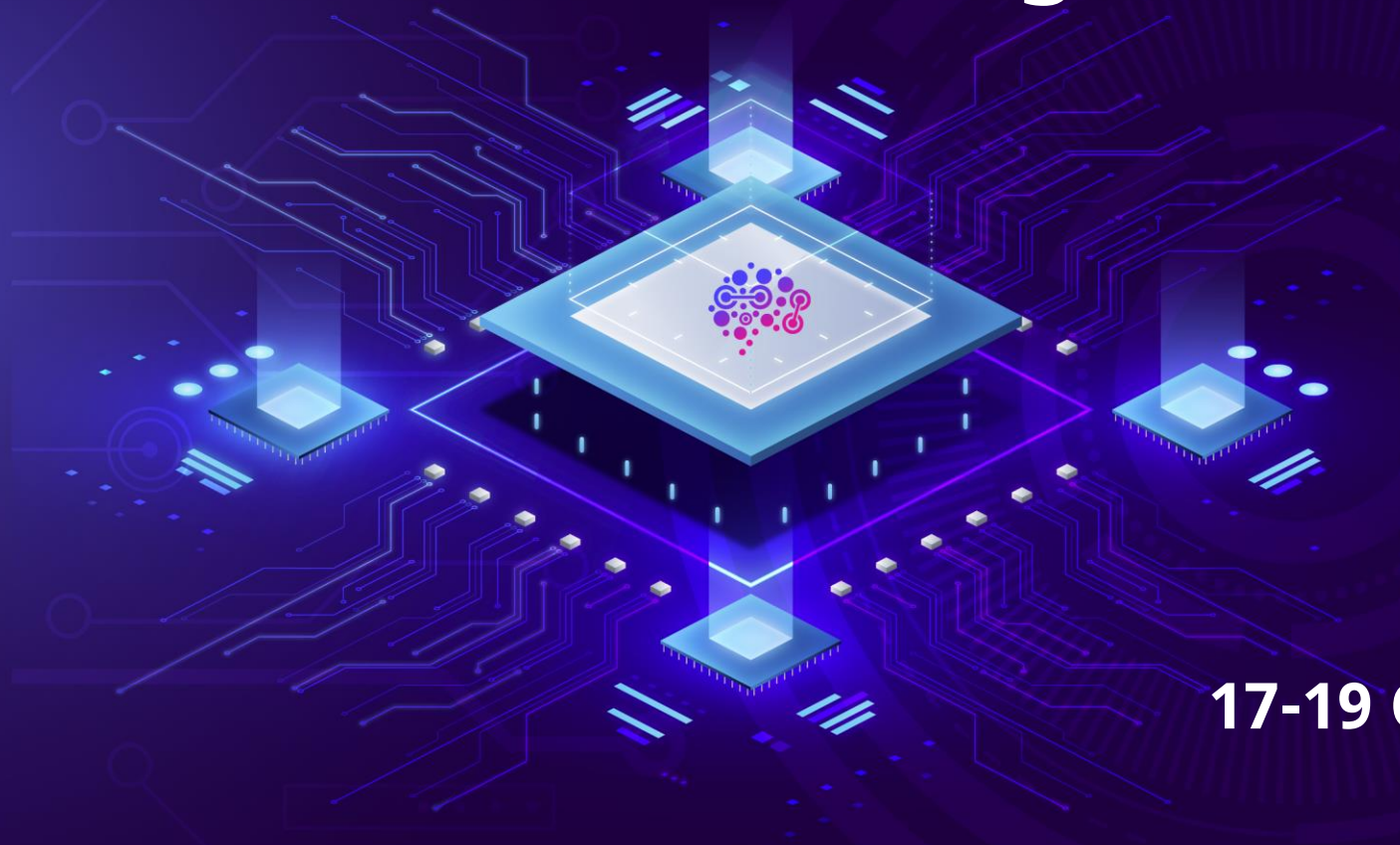
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI

