Dialectical Explanations

Francesca Toni

Department of Computing, Imperial College London, UK
ft@imperial.ac.uk

Abstract. The lack of transparency of AI techniques, e.g. prediction systems or recommender systems, is one of the most pressing issues in the field, especially given the ever-increasing integration of AI into everyday systems used by experts and non-experts alike, and the need to explain how and/or why these systems compute outputs, for any or for specific inputs. The need for explainability arises for a number of reasons: an expert may require more transparency to justify outputs of an AI system, especially in safety-critical situations, while a non-expert may place more trust in an AI system providing basic (rather than no) explanations, regarding, for example, items suggested by a recommender system. Explainability is also needed to fulfil the requirements of regulation, notably the General Data Protection Regulation (GDPR), effective from May 25, 2018. Furthermore, explainability is crucial to guarantee comprehensibility in human-machine interactions, to support collaboration and communication between human beings and machines.

In this talk I will overview recent efforts to use argumentative abstractions for data-centric methods in AI as a basis for generating dialectical explanations. These abstractions are formulated in the spirit of argumentation in AI, amounting to a (family of) symbolic formalism(s) where arguments are seen as nodes in a graph with relations between arguments, e.g. attack and support, as edges. Argumentation allows for conflicts to be managed effectively, an important capability in any AI system tasked with decision-making. It also allows for reasoning to be represented in a human-like manner, and can serve as a basis for a principled theory of explanation supporting human-machine dialectical exchanges and conversations.

Keywords: Explanation · Argumentation · Conversational AI