Chapter 2
Artificial Intelligence Accountability in Emergent Applications:
Explainable and Fair Solutions

Julia El Zini
https://orcid.org/0000-0001-7499-8668
American University of Beirut, Lebanon

ABSTRACT

The rise of deep learning techniques has produced significantly better predictions in several fields which lead to a widespread applicability in healthcare, finance, and autonomous systems. The success of such models comes at the expense of a trackable and transparent decision-making process in areas with legal and ethical implications. Given the criticality of the decisions in such areas, governments and industries are making sizeable investments in the accountability aspect in AI. Accordingly, the nascent field of explainable and fair AI should be a focal point in the discussion of emergent applications especially in high-stake fields. This chapter covers the terminology of accountable AI while focusing on two main aspects: explainability and fairness. The chapter motivates the use cases of each aspect and covers state-of-the-art methods in interpretable AI and methods that are used to evaluate the fairness of machine learning models, and to detect any underlying bias and mitigate it.

INTRODUCTION

The popularity of AI systems is motivated by the rise of deep learning models which have demonstrated significant performance gains in a plethora of areas. However, such models are seen as extremely opaque models whose predictions are notoriously hard to explain. Accordingly, the emergence of deep models brings to the front the trade-off between their accuracy and their accountability. In high-stake areas, AI accountability is of critical importance. For instance, healthcare systems require fair treatment of individuals regardless of their skin color, gender, or sexual orientation. Insurance applications must explain

DOI: 10.4018/978-1-6684-6937-8.ch002

their decision-making system to engender users' trust. Additionally, autonomous driving systems should deliver acceptable safety standards and legal guarantees on the rights, duties, and responsibilities of the user. Consequently, there has been increased attention recently dedicated to studying and enforcing the accountability of such models. This research is manifested in developing methods to ensure the proper functioning of AI systems through their design, development, and deployment phases.

These concerns led The US Federal Trade Commission to issue new guidelines requiring AI systems to be open, explainable, and fair. Moreover, the General Data Protection Regulation (GDPR) of the European Union mandates transparency for algorithms and fair representation and treatment in AI systems. Whether or not they operate in the European Union, industries that develop and use data-driven systems are moving into ensuring these regulations. That being the case, data and algorithmic accountability witnessed explosive growth mainly nurtured by the invasive use of autonomous systems and the regulations imposed by legal institutions on data and smart processes. Governments started to make sizeable investments in responsible and accountable AI systems. Researchers are extensively engaged in the fields of accountability, fairness, and explainability (Gade et al., 2019; Mehrabi et al., 2021). This is reflected in developing methods to explain AI decisions and learned representations for different data types. Additionally, researchers are working on providing fairness definitions and bias detection methods in numerous applications. This is mostly accompanied by several techniques to neutralize learned representations and mitigate bias in decision-making systems.

Covering all aspects of AI accountability is beyond the scope of this book. However, the nascent subfield of accountability should be an integral part of the discussion on any emergent AI application. This chapter presents a comprehensive study of critical areas that are moving into adopting AI-based solutions and integrating accountability guarantees. These requirements entail a transparent decision-making scheme and fair treatment of individuals.

This chapter focuses on the two aforementioned accountability aspects: explainability and fairness in AI applications and their fundamental interconnection. Explainability requires a meaningful explanation of AI's logic in reaching a decision concerning their data. This explanation should be clear, concise, and easily comprehensible format. Fairness ensures that AI systems handle individual's data fairly. This requires that AI systems do not generate outcomes that could negatively impact marginalized groups. Even if AI systems are not created with detrimental goals, fairness ensures that these systems do not unintentionally learn historical and social discrimination from unfair datasets. This chapter discusses state-of-the-art methods of explainable AI on different modalities and applications while highlighting different notions of algorithmic fairness and its applicability in different settings.

To set the expectations of the reader, the next section highlights the importance of explainability and fairness with the current breakthroughs in AI, Deep Learning (DL) specifically. The section also categorizes explainable AI methods according to (1) whether explainability is introduced into AI models before or after design and (2) the type of representations or decisions they explain. This categorization underlines the interdependency between the explainability aspect of AI systems and the detection of any underlying bias. The third section is devoted to cover the need for explainable AI in emergent applications while briefly covering state-of-the-art methods and discussing a novel line of work within Explainable AI (ExAI), counterfactual explainability. Lastly, the fourth section covers the bias sources in AI systems, the important fairness definitions in the literature, and successful ways to neutralize models and mitigate bias before concluding with final remarks.