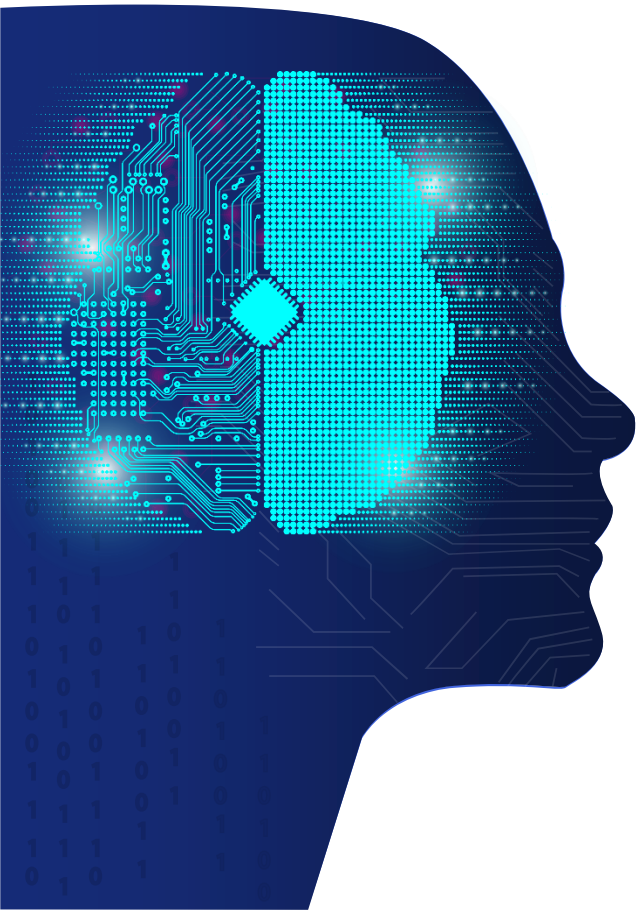
Reward Shaping for Safe Reinforcement Learning Controllers (WiP)

Lorenz Klampfl, AVL List GmbH, Austria on behalf of
Dr. Stefan Jaksic, AIT Austrian Institute of Technology, Austria

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline



- Introduction
  - Reinforcement Learning
  - Reward Shaping
- State-of-the-art
- Research Goal
  - A Systematic Approach in Reward Shaping
- Methodology
- Design and Implementation
- Experimental Results
- Conclusion & Future Research

Copyright © 2023

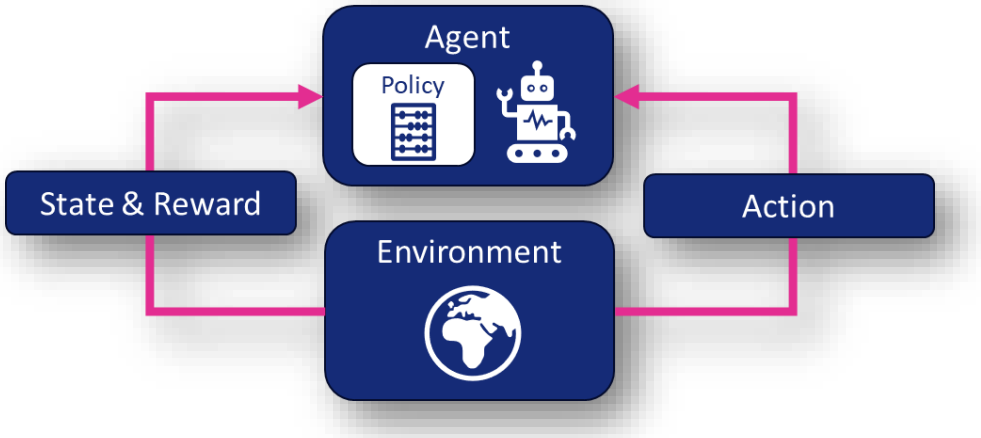
Introduction

Reinforcement Learning

Learning optimal control policies for intelligent agents by interacting with uncertain environments and received rewards.

Policy: Action space for an agent within a state.

Optimal Policy: Sequence of actions that guarantee that the maximum reward is achieved.



Reward Shaping

The performance of agents heavily depends on an appropriate reward function to avoid unsafe behaviors, convergence issues, inefficient learning.

