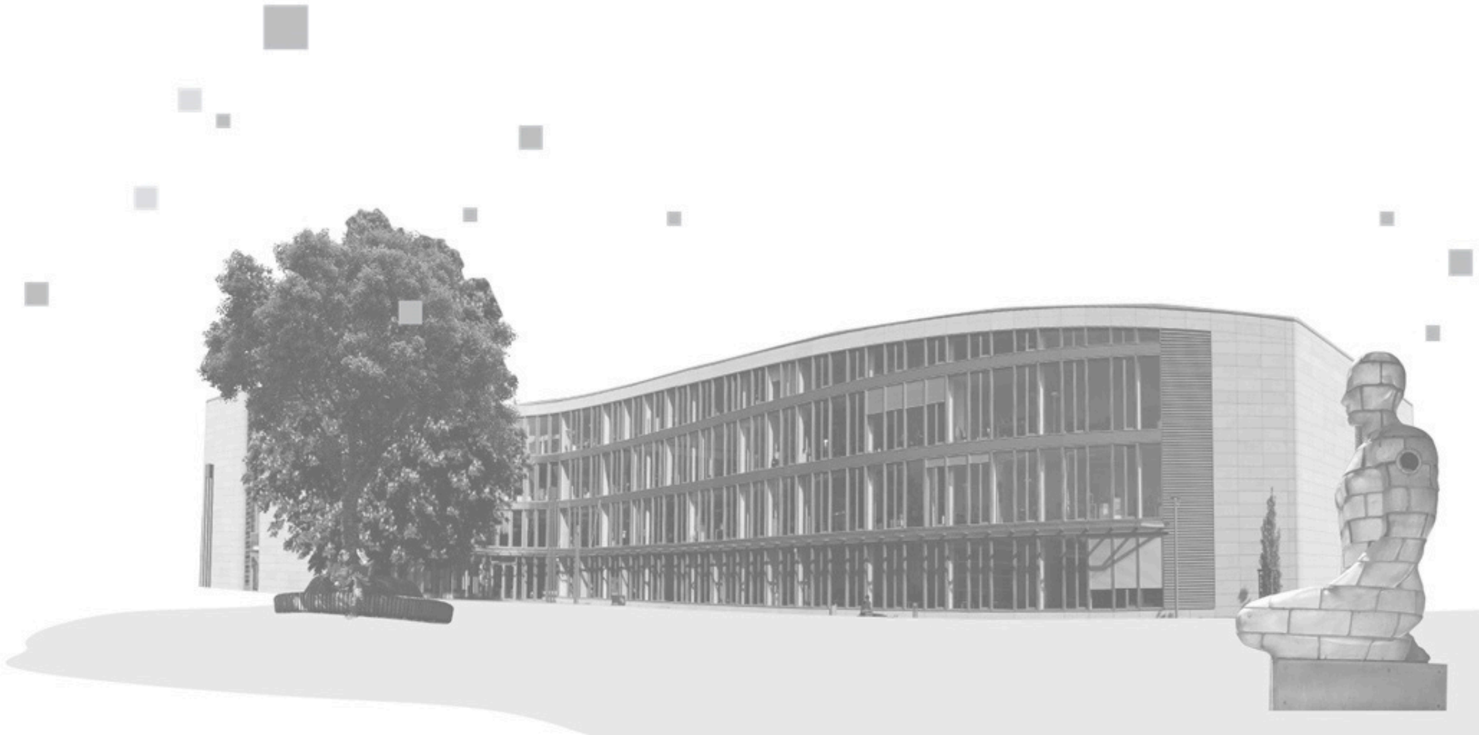
KI Service Zentrum
by Hasso-Plattner-Institut
HPI

Reinforcement Learning I –
How Agents Learn by Trial and Error



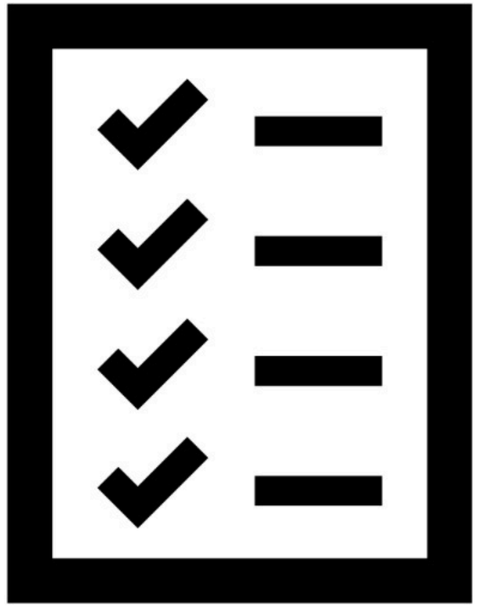
Design IT.
Create Knowledge.

www.hpi.de

Attendance List

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz



Four AI Service Centres in Germany

KI Service Zentrum by Hasso-Plattner-Institut

https://hpi.de/ki-servicezentrum/

Overview

KI-Servicezentren

Goal:
Reduce barriers to the implementation of AI applications in society and the economy

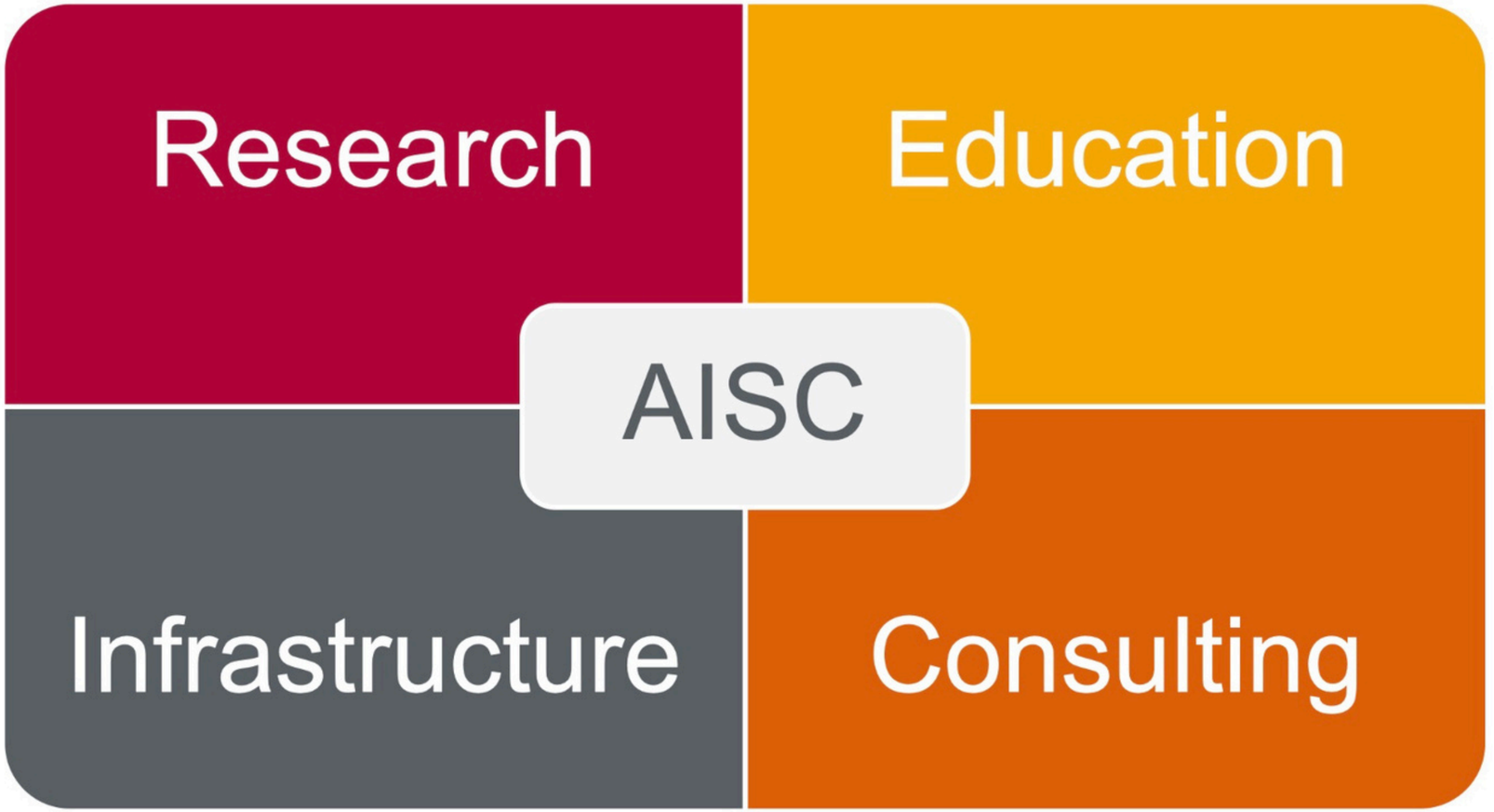
Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

