Eliminating Discriminatory Behaviour in Machine Learning - The Search for Objective Data Analysis

Machine learning is a subset of artificial intelligence that, in simple terms, is responsible for analysing complex data structures in order to demonstrate reasoning and decision-making without the intervention of humans. The output behaviour can be exploited by the computer to develop its own reasoning in the future. The powerful nature of this field in scientific research has made it an integral part of our daily life and has considerable uses yet to be discovered. ”Data is the lifeblood of all business.” Machine learning is widely utilised in practically every profession, from healthcare, manufacturing to even financial services. More contemporary technological advancements such as virtual personal assistance, self-driving cars and automated diagnosis detections are only a few examples. However, all scientific phenomena have both advantages and disadvantages, and machine learning is no different; this is where ethical concerns arise. Signs of bias and unfairness appear to have been identified in several algorithms that recommend unobjective proposals. Besides, various statistics demonstrate that the more algorithms based on machine learning are applied in data processing systems, the more imbalances in computations including, for example, racial or gender inequality emerge. So, having chosen an inadequate method to operate with machine learning can result in potentially serious problems. Thus, questions like these still persist: how can computational methods involving machine learning become discriminatory? How should businesses implement such algorithms without the risk of bias?

Bias in AI is oftentimes caused by the lack of variety in data or in already existing prejudice in creators. For example, if we desire to collect data about the average income of people in a certain city, the results received on Monday may differ from those on Sunday, supposing that different groups of people are outside during these times. Even if samples are collected every day of the week, a number of factors may well be overlooked. Furthermore, historical human bias in the form of male dominance has resulted in the emergence of discriminatory behaviour in AI. Ultimately, prejudices against underrepresented groups could potentially arise from previously collected biassed data that was not cleansed from databases. For instance, advertisements aimed at people’s genders had been used by Facebook, allowing targeted job recommendations, which merely led to jobs such as nurse and secretary being offered to women. An algorithm must be trained with the appropriate data for it to learn something. If this data is defective or incomplete, this can result in discriminatory behaviour toward certain groups. For instance, Twitter has developed and trained an AI that can detect hate speech. However, due to the different meanings of words in different language settings, the AI was flagging African American users more often than others. This suggests that the algorithms and the content moderators were unaware of the context of the tweets processed. In short, the data they were working with was insufficient, which impacted the results.

After highlighting the importance of machine learning while recognizing the major faults, setbacks, and biases in the field, it is significant to investigate ways to eliminate the surfaced issues. This negative side of AI is natural: randomly sampled data will inevitably have biases since we live in a biassed world where equal opportunity still ceases to exist. Unfortunately, the approaches might not be completely fool proof and hence cannot eliminate

the problem, however, they are a step forward towards change. Initially, it is significant that we thoroughly understand the root of the problem and why discrimination arises. Numerous applications have been created to detect software bias, one of them being Themis. Themis can be used to generate a test suite to check if, and how much, a software discriminates against race, age, or any other sensitive attributes. Depending on these results, action could be taken by a recruited team which is diverse and is able to provide new and innovative ideas, representative of each group's mentalities. Additionally, it is mandatory that we enforce data governance and regulations by having the algorithms periodically tested by a third party and ensure they are ethical. Finally, another proposed solution, by Soheil Ghili, assistant professor of marketing at Yale SOM is named "train then mask". In this system, sensitive attributes are available to the software while in the training phase but hidden when evaluating new data. Though this method accuracy has only dropped by 0.2%, however, this may not be viable for every situation.

In conclusion, all data is imperfect, and therefore a perfect algorithm cannot be developed with our current knowledge. Although machine learning is undeniably a powerful tool that is widely used in a variety of fields, developers must be aware of bias against underrepresented groups. Additionally, complex data structures ought to be made more transparent to represent all categories equally, and thus eliminate the risk of bias in its early stages. A heavily and rigorously arranged database might answer the problem for a while but would still succumb under the weight of human prejudice as shown in the examples previously mentioned. Further contingencies should be taken into consideration to resolve the underrepresented groups' misinterpretation on Twitter, for instance. Additionally, raising awareness to include anonymity policies for targeted advertisements on Facebook can help respect gender equality on the virtual landscape. Hence, large companies should be exemplary of this change that needs to be made. It is mandatory to fulfil computer logic's imperfections by morally driven human expertise until we achieve fully autonomous machines that adapt to ethically acceptable results. This ensures that previously experienced biases do not reoccur, and that machine learning does not enter a discriminatory cycle. As shown by the examples, the utilisation of algorithmic reasoning using machine learning is somewhat still uncertain. Soheil Ghili, as well as other researchers, emphasise the question: should developers compromise efficiency for impartiality?

**Special thanks go to Zafeiria Chazapi, Ștefan Bud, Martin Damyanov, Emre Gürmeriçliler for their valuable participations to this research paper.*

Section	References
Introduction	https://www.netapp.com/artificial-intelligence/what-is-machine-learning/ https://www.healthcareitnews.com/news/how-ai-bias-happens-and-how-eliminate-it https://www.glennstovall.com/how-software-can-be-racist/ https://www.javatpoint.com/applications-of-machine-learning
Second paragraph	https://www.glennstovall.com/how-software-can-be-racist/ https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter https://towardsdatascience.com/ai-is-flawed-heres-why-3a7e90c48878 https://technologyandsociety.org/bias-and-discrimination-in-ai-a-cross-disciplinary-perspective/ https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/
Third paragraph	https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070 https://www.healthcareitnews.com/news/how-ai-bias-happens-and-how-eliminate-it https://f8federal.com/overcome-and-prevent-ai-bias/ https://dl.acm.org/doi/pdf/10.1145/3106237.3106277 https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ https://insights.som.yale.edu/insights/can-bias-be-eliminated-from-algorithms
Conclusion	-