Reward Shaping: Using formal specifications to automatically provide locally shaped rewards that guide the reinforcement learning algorithm.

Proximal Policy Optimization (PPO) by OpenAI

Copyright © 2023

State-of-the-art

Reward shaping is not a new idea, but current state-of-the-art methods often lack on properties like scalability, density, and usability in Non-Markovian settings.

- Fragments of LTL for finite horizon tasks. [1]
- Custom specification languages for generating rewards. [2]
- Enhance MDP with n-step history for Non-Markovian rewards. [3], [4]

Innovation: *Non-sparse, p*otential-based rewards using symbolic automata.

Copyright © 2023

Research Goal

Research Topics & Goal

Development of a systematic approach for the design of a reward function that prevents reward hacking and enables the optimization for intermediate goals.

Reward shaping algorithm preserving optimal control policies by utilizing formal specifications and symbolic automata.

Showcasing its potential and capabilities as well as possible application in different domains.

Challenges

Very hard to hand-craft good reward functions that:

- *Enable multi-goal optimization:*
  - In autonomous driving the reward must be encoded to reach a sequence of goals but also ensure safe driving behavior.
- *Prevent reward hacking:*
  - Reward for progress ➡ driving in circles or breaking safety rules.
  - Penalty for unsafe behavior ➡ driving very slowly and hardly any progress.

Copyright © 2023

Methodology

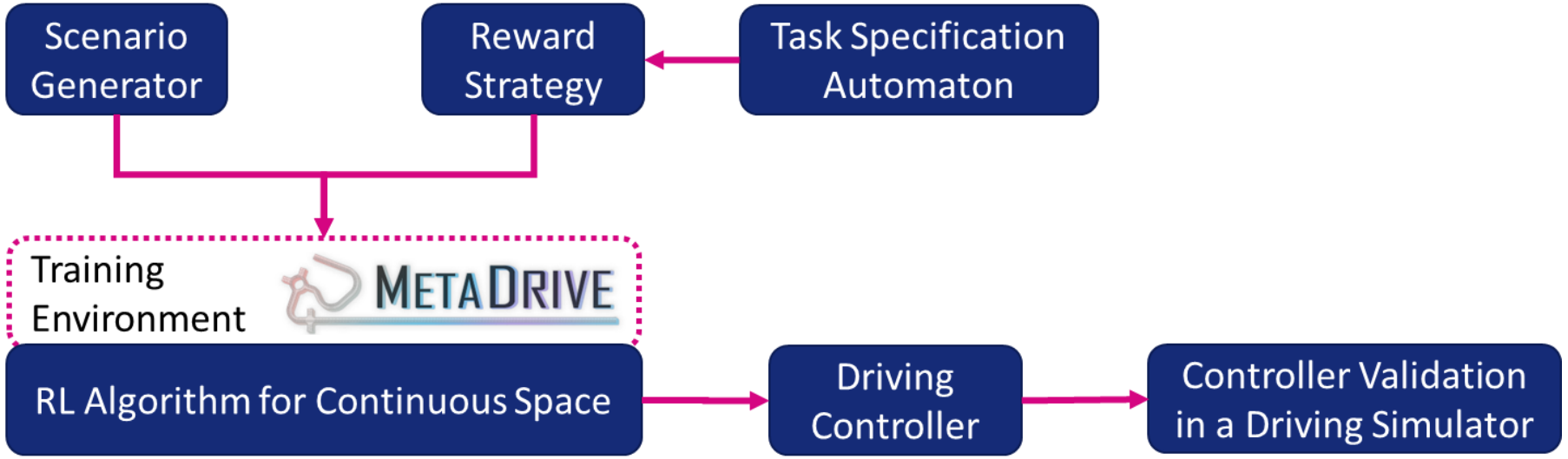
Goal:

Given a task *A,* learn a policy which maximizes the probability of satisfying task *A.*

- *Example:* ensure safe driving behavior while reaching *N* sequential goals.