AI Service Centre Berlin-Brandenburg

KI Service Zentrum by Hasso-Plattner-Institut HPI

Structure



Gefördert durch:

Bundesministerium für Forschung, Technologie und Raumfahrt

KI Service Zentrum
by Hasso-Plattner-Institut

HPI

Research

- AI methods research
- AI operations research

PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?

Sedigheh Eslami, Christoph Meinel, Gerard de Melo

Hasso Plattner Institute / University of Potsdam

{sedigheh.eslami, christoph.meinel, gerard.demelo}@hpi.de

Abstract

Contrastive Language-Image Pre-training (CLIP) has shown remarkable success in learning with cross-modal supervision from extensive amounts of image-text pairs collected online. Thus far, the effectiveness of CLIP has been investigated primarily in general-domain multimodal problems. In this work, we evaluate the effectiveness of CLIP for the task of Medical Visual Question Answering (MedVQA). We present PubMedCLIP, a fine-tuned version of CLIP for the medical domain based on PubMed articles. Our experiments conducted on two MedVQA benchmark datasets illustrate that PubMedCLIP achieves superior results improving the overall accuracy up to 3% in comparison to the state-of-the-art Model-Agnostic Meta-Learning (MAML) networks pre-trained only on visual data. The PubMedCLIP model with different back-ends, the source code for pre-training them and reproducing our MedVQA pipeline is publicly available at https://github.com/sarahESL/PubMedCLIP.

1 Introduction

Medical visual question answering (MedVQA) seeks answers to natural language questions about a given medical image. The development of MedVQA has considerable potential to benefit healthcare systems, as it may aid clinicians in interpreting medical images and obtaining more accurate diagnoses by consulting a second opinion. Thus, it has become a very active area of research, with competitive benchmarks and yearly competitions (Abacha et al., 2021). Yet, visual question answering in the medical domain in particular remains non-trivial as we suffer from a general lack of large balanced training data, in part due to privacy concerns. To solve the multimodal task of MedVQA, a system must understand both medical images and textual questions and infer the associations between them sufficiently well to produce a correct answer (An

Findings of the Association for Computati
May 2-6, 2023 ©2023 Associ

Exploring Paracrawl for Document-level Neural Machine Translation

Yusser Al Ghussin[1,2], Jingyi Zhang[3] and Josef van Genabith[1,3]

[1]German Research Center for Artificial Intelligence (DFKI),
Saarland Informatics Campus, Saarbrucken, Germany
[2]Department of Language Science and Technology, Saarland University, Germany
[3]Hasso-Plattner-Institut (HPI), Potsdam, Germany
yusser.al_ghussin/Josef.Van_Genabith@dfki.de,Jingyi.Zhang@hpi.de

Abstract

Document-level neural machine translation (NMT) has outperformed sentence-level NMT on a number of datasets. However, document-level NMT is still not widely adopted in real-world translation systems mainly due to the lack of large-scale general-domain training data for document-level NMT. We examine the effectiveness of using Paracrawl for learning document-level translation. Paracrawl is a large-scale parallel corpus crawled from the Internet and contains data from various domains. The official Paracrawl corpus was released as parallel sentences (extracted from parallel webpages) and therefore previous works only used Paracrawl for learning sentence-level translation. In this work, we extract parallel paragraphs from Paracrawl parallel webpages using automatic sentence alignments and we use the extracted parallel paragraphs as parallel documents for training document-level translation models. We show that document-level NMT models trained with only parallel paragraphs from Paracrawl can be used to translate real documents from TED, News and Europarl, outperforming sentence-level NMT models. We also perform a targeted pronoun evaluation and show that document-level models trained with Paracrawl data can help context-aware pronoun translation. We release our data and code here[1].

1 Introduction

The Transformer translation model (Vaswani et al., 2017), which performs sentence-level translation based on attention networks, has achieved great success and significantly improved the state-of-the-art in machine translation. Compared to sentence-level translation, document-level translation (Xu et al., 2021; Bao et al., 2021; Jaaregi Unanue et al., 2020; Ma et al., 2020; Maruf et al., 2019; Tu et al., 2018; Maruf and Haffari, 2018) performs translation at document-level and can potentially fur-

[1]https://github.com/YusserAl/Exploring-Paracrawl-for-Document-level-Neural-Machine-Translation

Proceedings of the 17th Conference of the European Chapter of
May 2-6, 2023 ©2023 Associ

Efficient Parallelization Layouts for Large-Scale Distributed Model Training

Johannes Hagemann
Aleph Alpha / Hasso Plattner Institute
johannes.hagemann@student.hpi.de

Samuel Weinbach
Aleph Alpha
samuel.weinbach@aleph-alpha.com

Konstantin Dobler
Hasso Plattner Institute
konstantin.dobler@hpi.de

Maximilian Schall
Hasso Plattner Institute
maximilian.schall@hpi.de

Gerard de Melo
Hasso Plattner Institute
gerard.demelo@hpi.de

Abstract

Efficiently training large language models requires parallelizing across hundreds of hardware accelerators and invoking various compute and memory optimizations. When combined, many of these strategies have complex interactions regarding the final training efficiency. Prior work tackling this problem did not have access to the latest set of optimizations, such as FLASHATTENTION or sequence parallelism. In this work, we conduct a comprehensive ablation study of possible training configurations for large language models. We distill this large study into several key recommendations for the most efficient training. For instance, we find that using a micro-batch size of 1 usually enables the most efficient training layouts. Larger micro-batch sizes necessitate activation checkpointing or higher degrees of model parallelism and also lead to larger pipeline bubbles. Our most efficient configurations enable us to achieve state-of-the-art training efficiency results over a range of model sizes, most notably a Model FLOPs utilization of 70.5% when training a LLAMA 13B model.

1 Introduction

The number of parameters and computational resources spent on training deep neural networks is growing rapidly ([1, 3, 14]). The largest models consisting of hundreds of billions of parameters do not even fit onto a single hardware accelerator. Thus, training these models requires various ways of reducing the memory requirements, such as ZeRO ([16]), activation checkpointing ([2]), and 3D-parallel (data, tensor, and pipeline parallel) training ([13]). 3D parallelism, in particular, has been demonstrated to be effective for the training of Transformer-based large language models (LLMs) with hundreds of billions of parameters ([13]).

However, training these models efficiently with 3D parallelism requires significant domain expertise and extensive manual effort to determine the ideal configurations. These configurations not only need to combine data, model, and pipeline parallelism most efficiently, but also consider complex interactions with other memory and compute optimizations. FLASHATTENTION [5] in particular has had a notable impact since its release, enabling us to train models at previously impossible degrees of training efficiency. In light of these developments, we conduct a systematic study via a large-scale training efficiency sweep of these interactions. We consider up to 256 GPUs and LLAMA models with up to 65 billion parameters.

Workshop on Advancing Neural Network Training at 37th Conference on Neural Information Processing Systems (WANT@NeurIPS 2023).

