Model's ability to be discriminatory

Part 1 of the reading "*Big Data's Disparate Impact*" briefly discusses the factors that can contribute to a data mining algorithm being discriminatory. To begin with, the authors point out that "discrimination can be an artifact of the data mining algorithm itself, rather than a result of programmers assigning certain factors inappropriate weight." While I agree that a data-mining algorithm can be biased as a result of incorrect learning, the authors put the entire onus of discrimination on the algorithm and overlook the human factor that has a role to play in defining the algorithm. For instance, decisions pertaining to the target variable, class labels, data collection, training data, feature selection have human involvement which introduces a certain level of subjectivity and as a result, bias and discrimination.

Further, the reading provides an example of hiring decisions based on academic credentials which tend to put weightage on the reputation of schools. This has been a prevalent scenario in India, where students earning their bachelors from the eminent IITs (Indian Institute of Technology) and MBA from IIMs (Indian Institute of Management) are given preference over students graduating from other colleges. Not only that, some universities in the US are also known to give undue preference to prospective masters/MBA candidates from these colleges. While it is true that it is extremely difficult to get admitted to these institutions, being an IIT or IIM graduate does not imply success in the corporate world. Yet companies have been, for long, discriminating against candidates on the basis of their alma mater.

The article also talks about "redlining". It is a well-known fact that India has a history of caste system. Banks and micro-finance institutions have been discriminating against the lower castes of the society and use it as a general criterion to distinguish between the customers. For example, in one of the studies[1] conducted in India, it was found that female loan applicants from a lower caste, relative to those from a higher caste, had lower odds of receiving loans when their credit scores were below the mean. This was not true for those from upper castes. Additionally, the study also found that those from lower castes with a potential to repay the loans did not have a higher chance of receiving the loan either. As the reading suggests, the above example shows how an algorithm easily discriminates amongst certain sections of a society.

Moreover, the article correctly mentions that the efficacy of data mining is dependent on the quality of training data. Bryson, in his article "*Three different sources of bias and how to fix them*" also mentions about poorly selected training data as one of the three reasons for bias. To give an example, one of my friends used images from various ecommerce platforms to train his machine learning model. However, when this model was used for validation, it failed to identify most of the images even though the training set had similar pictures. The issue here was that images on any ecommerce platform are usually with a white background, while that is not true in the real world. The model did not take into account these uncertainties as the training data did not account for the real-world scenario.

As Bryson mentions, it is not at all surprising that machines can be biased. Simply put, a machine just emulates the bias of its owner. All in all, it is important to avoid our own bias to make way into the initial data engineering steps in order to get a non-discriminatory, unbiased model.

[1] Patel, Lenka, Parida *Caste-Based Discrimination, Microfinance Credit Scores, and Microfinance Loan Approvals Among Females in India*, SAGE Journal (Dec 27, 2020)