International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 13, Issue 1 [January. 2024] PP: 103-106

Understanding and Mitigating Algorithmic Bias in Contemporary Applications

Ljubinko Stojanovic[1] , Vesna Radojcic[1] , Filip Jovanovic[2], Slavica Savic[3], Hasam Mostafa Ahmed Mletan[4]

[1]Faculty of computing and Informatics, Sinergija University, Bijeljina, Bosnia and Herzegovina, [2] Faculty of Project and Innovation Management, Educons, Belgrade, Serbia, [3]Faculty of Technical Sciences, Kosovska Mitrovica, [4]University Singidunum, Belgrade, Serbia
Corresponding Author: Vesna Radojcic

ABSTRACT: *This paper explores the pervasive issue of algorithmic bias in contemporary applications, focusing on its implications across various sectors. The increasing reliance on algorithms and machine learning has introduced a pressing challenge, as these technologies often reflect and perpetuate existing societal biases. Through an in-depth analysis of related studies and real-world examples, the paper sheds light on the far-reaching impacts of algorithmic bias in areas such as healthcare, education, criminal justice, and finance. The study also discusses the challenges posed by algorithmic bias, including its potential to exacerbate social inequalities, exclude marginalized communities, erode trust in algorithms, and invite legal scrutiny. Additionally, the paper reviews ongoing efforts to mitigate algorithmic bias, emphasizing the importance of diverse and representative data, inclusive algorithm development teams, and regular algorithm audits. By addressing algorithmic bias, the paper advocates for the responsible and equitable use of algorithms and machine learning systems.*

Keywords: *Artificial Intelligence (AI), Information Systems, Machine Learning, Algorithms, Social Bias*

Date of Submission: 05-01-2024 Date of acceptance: 17-01-2024

I. INTRODUCTION

The widespread adoption of algorithms and machine learning is transforming numerous sectors, spanning finance, healthcare, education, and even the criminal justice system. These technologies promise to enhance efficiency, reduce costs, and optimize decision-making processes. However, it is crucial to recognize that algorithms employed in these domains are not immune to the biases prevalent in our society. As companies increasingly prioritize data-driven decision-making, a significant challenge comes to the forefront — algorithmic bias [1]. This bias encapsulates systematic errors inherent in the processes and decisions executed by algorithms, with far-reaching consequences. It can lead to the unjust treatment of specific individuals or groups, perpetuating and exacerbating existing social inequalities.

II. RELATED WORK

In this section, we delve into previous research addressing algorithmic bias. The 2019 study titled "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," published in the prestigious journal Science, revealed that an AI algorithm predicting the need for healthcare, utilized by approximately 200 million American citizens, exhibits racial bias [2]. Specifically, the algorithm was tasked with assessing which citizens would benefit from entering a healthcare program in cases of high health risk, allocating additional healthcare to those individuals. The issue arose because the assessment of the need for healthcare, as a proxy measure – a measure of something we cannot directly measure, approximated by something believed to be related – utilized the citizen's previous healthcare system expenditure as a benchmark. The empirical study concluded that Black citizens, even when spending as much on healthcare as White citizens, more frequently incurred expenses for serious interventions, such as diabetes therapy or high blood pressure issues. This could potentially be a consequence of unequal access to healthcare services. Analyzing the group of citizens with high calculated algorithmic risk, it was found that among them, Black citizens had 26.3% more chronic illnesses. In other words, due to the use of a proxy measure that did not accurately reflect the intended measurement of the real need for healthcare, the algorithmic risk measure implicitly favored White citizens in the United States [3]. An excellent illustration of how algorithms have assimilated existing biases in information retrieval processes can be observed in the case of Facebook. In 2019, as reported by Wired, a phenomenon emerged that sparked both amusement and concern [4]. When users entered the search query 'photos of my female friends,' the search platform suggested refining the results with additional phrases such as 'in bikinis' or 'on the beach.' Interestingly, such suggestions were not presented when conducting a similar search for 'photos of my male friends.' This incident highlighted how the search algorithm simply mirrored the real biases and interests of its user base. In a study employing simulation methodology, published in the journal Significance, researchers explored how predictive policing software can become ensnared in a specific feedback loop, resulting in an intriguing form of algorithmic bias. Based on an analysis of historical crime data in specific areas of the city, the software would allocate a higher number of police patrols to those areas. However, the presence of more police forces in a particular territory logically leads to an increase in recorded criminal incidents. When new data is fed into the predictive system, it prompts additional allocation of police forces, creating a vicious cycle and reinforcing the bias that criminal activities appear where they have already been observed. This cycle may overlook the need for predictive allocation of police forces in other parts of the city, where there may be a genuine necessity [5].