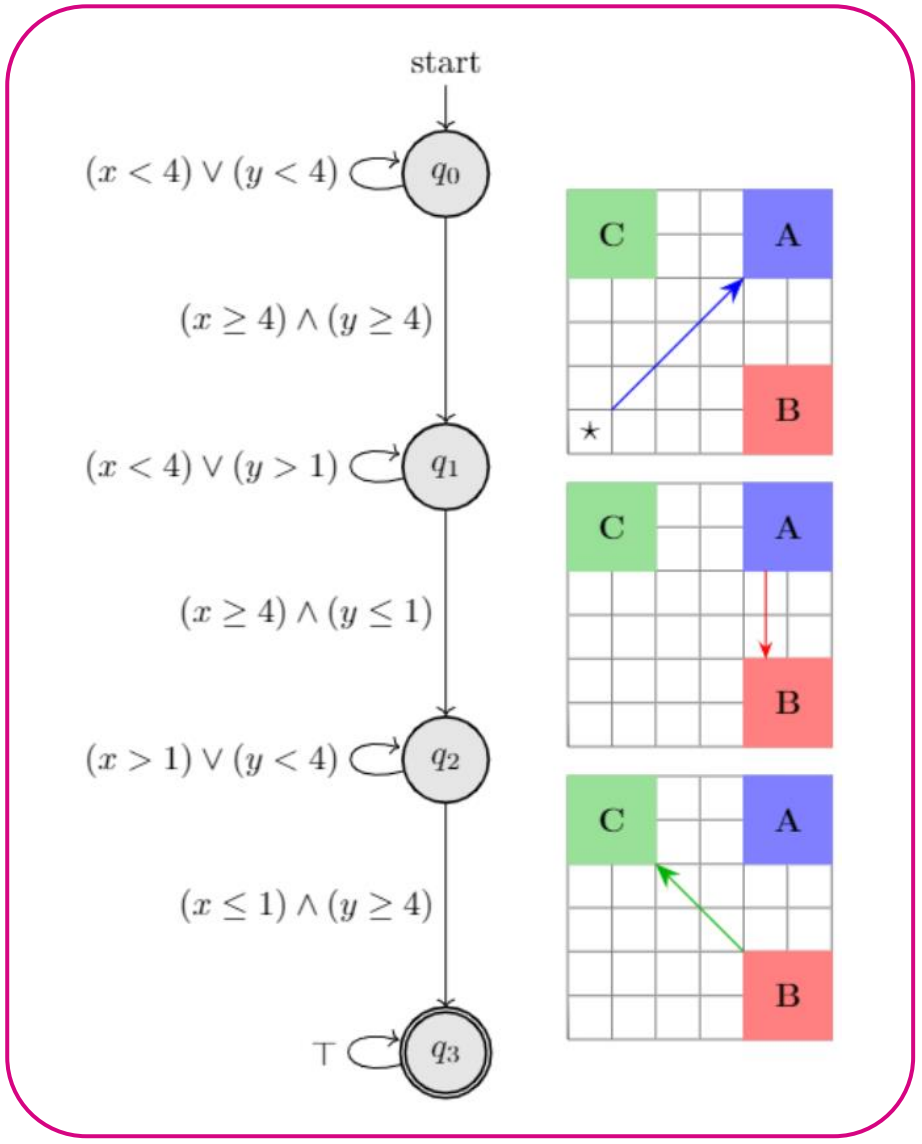
Requirements for the reward function:

- Density reward ➡ a reward is defined for every action
- Agent needs to move closer to goal ➡ positive reward
- Shaped rewards does not alter the set of optimal policies.



