Summary

Elena Beretta

January 2021

Nowadays, it is widely recognized that algorithms risk to reproduce and amplify human bias that historically have led to discriminatory functioning, especially towards disadvantaged groups. Evidence of such discrimination has been collected and reported in several fields: credit score, allocation problems, criminal justice, advertising, job placement, etc. Solutions to mitigate the effect of biased decision systems focused on metrics to measure the degree of equity of the algorithms and different notions of fairness have been introduced. As a consequence, achieving fairness don't merely involve the process of planning and engineering algorithms that satisfy mathematical and statistical properties. These algorithms indeed should also explicitly encode specific values and equity criteria.

As a result, a significant ethical and political challenge arises for those who are responsible to decide which measures of fairness and which values an algorithm should embody. Several recent studies have drawn attention to this issue related to the implementation of machine learning systems. Evidence emerging from these studies suggests that fairness should be considered as a trade-off process whereby the system background priorities are established. In fact, since the beginning of the first studies on fairness in the field of machine learning, the main challenge has been to define what fairness means: the large number of fairness measurements appeared in the literature is due to this effort, although conciliating different metrics of fairness might be mathematically not achievable, except under constrained special cases. As a consequence, choosing a fairness metric not only involves mathematical aspects or technical requirements the model is supposed to exhibit, but also conditions belonging to moral and political philosophy, as well as issues of human perception of fairness metrics, thereby shifting the focus from purely technical requirements to a multi-facet problem.

In such a context, the primary thesis goal is to investigate the role of fairness and bias in Automated Data-Driven Decision-Making Systems (ADMs). The current work lies at the interface of science, technology and society by offering an wide-ranging interdisciplinary perspective on fairness and bias in automated systems. The discussion about fairness and bias is approached from different perspectives across different application domains. In this vein, four case-studies are provided.

The thesis initially introduces three major Research Problems that constitute the ground on which the whole work is based. More specific Research Questions

are subsequently outlined for each of the case studies.

The first case-study analyses the limitations of the mainstream definition of Artificial Intelligence (AI) as a rational agent, which currently drives the development of most AI systems. In this work, the need of a wider range of driving ethical principles for designing more socially responsible AI agents is drawn.

In the second case-study we propose a method of data annotation based on Bayesian statistical inference that aims to warn about the risk of discriminatory results of a given data set. The method aims to deepen knowledge and promote awareness about the sampling practices employed to create the training set, highlighting that the probability of success or failure conditioned to a minority membership is given by the structure of the data available.

The third case-study a decision-making model to mitigate potential discriminatory effects of ranking systems is presented. We introduce AFteRS, an Automated Fair-Distributive Ranking System, that has the objective of determining the best top-N-ranking in a set of candidates while simultaneously satisfying fairness constraints and preserving the general utility of the system.

Lastly, in the fourth case-study we propose a Decision Support System that aims to ensure long-term fairness. The methodology extends Decision Theory to automated decision-making systems by introducing a theoretical model to apply fairness to a binary partition of the target population. In the spirit of promoting fairer and more effective automated decision systems, the role of individual dynamics in automated decision-making is explored and integrated in our theoretical formalization.

Based on the context, functioning, Research Problems and Questions analyzed throughout the work, and based on the results obtained in the case studies, the thesis ultimately suggests and outlines New Research Trajectories, *Cross-Disciplinary Validation, Multi-High-Interpretability and Systematic Ground Encoding.*