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

Zugangsanfrage
aisc.hpi.de

Infrastructure

- **Free access**
- No production operation
  - Data should be **anonymised** or **synthesised**
  - No **hosting** of products
- **Reporting** & **publication** by users
- **Old rights** remain with users
- **New rights** remain with users
  - Granting of rights of use for research and teaching

Training
- 64 NVIDIA H100 GPU

Inference
- 40 NVIDIA A30 GPU

ARM Server
- Ampere Altra Max M128-30 CPU
- 2 x NVIDIA L40 GPUs

GPU Server
- AMD Epyc CPU
- 8 x NVIDIA L40S GPU

Edge
- ARMv8 CPU
- NVIDIA Jetson AGX Module

Neuromorph
- 288 SpiNNaker2 Chips

Storage
- 1.5 PB NVMe

Network
- 400 Gb/s Infiniband
- 200 Gb/s Ethernet

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Newsletter

Education

Talks

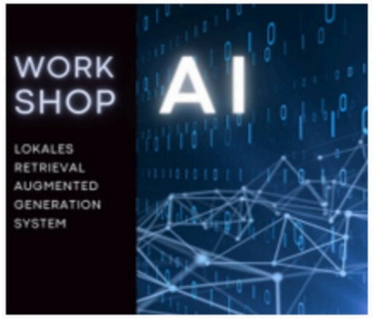
tele-task.de/series/1463

- Guest talks about research and innovation

Work shops

aimaker.community

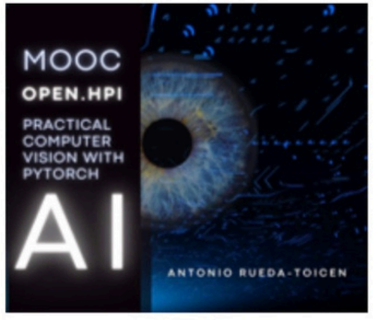
- Practical topics
- Example topics: Speech2summary, Docker for ML, semantic search



MOOCs

open.hpi.de/channels/ai-service-center

- ChatGPT: What does generative AI mean for our society?
- Profitable AI
- Understanding and avoiding AI biases



Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Sprechstunde buchen

Consulting

AI consultation hours

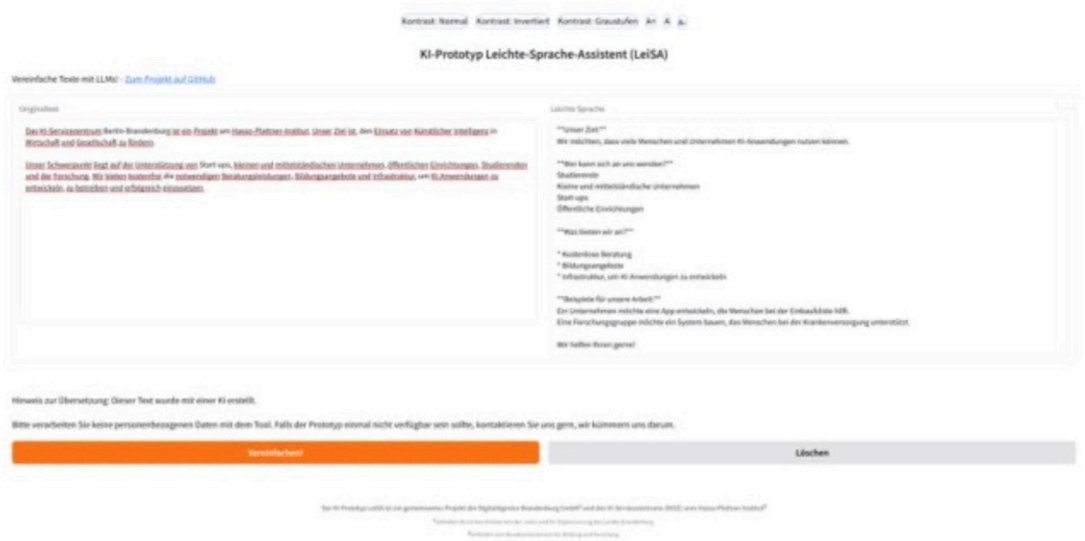
- Answering questions:
  - about AI infrastructure
  - about AI models & frameworks
  - about AI use cases

AI pilot projects

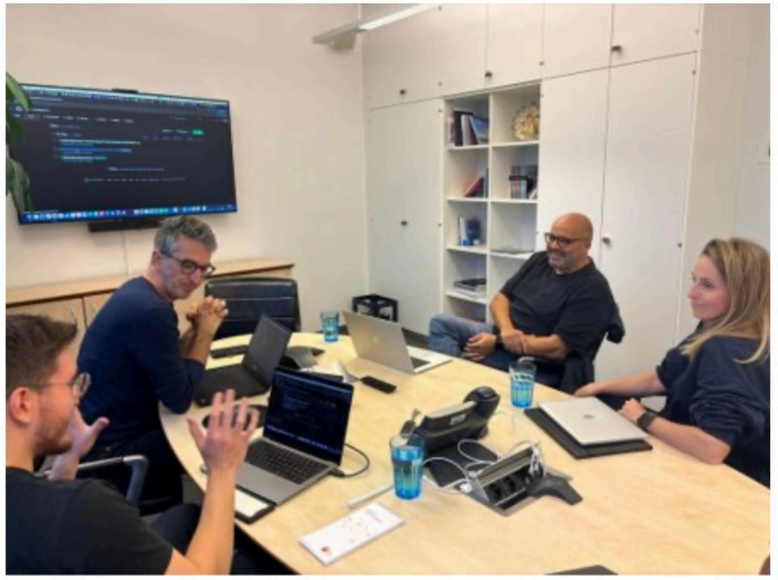
- Co-development of a prototype
- Application every three months
- Selection criteria, e.g. AI maturity, public benefit
- Publication of results

Cooperation

- Jointly organized network meetings



github.com/aihpi/leichte-sprache



Previous AI pilot projects

- Generation of math problems
- Generation of easy language
- Generation of upcycling suggestions
- Reduction of food waste
- Dating by handwriting

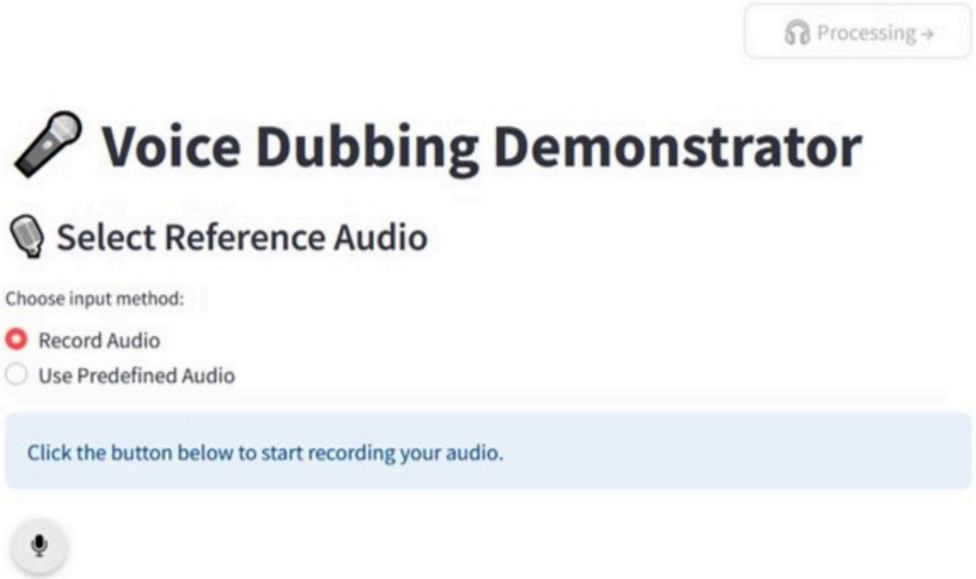
Jetzt bewerben!

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Speech processing

https://tts.aisc.hpi.de/
user: event
pw: aisc@hpi2022!

