Fairness Challenges in Artificial Intelligence

F

Shuvro Chakrobartty
Dakota State University, USA

Omar El-Gayar
https://orcid.org/0000-0001-8657-8732
Dakota State University, USA

INTRODUCTION

Massive amounts of data and its cheap storage combined with vast computing power, efficient and improved machine learning algorithms to process the data, powered the development of the various Artificial Intelligence (AI) and Machine Learning (ML) applications. However, concern has been raised about the decisions made by these AI systems, especially where the AI system's decisions impact human life. One such example is the recidivism risk prediction tool COMPAS used by the United States Department of Justice, where it has been observed that recidivism risk prediction was biased against black Americans. It should be noted that the COMPAS system used a proxy of prior arrests and friend/family arrests that measured risk for crime (Suresh & Guttag, 2020). Because of cases like this, there has been a focus in the AI research discipline on algorithmic fairness.

Fairness is a highly desirable human value in day-to-day decisions that affect human life. In recent years many successful applications of AI systems have been developed, and increasingly, AI methods are becoming part of many new applications for decision-making tasks that were previously carried out by human beings. Questions have been raised 1) can the decision be trusted? 2) is it fair? Overall, are AI-based systems making fair decisions, or are they increasing the unfairness in society?

Accordingly, this chapter presents a systematic literature review (SLR) of existing works on AI fairness challenges. Towards this end, a conceptual bias mitigation framework for organizing and discussing AI fairness-related research is developed and presented. The systematic review provides a mapping of the AI fairness challenges to components of a proposed framework based on the suggested solutions within the literature. Future research opportunities are also identified. The rest of the chapter is organized as follows: first, AI fairness is elaborated, then a conceptual framework for an AI fairness challenge category and bias mitigation framework is presented. Later, the review protocol applied in the study is described. Following that, the results are discussed. Finally, a set of future research directions and a summary of key findings are described.

BACKGROUND

In recent years discrimination through bias in AI systems has made headlines multiple times across multiple industries. For example, in 2018, Amazon's recruiting algorithm was flagged for penalizing applications that contained the word "women's" (Dastin, 2018). The AI models were trained to vet ap-

DOI: 10.4018/978-1-7998-9220-5.ch101

plicants by observing patterns in resumes submitted to the company over ten years. Amazon's AI system had taught itself that male candidates were preferable because most applications came from men, reflecting the tech industry's male dominance. Bartlett et al. (2021) investigated and found that the mode of lending discrimination has shifted from human bias to algorithmic bias in the USA, where even the online lending backed by algorithmic decision making caused the minority lenders to be charged higher interest rates for African Americans and Latino borrowers.

There is no apparent consensus within the literature as to what the definition of fairness is, and the fairness metrics for any given ML model should be given in each situation (Mehrabi et al., 2021; Verma & Rubin, 2018). This is because defining fairness is not easy, as stakeholders are unlikely to agree on "fair" in different spheres of life. Moreover, something may be deemed fair in one context but may seem unfair in another context. For example, ethnic affinity (one's affinity for a specific ethnic group without identifying their ethnicity) based ads targeting for selling products may not be wrong. However, the same would be deemed unfair when targeting the same ethnic affinity to advertising for credit, housing, jobs, or other opportunities that impacts human life. This would even be illegal if the algorithm could identify a person's actual ethnicity. However, it is understood that fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making (Makhlouf et al., 2021). Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice (Mehrabi et al., 2021). Multiple definitions of fairness and mathematical formulas have been proposed, such as equal odds, positive predictive parity, counterfactual fairness, etc. Verma and Rubin (2018) collected the most prominent definitions of fairness for the algorithmic classification problem explaining the rationale behind these definitions and demonstrated them on a single unifying case study. In real-life scenarios, fairness measurement may not be as simple since no algorithm can pass all of these notions of fairness tests because the critical notions of fairness are incompatible with each other (Kleinberg et al., 2016).

Fairness is closely related to bias, and bias can come from data used to train AI algorithms. An early review article on computer bias by Friedman & Nissenbaum (1996) identified three kinds of biases:

1. Preexisting biases based on social practices and attitudes.
2. Technical bias based on design constraints in hardware and software.
3. The emergent bias that arises from changing the use of context.

These biases remain as a guideline for building fair AI systems. More recently, Mehrabi et al. (2021) categorized biases in data, algorithm, and user interaction identifying the feedback loop that goes from data to algorithm, the algorithm to user interaction, and user interaction producing more biased data that gets fed into the algorithm again. In AI systems, data biases arise from sensitive or protected attributes. Protected attributes define the aspects of data that are socio-culturally sensitive for the application of ML. Some examples of such variables are age, ethnicity, gender, marital status, religion etc. Additionally, these sensitive variables' synonyms (i.e., proxy variables) should also be treated as protected. However, the notion of a protected variable can encompass any feature of the data that involves or concerns human beings. Most approaches to mitigate unfairness, bias, or discrimination are based on the notion of protected or sensitive variables and unprivileged groups (Caton & Haas, 2020). The dimension of fairness ensures that algorithmic decisions do not display unjust or biased behavior concerning sensitive or protected attributes. It accounts for the ethical and legal risk of discrimination against specific collectives or minority groups (Unceta et al., 2020). Further, ML algorithms operate by learning models from historical data and generalizing them to unseen new data (Suresh & Guttag, 2020). The increased