Human-Generated Explanations of Inferences in Bayesian Networks: A Case Study

Marko Tesic
Birbeck, University of London, London, United Kingdom

Ulrike Hahn
Birkbeck, University of London, London, London, United Kingdom

Abstract

As AI systems come to permeate human society, there is an increasing need for such systems to explain their actions, conclusions, or decisions. This is presently fuelling a surge in interest in machine-generated explanation. However, there are not only technical challenges to be met here; there is also considerable uncertainty about what suitable target explanations should look like. In this paper, we describe a case study which makes a start at bridging between machine reasoning, and the philosophical and psychological literatures on what counts as good reasoning by eliciting explanations from human experts. The work illustrates how concrete cases rapidly move discussion beyond abstract considerations of explanatory virtues toward specific targets more suitable for emulation by machines. On the one hand, it highlights the limitations of present algorithms for generating explanations from Bayesian networks. At the same time, however, it provides concrete direction for future algorithm construction.