DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Aspects of trustworthy Reinforcement Learning

Ann Nowé

The increasing power of Reinforcement Learning techniques, in particular Deep RL, comes with an increase in complexity, and the traditional convergence guarantees no longer hold. Moreover, the policies learned express more complex behaviour. While transparency and explainability of machine learning models have recently received quite some attention, this is much less explored in the context of Reinforcement Learning. In this talk I will discuss the different aspects to this problem and provide an overview of the state-of-art, including visualisation and some of our own work on policy distillation and formal guarantees.