On the Need of an Explainable Artificial Intelligence

Cecilia Zanni-Merk^(✉)

Normandie Université, INSA Rouen, LITIS, St-Etienne-Du-Rouvray, France
`cecilia.zanni-merk@insa-rouen.fr`

Abstract. This plenary talk will explore a fascinating nex research field: the Explainable Artificial Intelligence (or XAI), whose goal if the building of explanatory models, to try and overcome the shortcomings of pure statistical learning by providing justifications, understandable by a human, for decisions or predictions made by them.

Artificial Intelligence (AI) applications are increasingly present in the professional and private worlds. This is due to the success of technologies such as machine learning (and, in particular, deep learning approaches) and automatic decision-making that allow the development of increasingly robust and autonomous AI applications. Most of these applications are based on the analysis of historical data; they learn models based on the experience recorded in this data to make decisions or predictions.

However, automatic decision-making by means of Artificial Intelligence now raises new challenges in terms of human understanding of processes resulting from learning, of explanations of the decisions that are made (a crucial issue when ethical or legal considerations are involved) and, also, of human-machine communication.

To meet these needs, the field of Explainable Artificial Intelligence has recently developed.

Indeed, according to the literature, the notion of intelligence can be considered under four aspects: (a) the ability to perceive rich, complex and subtle information, (b) the ability to learn in a particular environment or context; (c) the ability to abstract, to create new meanings and (d) the ability to reason, for planning and decision-making.

These four skills are implemented by what is now called the Explainable Artificial Intelligence (or XAI) with the goal of building explanatory models, to try and overcome shortcomings of pure statistical learning by providing justifications, understandable by a human, for decisions or predictions made.

During this talk we will explore this fascinating new research field.