Copyright © 2023

Design and Implementation



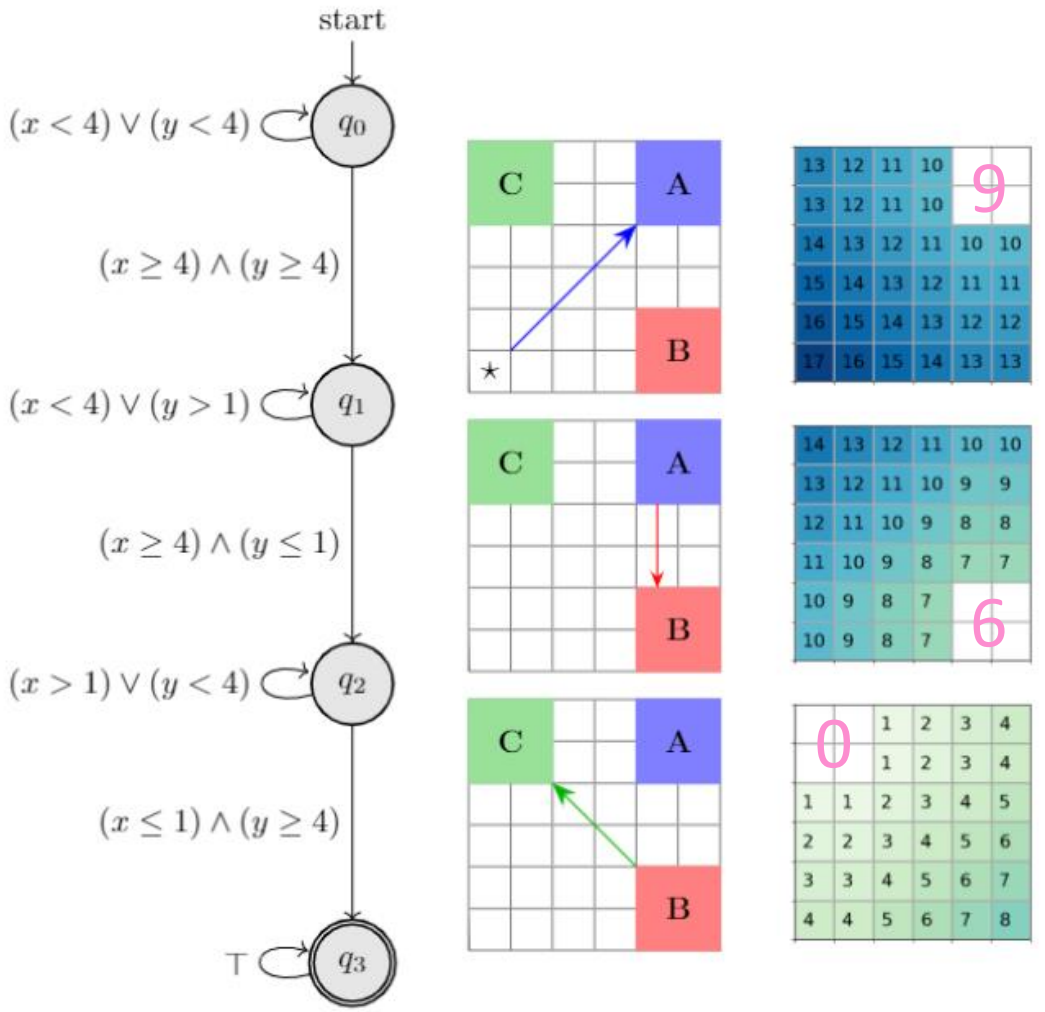
- *Experiment Design:*
  - Grid environment
- *Specification:*
  - Agent needs to go from *Init* ⟶ *A* ⟶ *B* ⟶ *C*.
  - Specification defined using automaton.
  - Transitions (*q, q'*) with guards.

- Automaton is used to produce a reward to guide the reinforcement learning algorithm to learn an agent that follows the specified path.

- Second automaton/Markov Decision Process is used to represent the environment.

