Chapter 4
A Survey on Explainability in Artificial Intelligence

Prarthana Dutta
National Institute of Technology, Silchar, India

Naresh Babu Muppalaneni
National Institute of Technology, Silchar, India

Ripon Patgiri
National Institute of Technology, Silchar, India

ABSTRACT

The world has been evolving with new technologies and advances everyday. With learning technologies, the research community can provide solutions in every aspect of life. However, it is found to lag behind the ability to explain its prediction. The current situation is such that these modern technologies can predict and decide upon various cases more accurately and speedily than a human, but has failed to provide an answer when the question of "how" it arrived at such a prediction or "why" one must trust its prediction, is put forward. To attain a deeper understanding of this rising trend, the authors surveyed a very recent and talked-about novel contribution, "explainability," which would provide rich insight on a prediction being made by a model. The central premise of this chapter is to provide an overview of studies explored in the domain and obtain an idea of the current scenario along with the advancements achieved to date in this field. This survey aims to provide a comprehensive background of the broad spectrum of "explainability."

INTRODUCTION

Researchers across generations have witnessed advancements in technology in every field of life. John McCarthy, in 1956 coined the term "Artificial Intelligence" (AI) (McCarthy, 1989), and even before that, it has been evolving as an elusive subject of concern in many research activities (Turing, 2009). Artificial Intelligence has been witnessing its utility in various fields for decades and attained desired

DOI: 10.4018/978-1-7998-7685-4.ch004

and satisfactory achievements with the applications of Machine and Deep Learning models (Iqbal & Qureshi, 2020; Jaouedi *et al.*, 2020; Tabassum *et al.*, 2020; Rani & Kumar, 2019), etc. There have been tremendous improvements in a wide range of domains such as commercial, medical, and marketing platforms, etc. Artificial Intelligence has paved the way such that no human intervention is needed in most aspects of decision-making. With more and more advancements in learning technologies and models, the decisions or predictions made by these systems are accurate approximately 100%. Also, machine learning algorithms can correctly predict, and their accuracy may rise to 100% in some favorable cases. Patgiri *et al.* (Patgiri *et al.*, 2019) reports an accuracy of 100% in conventional machine learning algorithms. Even though with such high and satisfactory accuracies, these technologies may face trust issues related to the model and the predictions (Patgiri *et al.*, 2019; Ribeiro *et al.*, 2016). Hence, the main risk that is associated with these predictions is that even though the model can predict accurately for all the test cases, it might not work in the same accurate manner when the model is left in the wild in their respective domains to track and make life-changing decisions (some of them are discussed in other sections). Thus a gap exists between the prediction or decision made by an intelligent system and the reason associated with these decisions or predictions.

For this, the researchers have been thinking of taking their research one step ahead by picturing if the system gave its end-user a valid reason for making such decisions or predictions. Thus they developed the idea that, along with making the predictions, if the intelligent systems (making these predictions) are also able to give a proper explanation of its prediction, it would prove to be much easier and more meaningful for end-users to rely upon and trust the systems and take further actions based on these explanations.

Thus, bridging the gap between the prediction and the reason for such a prediction for the end-users to understand and interpret better is what we call an "*explainable system*." The mechanism inbuilt in the model for explaining its decision is called "*Explainability.*" The motivation behind this survey is to provide new interested researchers a clear picture of the current scenario of explainability in learning technology. This chapter is intended to bring into limelight the recent trends which the research community is following to deal with explainability.

BACKGROUND

The term "eXplainable Artificial Intelligence" is usually abbreviated as XAI. The term was first formulated by Lent *et al.* in 2004 (Van Lent *et al.*, 2004). Before that, it was addressed simply as a "black-box."

Digging deep into the matter, it is found that explainability is not a recent topic that the researchers have been focusing upon. It has been discussed for many years and have recently gained more attention and importance due to their ability to explain the predictions made by the intelligent systems, which is indeed a major concern. Recently Das *et al.* provide an extensive survey of explainability employed in Deep Learning. They analyzed the various methodologies, algorithms along with the limitations (Das & Rad, 2020). Various researchers define "Explainability" in several ways. Ribeiro *et al*. described explainability as "presenting textual or visual artifacts that provide a qualitative understanding of the relationship between the instance's components (e.g., words in a text, patches in an image) and the model's prediction" (Ribeiro *et al.*, 2016). Whereas Guidotti *et al*. quotes, "an explanation is an 'interface' between humans and a decision-maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans" (Guidotti *et al.*, 2018). Another group of researchers, Gilpin *et al*., defined it in terms of "models that are able to summarize the reasons for neural network behavior, gain the trust