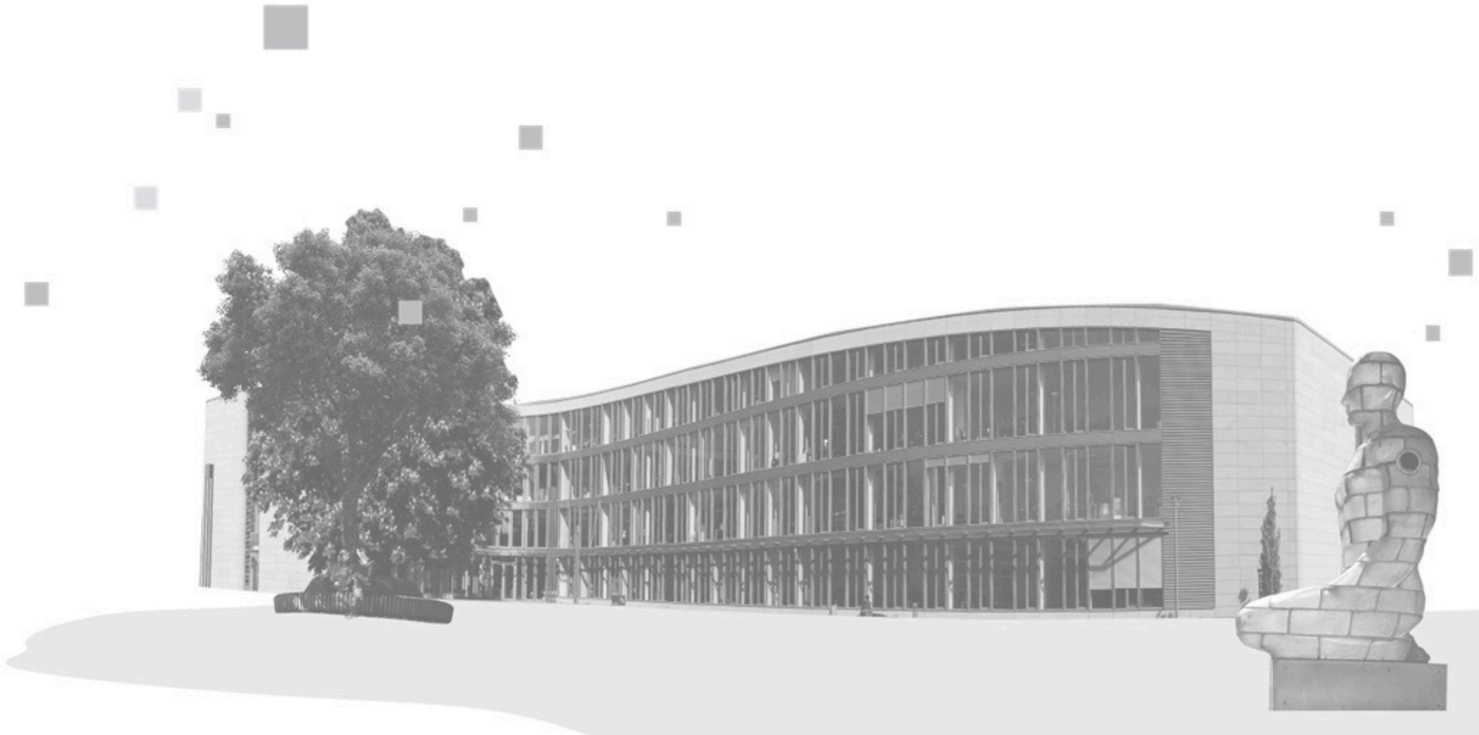
Prototypes



Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

Reinforcement Learning I –
How Agents Learn by Trial and Error



Design IT.
Create Knowledge.

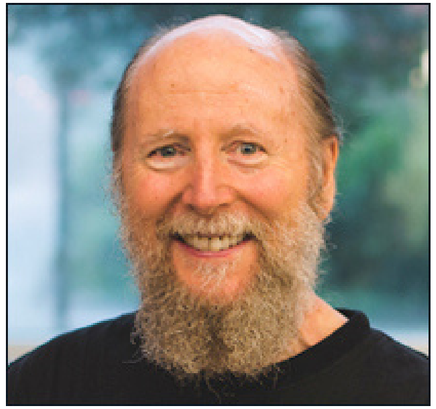
www.hpi.de

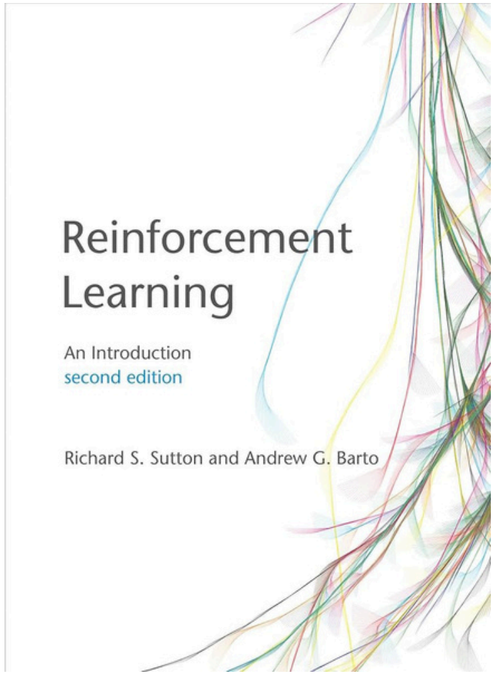
Motivation and context |

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz

“Of all the forms of machine learning, reinforcement learning is the closest to the kind of learning that humans and other animals do.” - Richard Sutton, Andrew Barto





source of pictures: https://amturing.acm.org

Motivation and context |

kisz@hpi.de
hpi.de/kisz

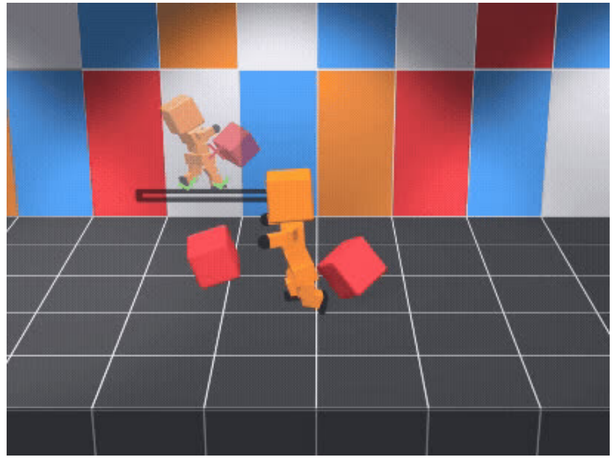
“Of all the forms of machine learning, reinforcement learning is the closest to the kind of learning that humans and other animals do.” -Richard S. Sutton

AI Warehouse - https://www.youtube.com/watch?v=v3UBlEJDXR0

Motivation and context ||: Applications

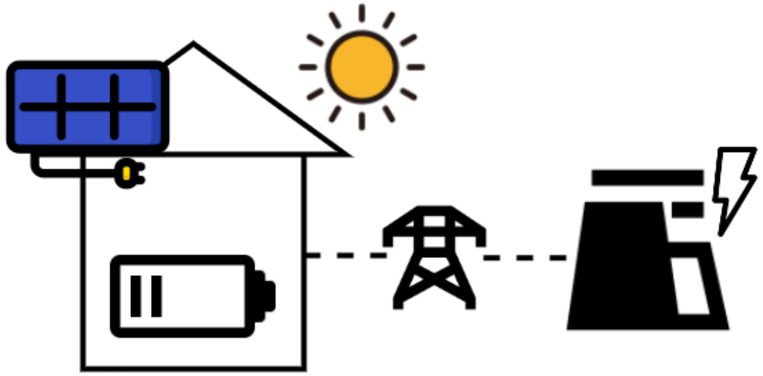
KI Service Zentrum by Hasso-Plattner-Institut
HPI
kisz@hpi.de
hpi.de/kisz

