Sai Teja Bandaru

Ms. Sarah Sowden

Data 605

29 November 2023

Beyond Binary: Exploring Ethical Dimensions of Algorithmic Partiality in AI

Imagine AI as a decision maker, equipped with an impressive amount of knowledge and data that guides its decisions. But what if this wizard's books are flawed, filled with errors and biases from the past? In terms of artificial intelligence and biased data, that is the issue at hand. It is akin to teaching these decision-makers with a textbook that is riddled with mistakes. And when they start making decisions based on this faulty information, they inadvertently perpetuate those mistakes, impacting crucial aspects of our lives, like who gets hired, who gets approved for a loan, or who ends up unfairly treated.

The algorithms depend on data and their results are as good as, or even more so than, the information that has been provided and labelled with a mathematical formula. Even in an unsupervised ML model working with raw data, the machine might find discriminatory societal patterns and replicate them. The computer can be used as a proxy for a human, relinquishing them of any moral responsibility.

The process of data mining, one of the ways algorithms collect and analyze data, can already be discriminatory as a start, because it decides which data are of value, which is of no value and its weight how valuable it is. Previous data, outcomes and the original weight given to a programmer are usually considered in decisions. One example can be when the word woman was penalized, by being given a negative or a lower weight, on a CV selection process based on the data of an industry traditionally dominated by men like the tech industry. The outcome ended discriminating women in the selection process.

Some machine learning models, such as supervised learning, learn from examining past cases and identifying how to label the data according to a classification called training. The discriminatory models of ML can arise from training data that are biased. It is possible to do it two ways. First, an incorrect or unfair valuation is given to a set of prejudicial examples from which the model has been trained or where underrepresented groups are concerned. Second, there is no or incomplete data on training.

The objective of the present paper is to delve into the impact of biased data on algorithmic decision-making and its wider societal implications. It aims to explore the integration of ethical theories in comprehending and tackling algorithmic biases. Moreover, the paper endeavors to offer practical and feasible solutions or strategies to counteract these biases within AI systems.

ALGORITHMIC DECISION-MAKING THAT DISCRIMINATES BASED ON GENDER

In the realm of algorithmic decision-making, there exists an alarming trend gender-based discrimination deeply embedded within these ostensibly objective systems. Despite the facade of impartiality that AI systems often portray, a closer examination reveals a disconcerting reality: the perpetuation of biases that favor or disadvantage individuals based on gender. This bias is conspicuous in pivotal spheres like recruitment processes, loan approvals, and judicial determinations, where AI systems have exhibited tendencies to exhibit gender-based disparities.

The emergence of such a divide has occasioned painful ethical debates, focusing on an intersection between progress in technology and social prejudices. In addition to statistical patterns, the consequences of discrimination based on sex within algorithmic decisions are much broader and encompass key aspects of individual rights and societal equality. These discriminatory results, which echo wider systemic biases, have an impact on fairness and equity. Using the following example, we can gain a clearer understanding of how artificial intelligence demonstrates gender-based discrimination.

Amazon Secret AI Recruiting Tool. Bias in machine learning can be a problem even for companies with plenty of experience with AI, like Amazon. Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers' rate products on Amazon.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those". But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way. That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And according to people familiar with the matter, it has reduced the number of graduates from two all-women's colleges. With a view to neutrality in these conditions, Amazon modified the programs. But that was not enough to prevent the machines from developing other ways of sorting candidates which could be discriminatory.

Although Amazon said it was never implemented, it did not confirm that recruiters had no access to the machine's recommendations. Thus consciously, or unconsciously, affecting the selection process.

ALGORITHMIC DECISION-MAKING THAT DISCRIMINATES BASED ON RACE

In the landscape of algorithmic decision-making, a disconcerting truth emerges: the presence of racial discrimination within these ostensibly neutral systems. Despite the promise of impartiality, artificial intelligence often mirrors and perpetuates societal biases, resulting in discriminatory outcomes favoring or disadvantaging individuals based on race.

The revelation of race-based discrimination within algorithmic decision-making unveils profound ethical quandaries, laying bare the intersection where technological advancement intertwines with systemic prejudices. This phenomenon extends beyond statistical discrepancies; it represents a threat to social fairness, reinforcing system inequality and diminishing confidence in what is supposed to be impartial decision-making processes.

The potential consequences of race-based discriminatory algorithmic decision-making for society are far-reaching. These include the deterioration of existing social inequalities between marginalized groups of people, which hinders progress towards equal opportunities. Bias algorithms make it more difficult to access important resources, such as employment, housing, education, and financial services, which result in economic hardship and immobility for millions of people. In addition, the distrust of artificial intelligence systems and institutions as well as a general lack of trust in fair decision-making processes are undermined by these unequal outcomes. Furthermore, they continue to perpetuate harmful stereotypes and raise legal and ethical concerns about accountability, transparency and the need for regulatory oversight thus requiring a comprehensive strategy in terms of correcting biases and ensuring equitable AI decision making.

COMPAS Risk-Assessment Tool. Correctional Offender Management Profiling for Alternative Sanction, known as COMPAS, is a predictive ML model designed to provide US courts with defendant's recidivism risks scores that oscillate between 0 and 10. Considering 137 variables, such as gender and age and criminal history, with a specific weight assigned to each, it forecasts the likelihood that the defendant will reoffend by perpetrating a violent crime.

It is a risk assessment tool for criminal justice organizations that enables them to carry out their activities and it complements other judicial information systems. Someone with a score of 8 would have twice the rate of recidivism compared to someone with four. Defendants waiting for a trial with a high-risk score are more likely to be imprisoned while waiting for trial than those with low risks, so the consequences of a wrong assessment can be dire. Someone can be wrongly imprisoned while awaiting trial who would not re-offend while a more dangerous individual more likely to offend would be let free.

