Bias and Fairness in Machine Learning Algorithms

Name

Institution

Course

Instructor's Name

Date

Bias and Fairness in Machine Learning Algorithms

From recommending movies to determining credit scores or even helping judges make parole decisions, machine learning algorithms are increasingly becoming gatekeepers of opportunity in modern life. Although these technologies are designed to ensure efficiency and objectivity, they are not immune to human weaknesses. As a matter of fact, machine learning algorithms tend to reinforce and even increase the social biases of the data with which they are trained. Consequently, unfair treatment of many individuals occurs not because of their actions, but due to systemic inequalities that infiltrate algorithms under the guise of neutrality. This is a mounting worry that has come to establish bias and fairness in machine learning as one of the top ethical challenges of the digital era. Machine learning bias is caused by biased training data, decision made by humans in the process of designing algorithms, and the absence of agreed-upon definitions of fairness collectively of which needs to be addressed to produce fair and trustworthy AI.

Crucial to the algorithmic bias issue is the data that the systems are trained on, and often, these data reflect their historical and societal biases. Machine learning algorithms only learn patterns based on given data and thus when the data used is biased as in an underrepresentation of a particular group or past discrimination of the group, the machine learning algorithm can act to be biased itself (which causes the same effects or results) (Turner Lee, 2018). There is one notable example of an artificial intelligence recruiting app that was trained on historical data of hiring that were male dominated, and therefore, because of this historical data, the app began to favor male resumes over female without any bias being incorporated. This shows how discriminatory outputs may occur due to biased input and may even occur in a circumstance wherein the algorithm is not written to output a discriminatory result.

Beyond biased data, the very design and development of algorithms can introduce or exacerbate unfairness. There are many decisions developers must make when modeling: what features to add, how to prioritize these features, and how to measure success (de Souza Nascimento, 2019). All of these choices may incorporate biased judgment and implicit preferences. As an example, facial recognition systems have never been able to perform as well on individuals with darker skin colors, which is in no small part due to a lack of adequate diversity in the dataset that the algorithms were trained on. It is an example of how the insufficient diversity in a development team may negatively affect marginalized communities.

Fairness in machine learning is a technical as well as an ethical issue that also involves the whole society. Algorithms such as fairness-aware modeling, re-sampling data, and transparency tools (examples include model interpretability and auditability) can limit this bias, although they cannot be used in all cases (Mohanarajesh, 2024). Furthermore, the developers, stakeholders, and policymakers need to work together to establish a definition of fairness in a specific setting. An example is that given the same healthcare algorithm, fairness could be defined differently as compared to fairness in loan approvals. Well-intended models would not be sufficient to guarantee outcome equity to all users without inclusive dialogues and interdisciplinary monitoring.

To sum up, the problem of bias and fairness in machine learning is determined by a complex interplay of historical prejudices embedded in the data, human subjectivity in algorithm creation, and the lack of contextual and inclusive definitions of fairness. All these problems have a relationship with one another in that biased information feeds skewed models, and the latter can negatively affect underrepresented populations at much higher and disproportionate rates, unless proactive and ethically-based action plans are implemented. The only way that machine learning can deliver on its promise is the ability of society to

address these biases squarely and pledge to establish AI systems that are transparent, inclusive, and accountable, and conform to the held value of fairness and equity. It is only then that we can use technology not only to bring progress, but justice as well.

References

de Souza Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., & Conte, T. (2019, September). Understanding development process of machine learning systems: Challenges and solutions. In *2019 acm/IEEE international symposium on empirical software engineering and measurement (esem)* (pp. 1-6). IEEE.

Mohanarajesh, K. (2024). Develop New Techniques for Ensuring Fairness in Artificial Intelligence and ML Models to Promote Ethical and Unbiased Decision-Making.

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252-260.