Argumentation-based justification and explanation in AI ethics

Beishui Liao

For an autonomous system, the ability to justify and explain its decision making is crucial to improve its transparency and trustworthiness. In this talk, I will first discuss some basic notions of explainable AI and ethical AI, and then introduce an argumentation-based approach to represent, justify and explain the decision making of a value driven agent (VDA). By using a newly defined formal language, some implicit knowledge of a VDA is made explicit. The selection of an action in each situation is justified by constructing and comparing arguments supporting different actions. In terms of a constructed argumentation framework and its extensions, the reasons for explaining an action are defined in terms of the arguments for or against the action, by exploiting their defeat relation, as well as their premises and conclusions.