Copyright © 2023

Design and Implementation

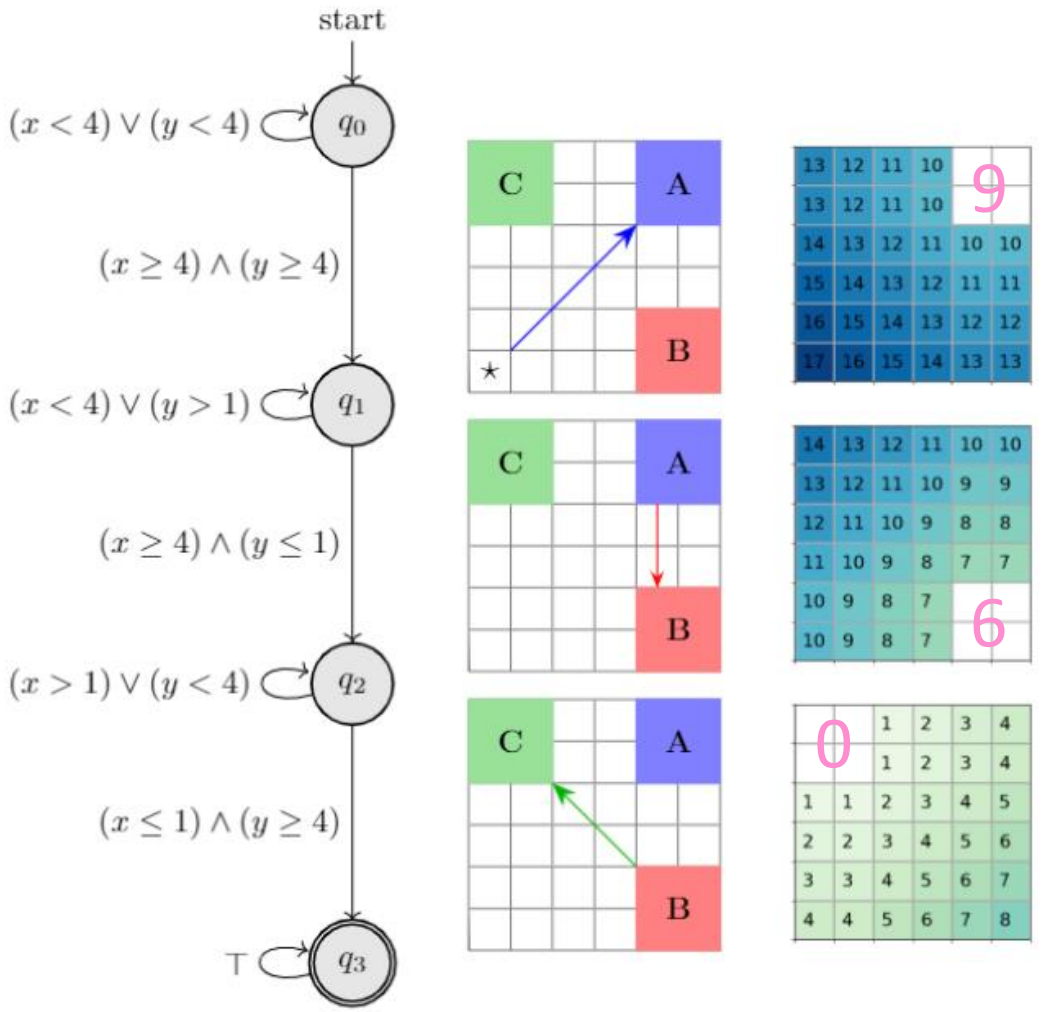


Reward Ingredients:

- Agent is in a certain state *(S, q)*, w.r.t. the environment and the specification.
- Assume an agent reached state *q* of the specification automaton and state *S* of the environment:
  - How good is this state ↔ what is the potential of this state?
- 2 ingredients the agent needs:
  - How good am I doing w.r.t. the specification?
    - Local specification goals vs. global specification goals.
  - How good am I doing in the environment?
    - State in the environment vs. next goal.

Copyright © 2023

Design and Implementation



Symbolic Subtask Progress (SSP): *[specification metric]*

- Approximate the distance for an automaton state to the final state:

$$SSP(q, q') = distance_to_q' + SSP(q")$$

- Final state: $SSP(q_3) = 0$
- *Progress* = estimate of remaining costs to reach final state.

