Data-Centric Treatment of Bias for Fair Decision Support Systems [Vision]

Agathe Balayn, Christoph Lofi

Delft University of Technology
{a.m.a.balayn, c.lofi}@tudelft.nl

Machine learning is increasingly used for human-centered applications in the context of decision support systems. For example, US judges use machine learning models which assess the probability for someone who committed a crime to re-offend, banks use models that predicts the probability of someone to reimburse a potential loan. One reason for this development is that it is commonly argued that machine learning systems are more objective (unbiased) because of their algorithmic nature, and thus should be fairer than human decisions. However, recent studies are showing that the outputs of these systems might exhibit biases creating unfairness, and that their implementation might not respect ethical requirements (e.g. infringement of privacy, use of "unfair" features). For instance, the model for re-offense risk assessment was accused of being racist because its outputs were biased towards certain sub-populations based on their race.

Most work on fairness in these data-driven decision support systems has been developed until now by the machine learning and data mining communities. However, we assume that the main source of bias and unfairness in the outputs of such systems is due to biases in the training datasets. Hence, we foresee a lot of research challenges that the data management community could tackle in an effort to create fairer systems.

In this presentation, we describe our *vision of a data-centric approach for designing fair decision support systems.* First, we outline the few related works in this domain. Then we discuss our vision, and argue that a central challenge lies in eliciting fairness and bias requirements and transforming them into data constraints for the input and output data. Specifically, we argue for the need of a formal framework allowing the specification of data-centric fairness constraints inspired by data modeling paradigms. Furthermore, we illustrate the need of a framework for automating the testing of these fairness constraints for structured and multimedia data. We propose to develop debugging methods in order to rectify the biased datasets, as well as guidelines for practitioners to initially build unbiased datasets.

In addition, we present current formalisms coming from machine learning research on fairness through a relational dataset perspective, in order to help envision research opportunities. After briefly summarizing the work that other research communities have tackled, we describe in more detail the research challenges to overcome in order to address the main points of the vision we proposed above.

We hope this presentation can foster discussions on the non-functional requirements of such data-driven decision support systems from a data management view.