Explainable Reinforcement Learning *

Amber E. Zelvelder[1], Amro Najjar[1], and Kary Främling[1]

Umeå University, Umeå , Sweden {amberz, najjar, framling}@cs.umu.se

Abstract. As has been shown by recent empirical user studies, its in human nature not to trust decision that we dont understand, regardless of whether they were made by other humans, or an Artificial Intelligence(AI). While we have learned how to evaluate the intent of other humans by their manner or qualifications, most people are unfamiliar with how AIs make their decisions and thus most of these people feel anxious about AI decision-making. A result of this is that AI methods suffer from trust issues and this hinders the full-scale adoption of them.

Machine Learning (ML) algorithms are a type of AI that predicts outcomes without being explicitly programmed. Goal-driven Machine Learning based on the natural learning progress is known as Reinforcement Learning (RL)[5]. The natural learning process is simulated in these algorithms by providing rewards if the desired state is reached or punishments when a wrong state is reached. Using this system, RL algorithms try to optimise the total reward, or reward over time.[2][3]

To make RL less opaque, we plan to use Explainable AI (XAI) methods [1][**?**][4]. XAI refers to AI and Machine Learning techniques that can provide understandable justifications and interpretations for their behaviour. If the AI can be clear about the reasons for its actions, this would help build rapport, confidence and understanding between the AI agent and its human operator, thereby increasing the acceptability of the systems , solving any ethical consideration raised by lack of transparency of decisions, meeting recent legal obligations, and enhancing end-user satisfaction.

In this research project, we try to determine what the main application domains of RL are, and to what extent research in those domains has explored explainable solutions. In addition, we seek to establish the key factors to determine the need for explainability that can be used as a guideline for increasing the explainability for RL algorithms, and with that also increase the trust people can put in RL. Building on that, we will then assess the efficiency and viability of a selection of RL methods. This is followed by doing the same for a selection of XAI methods. Finally we will try to optimise the RL and XAI methods to work together and define a place for them individually or as a combination within the larger space of Machine Learning and AI.

* This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

References

1. Anjomshoae, S., Najjar, A., Calvaresi, D., Framling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proc. of the 18th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2019)
2. Barto, A., Thomas, P., Sutton, R.: Some Recent Applications of Reinforcement Learning. Workshop on Adaptive and Learning Systems (2017)
3. Främling, K.: Light-Weight Reinforcement Learning With Function Approximation for Real-Life Control Tasks. Proceedings of the 5th International Conference on Informatics in Control, Automation and Robotics, Intelligent Control Systems and Optimization (ICINCO-ICSO) (2008), http://www.cs.hut.fi/u/framling/Publications/KF_ICINCO2008_final.pdf
4. Hendricks, L.A., Akata, Z., Rohrbach, M., Schiele, B., Darrell, T.: Generating Visual Explanations
5. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction - second edition. The MIT Press, 2 edn. (2018)