SCIENCES &
BIOENGINEERING SCIENCES

Explainable AI - Towards Explainable Reinforcement Learning

Promotor(s)

Ann Nowé ✉ ann.nowe@vub.be

Advisor(s)

Hélène Plisnier ✉ helene.plisnier@vub.be

Research Lab

AI

Context

Reinforcement learning (RL) is a machine learning technique capable of handling difficult sequential decision problems as well as control problems. In recent years, the development of powerful RL techniques has assured that **RL will become an indispensable component in the industry** (e.g., manufacturing, electric power systems and grid management). Additionally, RL-based solutions for applications such as (semi-)autonomous cars, socially assistive robotics, household solar storage management, also means that **RL will find its way into daily human activities**. Recent successes like AlphaGo, prove that RL's time has come. However, any powerful RL technique is built around complex function approximation procedures that transforms this framework in a **black box approach**. This may prevent its application in future critical domains, considering the upcoming European General Data Protection Regulation (GDPR). We believe it is a crucial moment to shift research efforts into **creating an explainable RL framework**. This will be realised by opening the learning process to the user, by explaining what has been learned, how the learning progressed and how the knowledge was applied. The approach taken is to augment existing state-of-the-art reinforcement learning techniques, such that no loss in learning capacity is inflicted.

Until now, insights in the learned policy were typically limited to showing a typical trace, i.e., an execution of the policy or through analysing and visualizing agent reward structures (1). Relevant efforts to mention are (3), that visualises action versus parameter exploration and (4) that analyses and visualizes the usefulness of heuristic information in a state space. Relational RL (RRL) could be seen as an interpretable RL approach "by design" (2) as it allows to express the policy as well as the quality of the policy using relational expressions. Unfortunately, RRL never scaled up because it is extremely sample expensive.

A comprehensive approach to Explainable RL is currently not existing. Therefore, there is a need to build upon the powerful existing state-of-the-art RL techniques, and to make them more interpretable through the introduction and exploitation of **meta-information**.

Research Objectives

The main challenge of Explainable RL is to develop a methodology to make RL, even when applied to very complex environments, understandable for users.

- Develop an approach that allows to **communicate a learned strategy to a user**. Representations used by RL might become very complex in order to successfully learn the optimal strategy. A careful a posteriori simplification and communication of the strategy is thus needed.
- Implement an approach allowing the RL agent to **explain the actual execution of a policy**. For example, to what extend does the policy unfolds as expected? Is the expected reward at stake?
- Augmenting RL with the necessary meta-information**meta-information**.
- Allow the user to **guide the policy being learned**. Unlike the current reward shaping and demonstration approaches that are oriented towards speeding up the learning, without losing the guarantees of convergence to the optimal policy, can we allow a limited reduction in optimality for the sake of capturing the user's intuition, preference or tacit knowledge, to obtain a policy that feels more natural to the user?
- Ground the above interactions in **natural language**.

Research Position

4 years of research full-time, with some occasional Teaching Assistant tasks.

References

[1] Agogino, A. K., Tumer, K. (2008). Analyzing and visualizing multi-agent rewards in dynamic and stochastic domains. Autonomous Agents and Multi-Agent Systems, 17(2), 320-338.

[2] Tadepalli, P., Givan, R., Driessens, K. (2004). Relational reinforcement learning: An overview. In Proceedings of the ICML-2004 Workshop on Relational Reinforcement Learning (pp. 1-9).

[3] Rückstiess, T., Sehnke, F., Schaul, T., Wierstra, D., Sun, Y., Schmidhuber, J. (2010). Exploring parameter space in reinforcement learning. Paladyn, 1(1), 14-24.

[4] Brys, T., Nowé, A., Kudenko, D., Taylor, M. E. (2014). Combining Multiple Correlated Reward and Shaping Signals by Measuring Confidence. In AAAI (pp. 1687-1693).