Trustworthy Reinforcement Learning: Opportunities and Challenges

Ann Nowé
Vrije Universiteit Brussel
Brussel, Belgium
ann.nowe@ai.vub.ac.be

ABSTRACT

Reinforcement Learning (RL) has long outgrown the traditional representations that guaranteed policy convergence but severely limited its application to complex domains. Modern Deep RL enables far richer and complex behaviour, yet at the cost of transparency and explainability. While these latter issues have recently received much attention in Machine Learning, they are underexplored in RL. In this talk, I will discuss them from multiple angles, survey state-of-the-art approaches, including recent developments in policy distillation and formal guarantees, and touch upon the related question of fairness.

KEYWORDS

Reinforcement learning, Trustworthiness

ACM Reference Format:
Ann Nowé. 2024. Trustworthy Reinforcement Learning: Opportunities and Challenges. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 1 page.

1 BIOGRAPHY

Ann Nowé is professor of Computer Science and director of the AI Lab at the Vrije Universiteit Brussel (VUB). Her main research interest is Reinforcement Learning (RL), including Multi-Agent and Multi-Objective RL. She strongly believes in the interplay between theory and applications. Her team has developed novel algorithms and tested them in domains such as smart grids, communication networks, mechatronics and scheduling problems. Ann Nowé is a former board member of EurAI and chairman of the BNVKI, and a current board member of IFAAMAS. She was PC co-chair of AAMAS'21 and general chair of ECAI'23 and EWRL'23. Recently she was elected as an EurAI fellow.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).