Formal Methods in Computer-Aided Design 2023

Formal Methods for Trused AI

Bettina Könighofer
Graz University of Technology
Graz, Austria
bettina.koenighofer@tugraz.at

***Abstract*—The enormous influence of systems deploying AI is contrasted by the growing concerns about their safety and the relative lack of trust by the society. This talk will focus on a few aspects of trustworthy AI: safety, accountability, and explainability. First, we will discuss recent work on evaluating safety for systems deploying deep learning, and correct-by-construction runtime assurance methods to enforce safety during runtime (aka shielding). For accountability, we outline the potential of formal computing tools to analyse the decisions of autonomous agents and to assign responsibility. Finally, we approach explainability from the automata learning perspective. We will discuss recent automata learning approaches which are able to learn compact probabilistic models for high-dimensional environments and outline how learned environmental models can effectively be used to understand and to evaluate the decisions of the agent.**

https://doi.org/10.34727/2023/isbn.978-3-85448-060-0_3