Symbolic Potential Function (SPF): *[environment metric]*

- $SPF(s, q)$ = minimum number of steps to reach the next goal from environment state $s + SSP(q, q')$.
- Final goal: $SPF = 0$

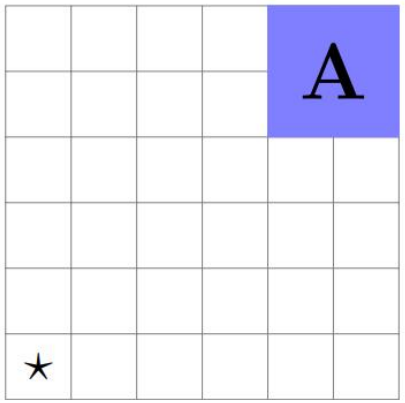
Shaped Reward:

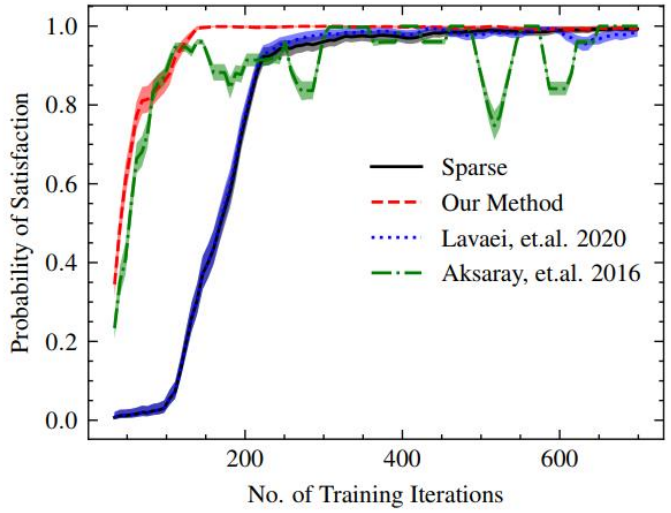
$$reward\left((s,q),(s',q')\right) = final_goal_reached\ ?\ \max : 0 + SPF\ (s,q)\ - SPF\ (s',q')$$