Some studies, such as "The Risk of Racial Bias in Hate Speech Detection," have discovered that tweets from the black population, written using distinctive dialects, are twice as likely to be labeled as toxic in common architectures of neural networks trained to recognize hate speech and language toxicity. This stems from the bias inherent in the annotated text corpora on which neural networks are trained [6].

III. EXISTING SOCIAL BIASES

Algorithmic bias often reflects the societal prejudices existing in our community. According to Barocas and Selbst, "the data used to train machine learning systems are often biased, perpetuating the social dynamics and inequalities of the society in which they are produced [7]." For instance, the data used to train algorithms may be biased towards certain groups, resulting in the unfair treatment of others. In criminal justice, for example, data used to train algorithms may be biased towards specific racial groups, leading to unjust treatment of individuals from those groups. Similarly, algorithms used in hiring processes may exhibit bias towards specific education or work experience, resulting in the exclusion of qualified candidates from marginalized communities. It is important to acknowledge the potential unintended consequences they can have, particularly in perpetuating bias and discrimination.

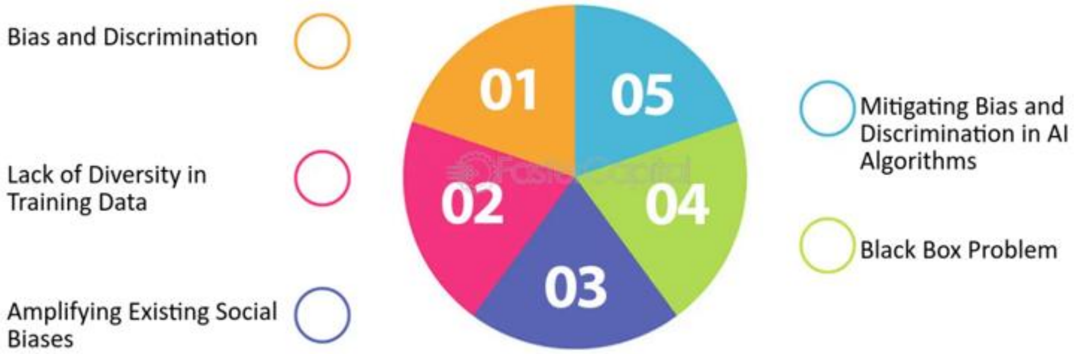
Bias and Discrimination
Lack of Diversity in Training Data
Amplifying Existing Social Biases
01
05
02
04
03
Mitigating Bias and Discrimination in AI Algorithms
Black Box Problem

Fig. 1 Unintended Outcomes of AI Algorithms - Ramifications: Grasping the Limitations of Artificial Intelligence [8]

IV. IMPACTS ON DIFFERENT SECTORS

Social sciences assert that discrimination is the "categorization of different social groups with different positions [9]," which can be used as a justification for the mistreatment of individuals. A distinction is made between direct discrimination, i.e., treating a person based directly on their characteristics (gender, age, appearance, etc.), and indirect discrimination, i.e., treating a person not directly linked to their characteristics but correlated with them. Indirect discrimination is also referred to as systematic or unintentional discrimination [10]. Algorithmic discrimination can occur unintentionally through correlations with a person's characteristics, making the detection of discrimination extremely challenging.

Communication among people has undergone a drastic shift from traditional media such as face-to-face interactions, telephone, television, or print to online social media platforms like Facebook, Instagram, Twitter, etc. Unlike previous traditional media, internet social networks control the information users see and share through filtering algorithms. These algorithms track individual information about user preferences and then filter the data they display based on those preferences. As a result, people tend to be exposed to opinions with which they already agree, leading to the aforementioned algorithmic bias.