Robotics



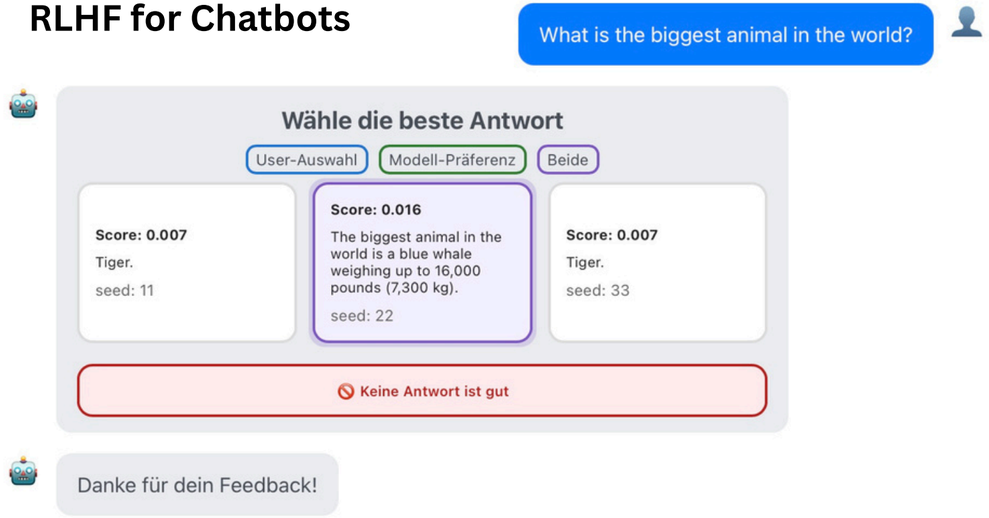
AI Warehouse
https://www.youtube.com/watch?v=L_4BPjLBF4E&t=60s

Control Energy Systems



https://github.com/tobirohrer/building-energy-storage-simulation

RLHF for Chatbots



Motivation and context |||

kisz@hpi.de
hpi.de/kisz

Supervised Learning

- Models are **trained on labeled data** to predict outputs

Reinforcement Learning

- Learning optimal behavior by **interacting with an environment**

Unsupervised Learning

- Learn to **identify previously unknown patterns** and relationships in unlabeled data

used when **optimal actions are too complex or unknown to define** but can be learned through experience

What will we cover today?

KI Service Zentrum by Hasso-Plattner-Institut HPI

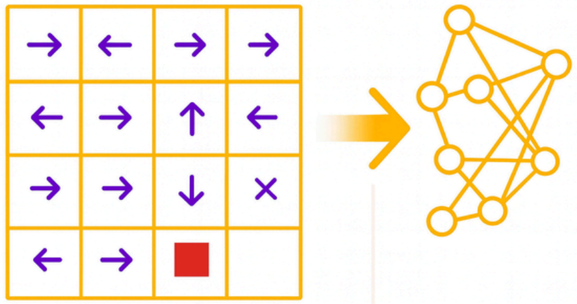
kisz@hpi.de
hpi.de/kisz

Today

- Basic concepts
- ”Hello World” of RL:
  Q-learning on FrozenLake environment
- Build intuition about learning parameters
- First Coding

Following workshops

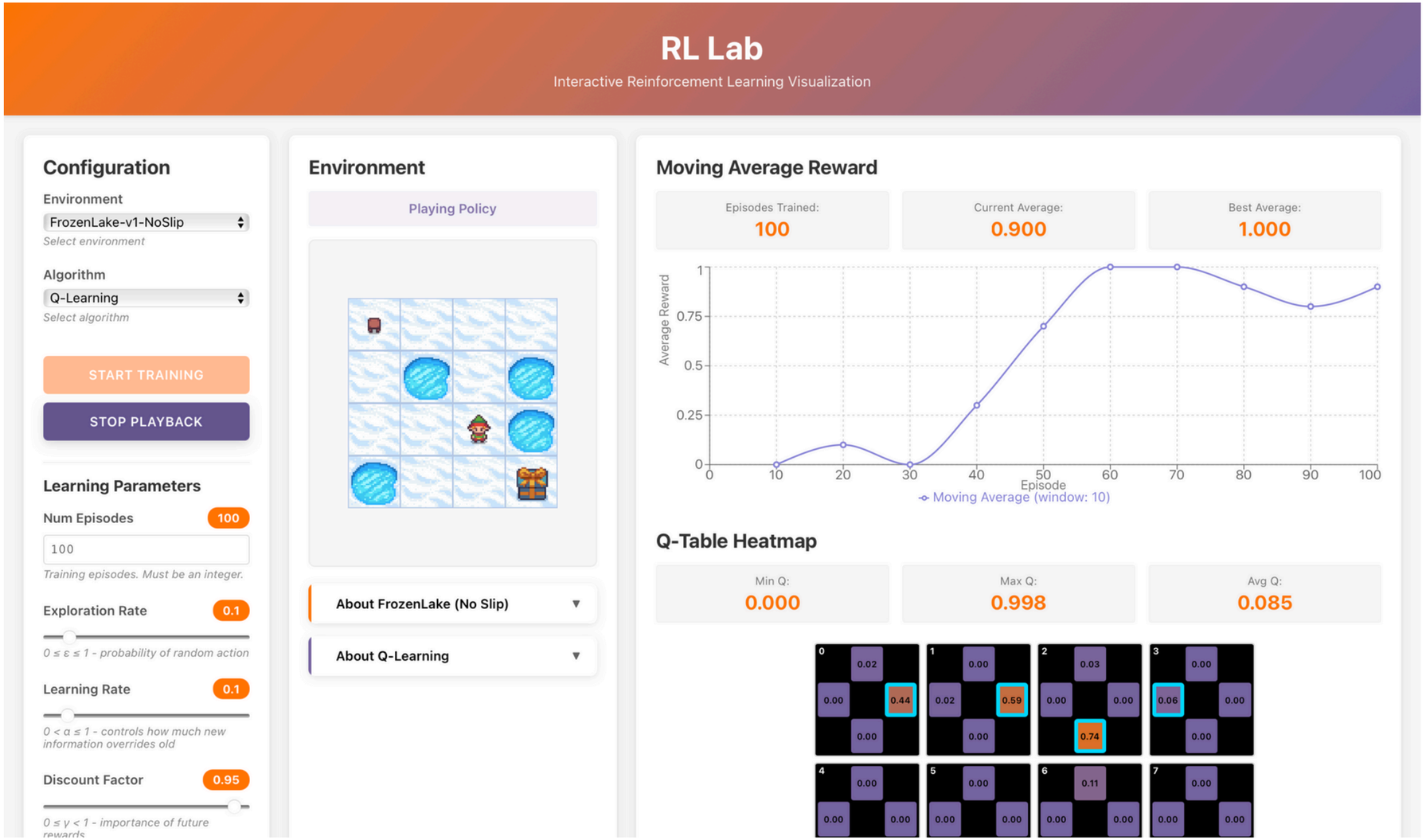
- Deep RL
- When to apply which algorithm
- How to set up a RL problem
- Real-world applications



What will we cover today?

