como padrao no aprendizado de máquina. *II COPENE Nordeste: Epistemologias Negras e Lutas Antirracistas*, pages 29–31.

Sowell, T. (2019). *Discrimination and Disparities*. Basic Books.

Sperrle, F., Schlegel, U., El-Assady, M., and Keim, D. (2019). Human trust modeling for bias mitigation in artificial intelligence. In *ACM CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Vilarino, R. and Vicente, R. (2021). An experiment on the mechanisms of racial bias in ml-based credit scoring in brazil.

von Eschenbach, W. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34.

York, C. (2016). Three black teenagers: Is google racist? it's not them, it's us. `https://www.huffingtonpost.co.uk/entry/three-black-teenagers-google-racism_uk_575811f5e4b014b4f2530bb5`. Accessed: 2022-02-05.