This phenomenon, where people are divided into groups with opposing views that rarely intersect, is becoming more common and is known as the "filter bubble" or "echo chamber." The use of algorithmic systems in processes can lead to increased indirect discrimination, as evidenced by ADM systems that can operate based on biased data [11]. In the healthcare sector, physicians are increasingly relying on health algorithms based on mathematical models to assist in diagnosing conditions and making treatment decisions. Artificial intelligence is employed for disease diagnosis, treatment recommendations, risk predictions for health and life, and more. However, the use of such algorithms may, in certain instances, worsen a patient's condition due to inaccurate results, often stemming from the algorithms relying on data from a specific group of individuals. For example, the 2007 VBAC algorithm aimed to evaluate the likelihood of a safe natural delivery after a cesarean section by considering factors such as a woman's age, the reason for a previous cesarean section, and the time elapsed between deliveries. Nevertheless, a 2017 study revealed inaccuracies in the algorithm, particularly in predicting lower chances of successful delivery after a cesarean section for Black, Hispanic, and Latina women, resulting in a higher rate of cesarean sections for these groups compared to white women [12].

Within the educational system, algorithms have gained extensive application in recent years, facilitating tasks like automated grading, statistical calculations for predicting dropout rates, and assessing eligibility for higher education, among others. In the UK in 2020, a manual algorithm was employed to assign grades based on teachers' estimates, revealing a bias as it assigned lower grades to students in state-funded schools and higher grades to those in private schools (sometimes significantly higher than teachers' estimates) [13].

In the financial sector, unlike industries prioritizing fairness as a fundamental principle, there is an ongoing balance between risk and reward. While healthcare systems prioritize individuals based on their needs rather than financial capacity, the banking sector uses data to distinguish between creditworthy and non-creditworthy individuals. This allows financial institutions to roughly gauge the risk associated with a specific user when granting a loan and, occasionally, to set interest rates based on the assumed risk. Enhanced accuracy in these processes contributes to the efficiency of loan approval, thereby increasing the competitiveness of banks.

Discrimination arises when certain groups receive systematic advantages, while others face systematic disadvantages, often resulting from conscious or unconscious biases introduced by individuals developing these algorithms. There are two instances where algorithms can exhibit biases or individual prejudices:

1. Algorithms are written by people, and people come with biases and prejudices.
2. Individuals may introduce biases by using incomplete, inaccurate, or biased datasets for training algorithms (types of biases outlined below).

Sampling bias occurs when one population is overrepresented or underrepresented in the training dataset. An example is a digital credit application where men are more dominant compared to women. If user data is used to train the algorithm, it will rely more on male data than female data. Labeling is the process of marking and classifying individuals based on characteristics and distinctive points to facilitate easier identification through algorithms. For instance, labeling clients' occupations for credit – using "director" instead of "professor." Soon, "rector" and "professor" could become interchangeable for gender bias among clients who applied for credit, while a university worker might face the same bias. Outcome bias occurs when machine learning is poorly defined. For example, if an algorithm uses the residential address as a determinant for predicting the likelihood of loan default. Algorithmic bias refers to systemic and systematic errors that can occur in the processes and decisions made by algorithms.

V. EFFORTS TO MITIGATE ALGORITHMIC BIAS

Several efforts have been made to mitigate algorithmic bias. One approach is to ensure that the data used to train algorithms are diverse and representative of different groups. This can be achieved by using data from various sources and ensuring that the data is not biased towards specific groups. According to O'Neil, "the key to avoiding biased data is to have diverse data and be transparent about how the data was collected and cleaned [14]."

Another approach is to involve people from different backgrounds in the design and development of algorithms. This can help identify potential biases and ensure that algorithms are fair and inclusive. Crawford suggests that "diverse teams are better at identifying potential biases and developing algorithms that are more inclusive". Finally, it is important to regularly review algorithms to ensure that they are not biased toward certain groups. Barocas and Selbst suggest that "algorithmic bias audits can be an effective way to identify and correct errors before they become entrenched."

VI. CHALLENGES

Algorithmic bias poses several challenges. Firstly, it can perpetuate existing social inequalities, leading to unjust treatment of certain individuals or groups. Kleinberg, Mullainathan, and Raghavan argue that "fairness may be in tension with accuracy, so that focusing on fairness may inevitably sacrifice a certain degree of accuracy [15]." Secondly, it can result in the exclusion of qualified candidates from marginalized communities. Buolamwini and Gebru found that "commercial gender classification systems have higher error rates for dark-skinned individuals and women, especially women with darker skin [16]."

