# *Preference Based Reinforcement Learning*

Researchers
Christian Wirth

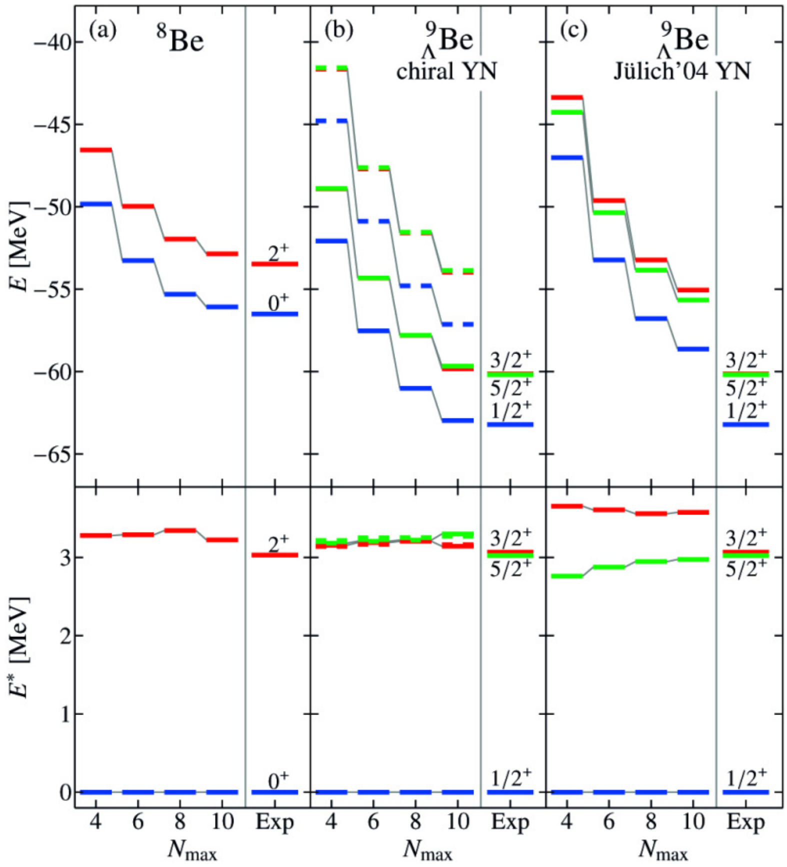
Principal Investigator
Prof. Dr. Johannes Fürnkranz

Project Term
2015 - 2015

Project Areas
Computer Science, Mathematics

Clusters
Lichtenberg Cluster Darmstadt

University
Technische Universität Darmstadt



## Introduction

Reinforcement Learning [1] is a common approach for solving sequential decision problems. The basic idea is to view a problem as a series of decisions that have to be performed with the goal to find a policy for the decision making process which leads to the best possible outcome.

## Methods

Most algorithms are utilizing numeric feedback over the performed decisions, but this hard to define or even unavailable in some domains. Hence, we are relaxing this assumption by learning from pairwise comparisons of two different decision sequences. As example, consider a medical treatment scenario, where it is hard to define a value for the death of a patient. But it is easily possible to determine that all sequences where a patient survives are preferred over deadly ones. One testing domain for the algorithms is chess, [2,3] where large scale annotated databases are available. Those annotations are describing a qualitative evaluation of states or actions, based on an experts opinion. They are relative, because different annotators may evaluate the same state differently. Hence, it is reasonable to use this information in a pairwise manner by defining preferences based on the qualitative annotations. This means states or actions with a good evaluation are preferred over bad ones. This is then used to calculate a numerical evaluation function for states.

## Results

A high computational budget is required for solving the problem, because of the high amount of data required for computing a good solution. As a first step, this enabled research in a setting where batch data is widely available and we showed how to compute a policy for solving the sequential decision problem based on this. In a second phase, this work was extended to the case where sequences and their evaluation are not readily available, but where it is required to propose new sequences and request a pairwise evaluation from an expert. This is a more difficult setting, because it is unknown which kind of sequences are most beneficial for the learning process. But this is also a more practical setting as it is possible to reduce the amount of required preference, and therefore the workload for the expert, by an intelligent sequence creation algorithm. Those algorithms, as well as the learning algorithm itself, are usually subject to a high amount of parameters which have to be tuned. This means a high amount of experiments is required for evaluating new approaches.

## Outlook

This is still ongoing work and several approaches [4,5] have been proposed for solving the aforementioned problem. A preliminary survey was also created to show the current state of research [6].  In the future, we are planning to improve the theoretical foundation of this domain and to continue the development of new algorithms. Additionally, we are also trying to find a fast, generic search algorithm for the parameter tuning problem, because this will ease the application of those algorithms. Comparable research in this area was carried out by Sebag et al. [7]

## Reference

[1] R.S. Sutton and A. Barto (1998), Reinforcement Learning: An Introduction, MIT Press.

[2] C. Wirth and J. Fürnkranz (2012), First Steps Towards Learning from Game Annotations. In: Workshop Proceedings - Preference Learning: Problems and Applications in AI at ECAI 2012, Montpellier: 53-58.

[3] C. Wirth and J. Fürnkranz (2014), On Learning from Game Annotations. In: IEEE Transactions on Computational Intelligence and AI in Games. http://dx.doi.org/10.1109/TCIAIG.2014.2332442

[4] C. Wirth and J. Fürnkranz (2013), A Policy Iteration Algorithm for Learning from Preference-based Feedback, in: Advances in Intelligent Data Analysis XII: 12th International Symposium (IDA-13), Springer. https://doi.org/10.1007/978-3-642-41398-8_37

[5] C. Wirth and J. Fürnkranz (2013), EPMC: Every Visit Preference Monte Carlo for Reinforcement Learning, in: Proceedings of the 5th Asian Conference on Machine Learning, (ACML-13): 483-497, JMLR.org.

[6] C. Wirth and J. Fürnkranz (2013), Preference-Based Reinforcement Learning: A Preliminary Survey, in: Proceedings of the ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards.

[7] R. Akrour, M. Schoenauer, and M. Sebag (2011), Preference-Based Policy Learning, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD-11): 12-27, Springer. https://doi.org/10.1007/978-3-642-23780-5_11

*Last Update:* 2022-09-02 12:03