Copyright © 2023

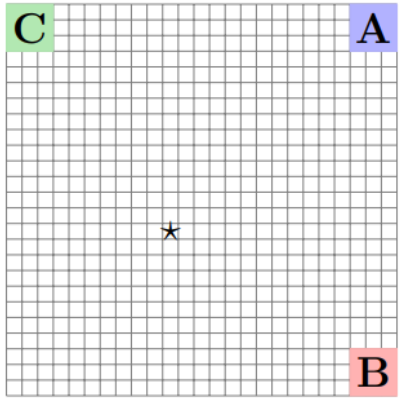
Experimental Results

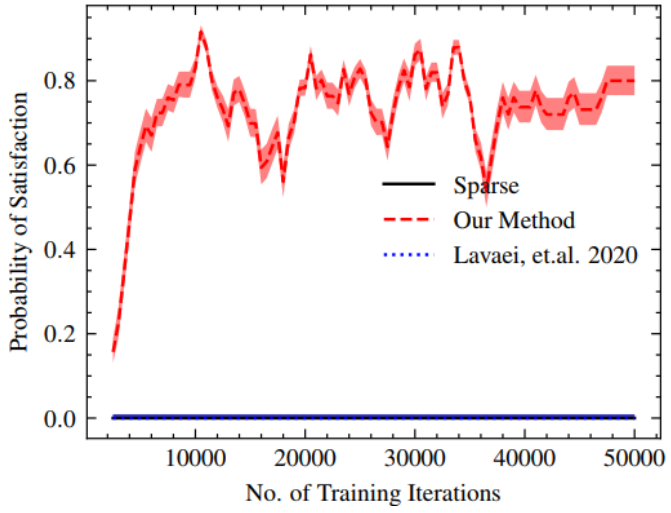
Simple "Reach" Task





Sequential Goals Task





Copyright © 2023

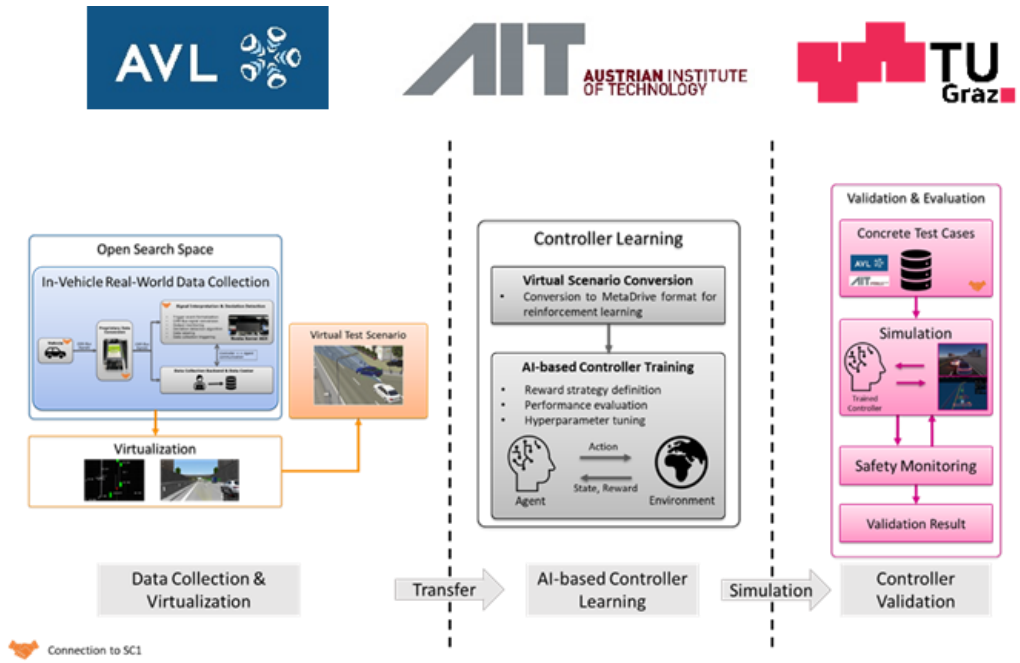
Conclusion & Future Work

Conclusion

- Having appropriate reward functions is essential for the success of learning a controller using reinforcement learning.
- Hand-crafted rewards do not provide guarantees and usually introduce a set of potential problems.
- Potential-based reward using formal specifications is a possible solution to those problems.

Future Work

- Apply this reward function to the continuous settings.
- Showcasing its potential and capabilities as well as possible application in different domains, especially in the autonomous driving domain within the AI4CSM project.



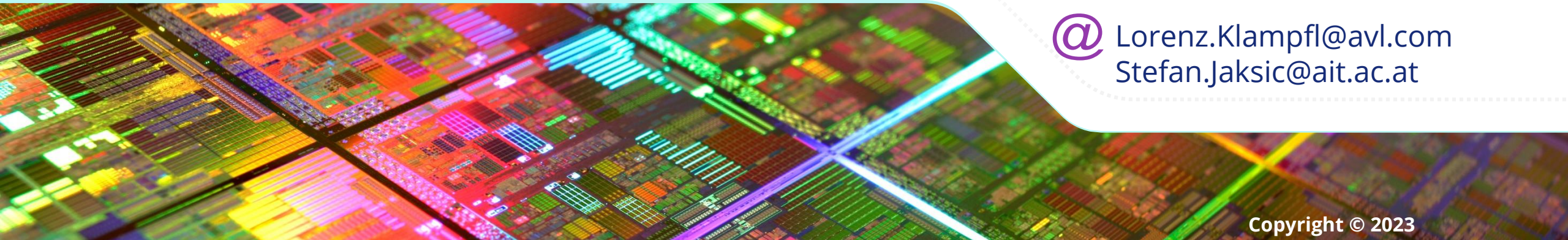
Copyright © 2023

EDGE AI

AI$_4$CSM

Thank You

For your attention



@ Lorenz.Klampfl@avl.com
Stefan.Jaksic@ait.ac.at

Copyright © 2023

References

[1] X. Li, Y. Ma, and C. Belta, “Automata-Guided Hierarchical Reinforcement Learning for Skill Composition,” *arXiv:1711.00129 [cs]*, May 2018.

[2] K. Jothimurugan, R. Alur, and O. Bastani, “A Composable Specification Language for Reinforcement Learning Tasks,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] X. Li, Y. Ma, and C. Belta, “A policy search method for temporal logic specified reinforcement learning tasks,” in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 240–245.

[4] A. Balakrishnan and J. V. Deshmukh, “Structured reward shaping using signal temporal logic specifications,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Nov. 2019, pp. 3481–3486.

Copyright © 2023