One such real instance, (Human 1/black male), prior offence: 1 resisting arrest without violence, given a high-risk assessment of 10. Subsequent offences: none. (Human 2/white male), prior offence: 1 attempted burglary, given a low risk assessment of 3. Subsequent offences: 3 drug possessions.

Northpointe, renamed Equivant, the company that created COMPAS, claimed that they do not use race as one of the factors. However, a study of defendants in Broward County, Florida, showed that black individuals are much more likely to be classified as high risk. The same paper indicates that black people who did not re-offend were twice as likely to be classified as high risk compared to a white person as the risk score assessment.

ETHICAL THEORY: STAKEHOLDER IN ALGORITHMIC DECISION-MAKING THAT DISCRIMINATES BASED ON GENDER

A robust framework to understand and deal with algorithmic bias is provided by stakeholder theory when applied to algorithmic decision making which discriminates based on gender in recruitment processes. The ethical theory considers the impact of biased AI systems on various stakeholders involved in or affected by such decisions, and recognizes that their effect is not limited to immediate users. Job applicants, recruitment managers, organizations and society at large are stakeholders in the context of discriminatory algorithms that influence candidates' choice on grounds of gender.

An examination of the effects that bias algorithms have on gender in recruitment processes requires looking at algorithmic bias through a stakeholder's point of view. It includes assessing the impact on gender minorities, women, men, and non-binary individuals applying for jobs. Stakeholder Theory The analysis of how bias algorithms help perpetuate or strengthen existing biases in the recruitment process, creating unequal opportunities and contributing to gender inequality at work has been prompted by stakeholder theory.

ETHICAL THEORY: UTILITARIANISM IN ALGORITHMIC DECISION-MAKING THAT DISCRIMINATES BASED ON RACE

Utilitarianism, a consequentialist ethical theory centered on maximizing overall happiness or utility, offers a critical perspective to scrutinize and mitigate algorithmic bias in race-based risk assessment tools like the Correctional Offender Management Profiling for Alternative Sanction (COMPAS). The analysis of COMPAS in a utilitarian way is to assess the consequences for individuals, society, and criminal justice systems due to biased algorithms. This examination delves into how these algorithms impact individuals of diverse racial backgrounds within the system, probing if biased outcomes lead to unfair treatment or exacerbate disparities in sentencing or parole decisions. Utilitarianism, in the light of the general welfare and fairness of using these algorithms for law enforcement and rehabilitation work, also leads to an assessment of wider societal implications.

To eliminate algorithmic bias in COMPAS via utilitarianism, it is necessary to adopt measures that maximize the welfare and fairness of society. It wishes to assess whether biased algorithms have a detrimental effect on overall social happiness by fostering inequalities, undermining trust in the judicial system and hindering successful efforts of rehabilitation. Ethical solutions within the framework could include a recalibrating of algorithms, increasing transparency and setting up oversight mechanisms to ensure that COMPAS decisions meet principles of fairness and positively contribute to societal welfare. To achieve equitable and just results in the Criminal Justice System, this approach advocates revising training information, carrying out regular reviews of bias against certain races, and implementing fairness indicators.

ACTIONABLE SOLUTIONS

Regulatory Oversight. Advocate for robust regulations mandating algorithmic transparency, fairness audits, and accountability in AI systems used in critical domains like hiring, finance, and criminal justice.

Bias Mitigation Policies. Implement policies that enforce diverse and representative training datasets, ensuring algorithms do not perpetuate biases.

Diversity in Development. Encourage diverse teams in AI development to bring varied perspectives and reduce biases in algorithm creation.

Fairness Metrics. Introduce industry-wide standards for assessing algorithmic fairness, enabling companies to regularly evaluate and mitigate biases.

AI Ethics Training. Offer comprehensive education on AI ethics to developers, policymakers, and users, emphasizing the ethical implications of biased algorithms and ways to address them.

Demand Transparency. Advocate for transparency in algorithmic decision-making processes, encouraging companies and institutions to disclose how AI decisions are made.

CONCLUSION

To conclude, the disturbing fact is that, when examining the complicated interaction of bias and algorithmic decision making, flawed data leading to AI decisions contributes to social biases which affect critical aspects of our lives. The illustrations of gender-based discrimination in recruitment tools and racial biases within criminal justice systems are emblematic of the ethical quandaries facing technological advancements today. These biases, which affect societal fairness, opportunities and confidence in decision making processes, are much broader than statistics mismatches. This paper does not simply identify this problem, but lays down a road map for resolving it. Advocating for regulatory oversight, diverse development teams, transparent decision-making processes, and comprehensive AI ethics training emerges as vital steps towards rectifying biases and fostering equitable AI systems.

Works Cited

Alanis Business Academy. (2014, January 6). *Social Responsibility Perspectives: The Shareholder and Stakeholder approach* [Video]. YouTube. https://www.youtube.com/watch?v=vD9XJKZmXEs

Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI And Ethics*, *2*(4), 771–787. https://doi.org/10.1007/s43681-022-00138-8

CrashCourse. (2016, November 21). *Utilitarianism: Crash course Philosophy #36* [Video]. YouTube. https://www.youtube.com/watch?v=-a739VjqdSI

Dastin, J. (2018, October 10). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Mattu, J. a. L. K. (2023, August 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Vincent, J. (2018, October 10). Amazon reportedly scraps internal AI recruiting tool that was biased against women. *The Verge*. https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report