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz

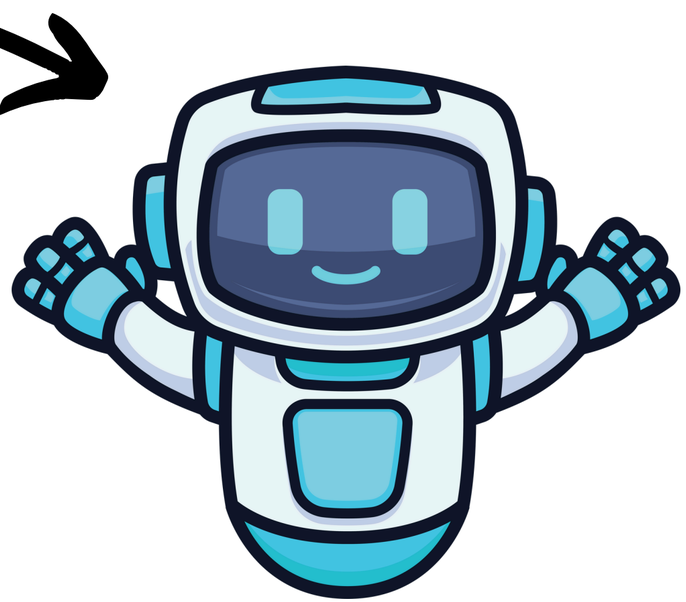


Key terms

kisz@hpi.de
hpi.de/kisz

Agent

An **agent** takes actions, gets feedback, and learns what works best

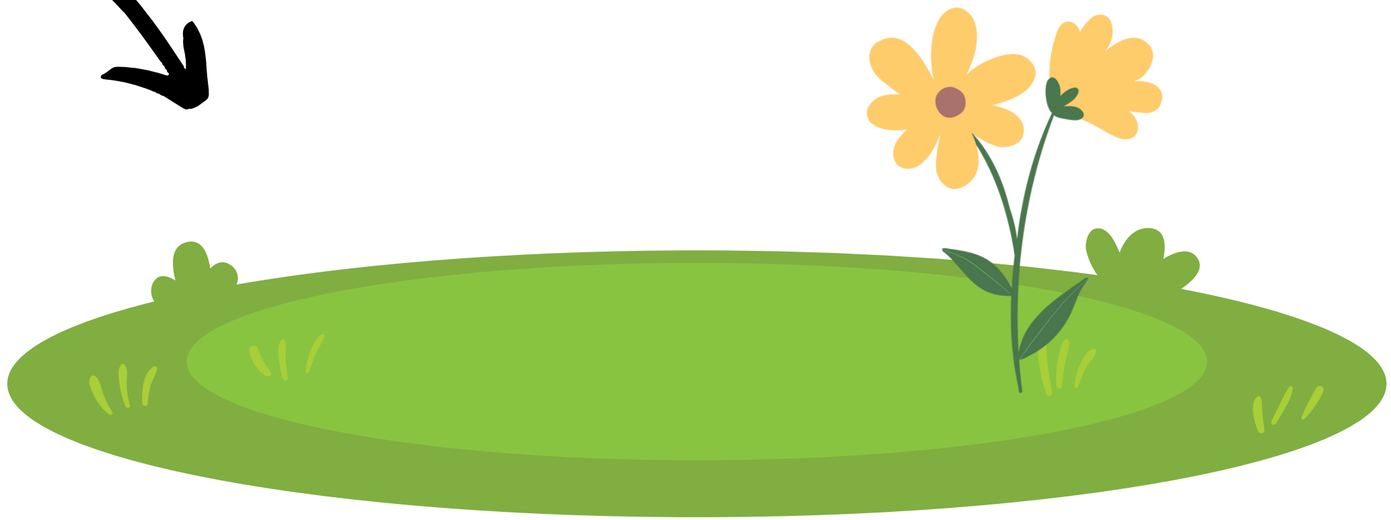


Key terms

kisz@hpi.de
hpi.de/kisz

Environment

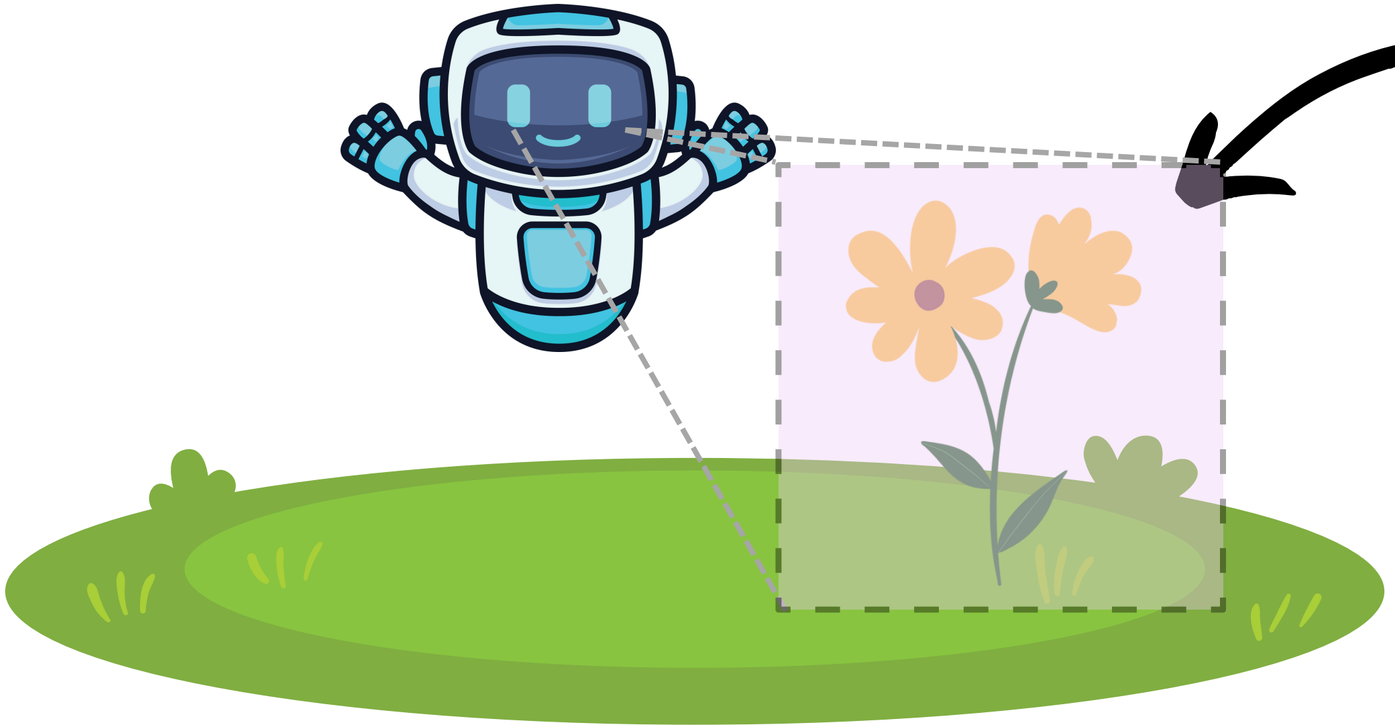
The environment is **the world the agent interacts with**. It defines what happens when the agent takes an action.