Thirdly, it can erode trust in algorithms and machine learning, leading to resistance to their use in various fields. Crawford notes that "algorithmic decision-making is often non-transparent and irresponsible, resulting in those affected by decisions having no way to understand or challenge them [17]." Fourthly, it can lead to legal challenges, as individuals from marginalized communities may contest decisions made by biased algorithms against them.

VII. CONCLUSIONS

Algorithmic bias is a significant issue that needs to be addressed to ensure algorithms and machine learning are used fairly and inclusively. Existing societal prejudices leading to algorithmic bias should be identified and rectified through diverse and representative data, including individuals of different backgrounds, and regular algorithm reviews. By mitigating algorithmic bias, we can ensure that algorithms and machine learning are employed to enhance efficiency, reduce costs, and improve decision-making in a fair and just manner. Several efforts have been made to alleviate algorithmic bias, such as ensuring data from various sources are unbiased toward specific groups. To make progress in this field, it is crucial to engage experts, decision-makers, and the community comprehensively in understanding, recognizing, and addressing algorithmic bias. Only through a collective effort can we create an environment in which algorithms contribute to society in a fair, inclusive, and responsible manner.

REFERENCES

[1]. K. Nima and G. Maryam, "Algorithmic bias: review, synthesis, and future research directions," European Journal of Information Systems, pp. 388-409 DOI: https://doi.org/10.1080/0960085X.2021.1927212, 2022.

[2]. O. Ziad, P. Brian, V. Christine, and M. Sendil, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453 DOI: 10.1126/science.aax234, 2019.

[3]. S. M. Goran, "Izazovi i pretnje algoritamske pristrasnosti," 2 December 2021. [Online]. Available: https://talas.rs/2021/12/02/izazovi-i-pretnje-algoritamske-pristrasnosti/. [Accessed 3 November 2023].

[4]. M. Louise, "A 'Sexist' Search Bug Says More About Us Than Facebook," 2019. [Online]. Available: https://www.wired.com/story/facebook-female-friends-photo-search-bug/.

[5]. L. Kristian and I. William, "To predict and serve? " Signifiance, pp. DOI: https://doi.org/10.1111/j.1740-9713.2016.00960.x, 2016.

[6]. S. Maarten, C. Dallas, G. Saadia, C. Yejin and A. S. Noah, "The Risk of Racial Bias in Hate Speech Detection," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.

[7]. B. Solon and A. D., "Big Data's Disparate Impact," California Law Review, vol. 104, no. 3, pp. 671–732. JSTOR, http://www.jstor.org/stable/24758720, 2016.

[8]. FasterCapital, "Existing Biases," [Online]. Available: https://fastercapital.com/keyword/existing-biases.html.

[9]. K. Alma and O. Carsten, "Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen," Digitale Gesellschaft und Wirtschaft, 2020.

[10]. O. Carsten, "Risks of Discrimination through the Use of Algorithms," A study compiled with a grant from the Federal Anti-Discrimination Agency.

[11]. A. Banu, D. Nancy, and I. Deniz, "Preventing Algorithmic Bias in the Development of Algorithmic Decision-Making Systems: A Delphi Study," in 53rd Hawaii International Conference on System Sciences, 2020.

[12]. D. Vyas, D. Jones, A. Meadows, K. Diouf, N. Nour, and J. Schantz-Dunn, "Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator," in Womens Health Issues, DOI: doi: 10.1016/j.whi.2019.04.007, 2019.

[13]. W. Bedingfield, "Everything that went wrong with the botched A-Levels algorithm," 19 September 2020. [Online]. Available: https://www.wired.co.uk/article/alevel-exam-algorithm.

[14]. O. Cathy, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books, 2016.

[15]. J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in Proceedings of Innovations in Theoretical Computer Science, 2016.

[16]. B. Joy and G. Timnit, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Conference on Fairness, Accountability, and Transparency, 2018.

[17]. C. Kate, "Artificial Intelligence's White Guy Problem," 2016. [Online]. Available: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.