Key terms

KI Service Zentrum by Hasso-Plattner-Institut HPI

kisz@hpi.de
hpi.de/kisz

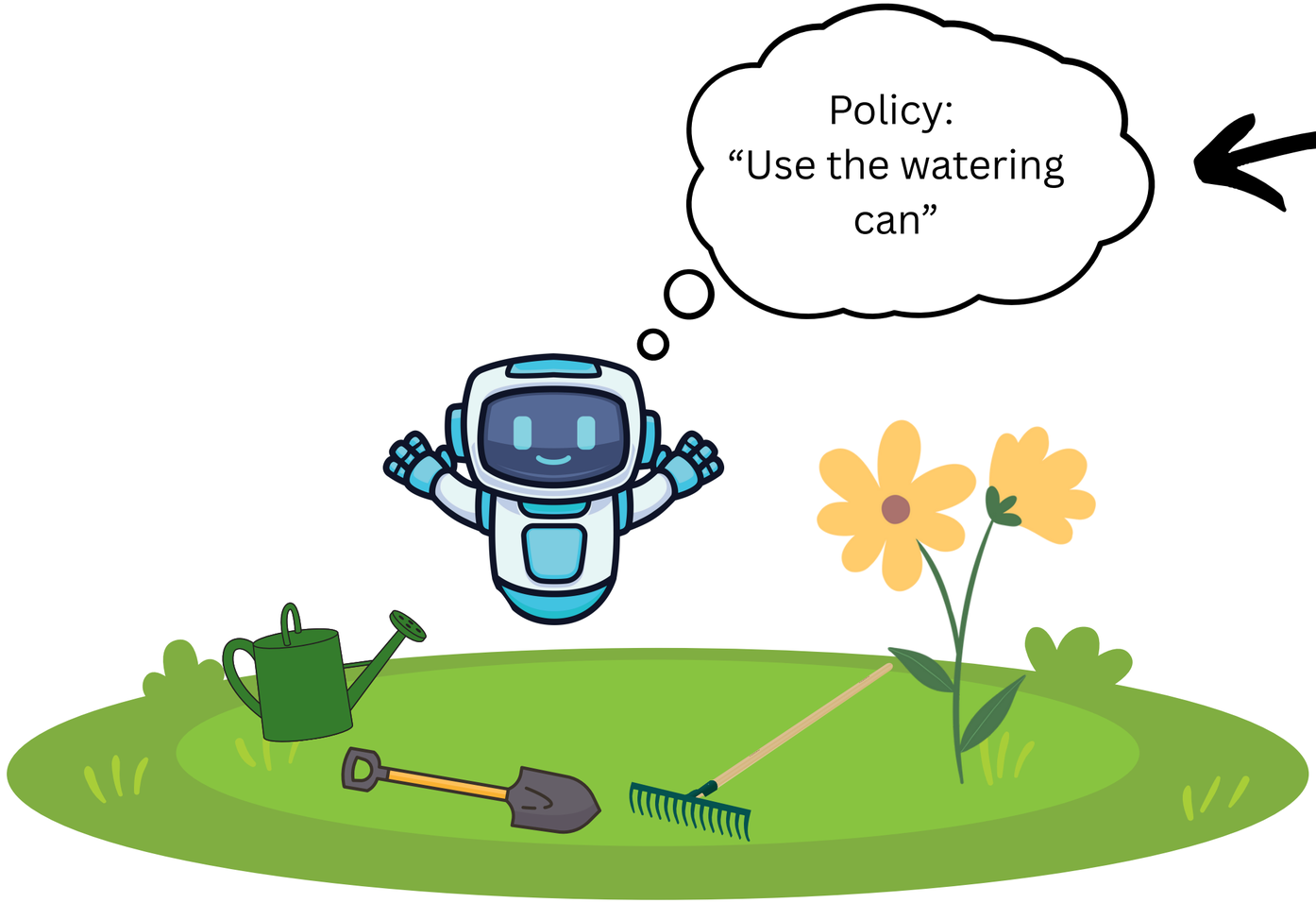


State

The state is the **agent's current situation**, what it perceives in the environment

Key terms

kisz@hpi.de
hpi.de/kisz

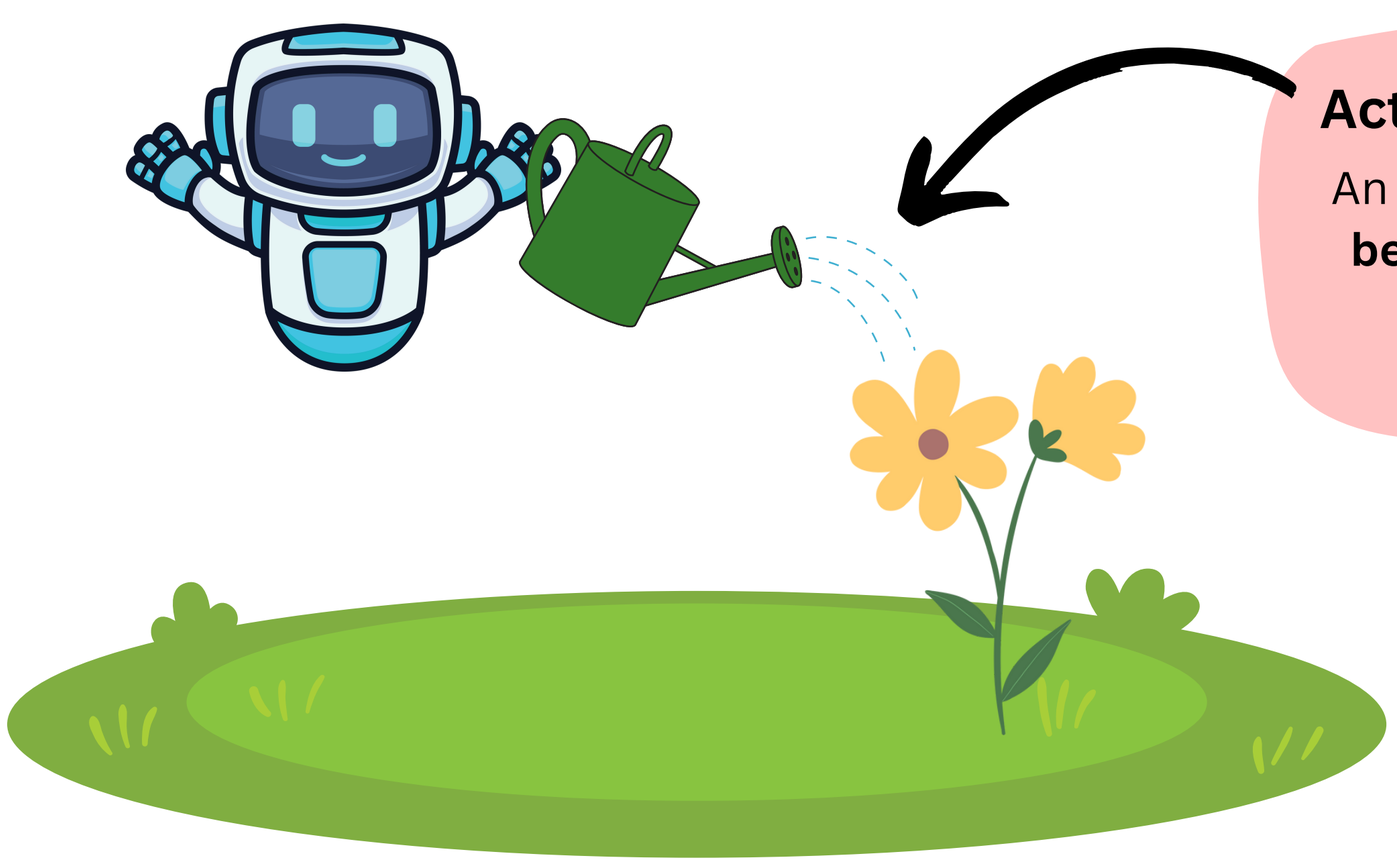


Policy

The policy is the strategy or rule that **tells the agent which action to take** in a given state

Key terms

kisz@hpi.de
hpi.de/kisz

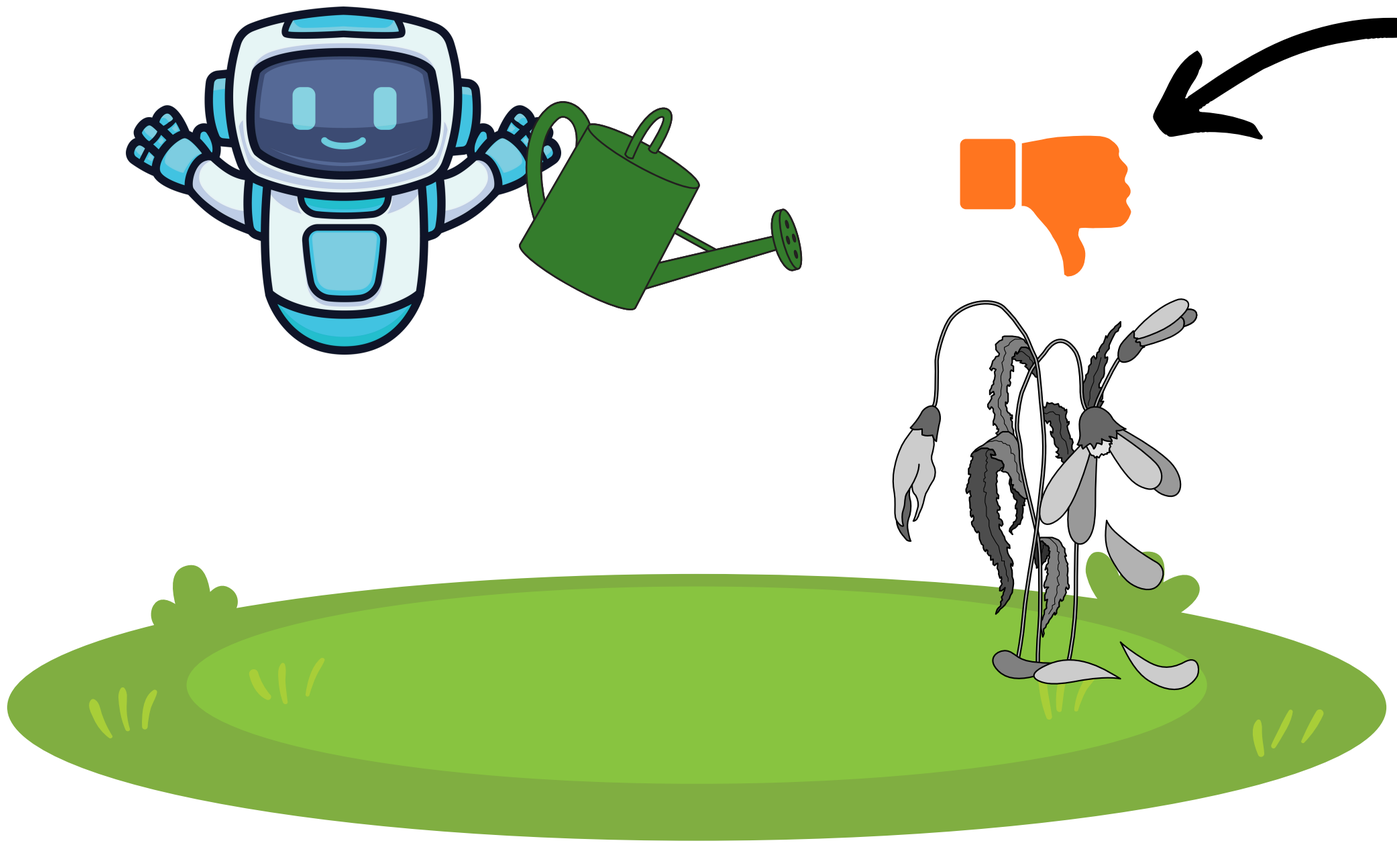


Action

An action is the **agent's behavior** at a specific moment

Key terms

kisz@hpi.de
hpi.de/kisz

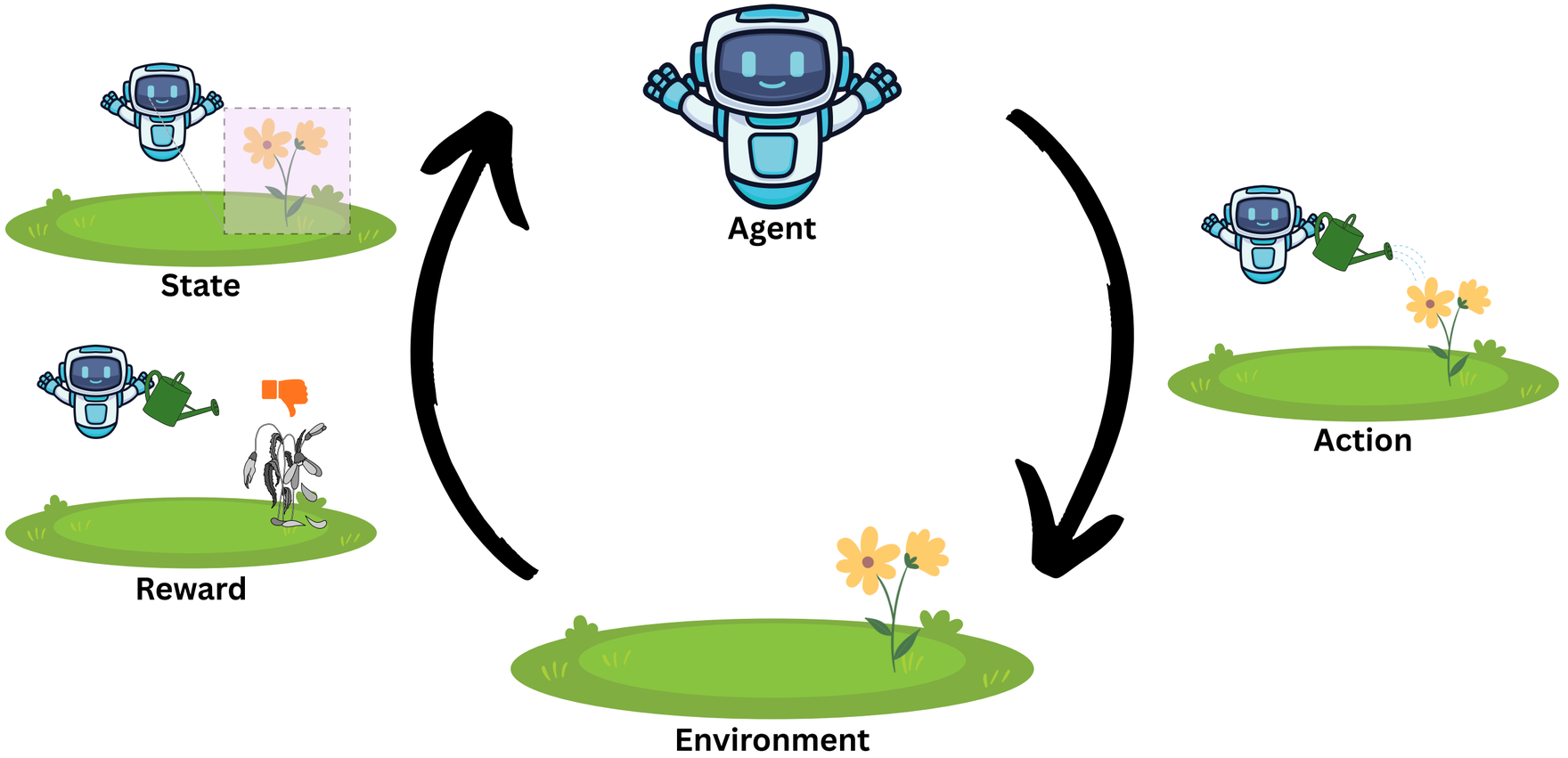


Reward

A reward is **feedback** indicating if an action was good or bad

Key terms

KI Service Zentrum by Hasso-Plattner-Institut
HPI
kisz@hpi.de
hpi.de/kisz



Timestep

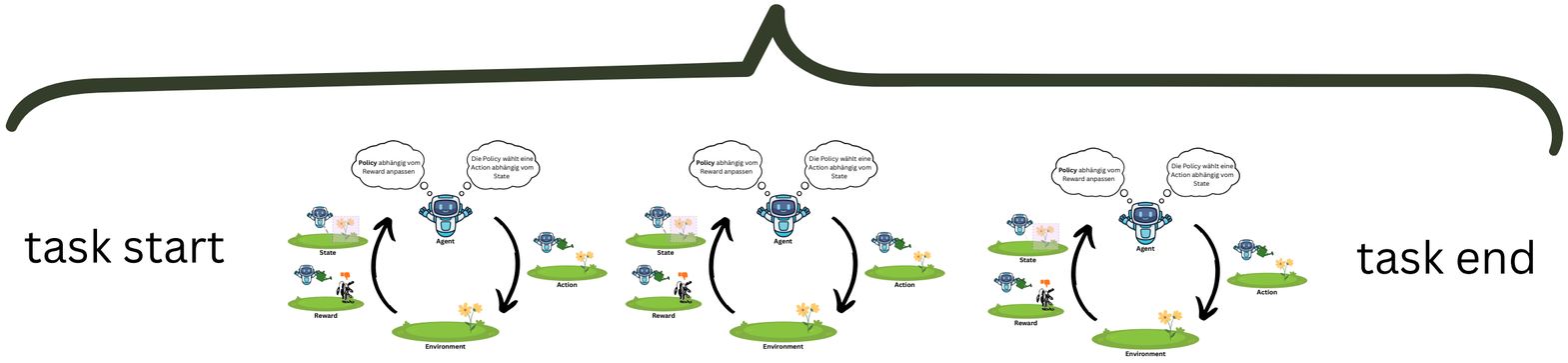
A timestep is one interaction with the environment

Key terms

kisz@hpi.de
hpi.de/kisz

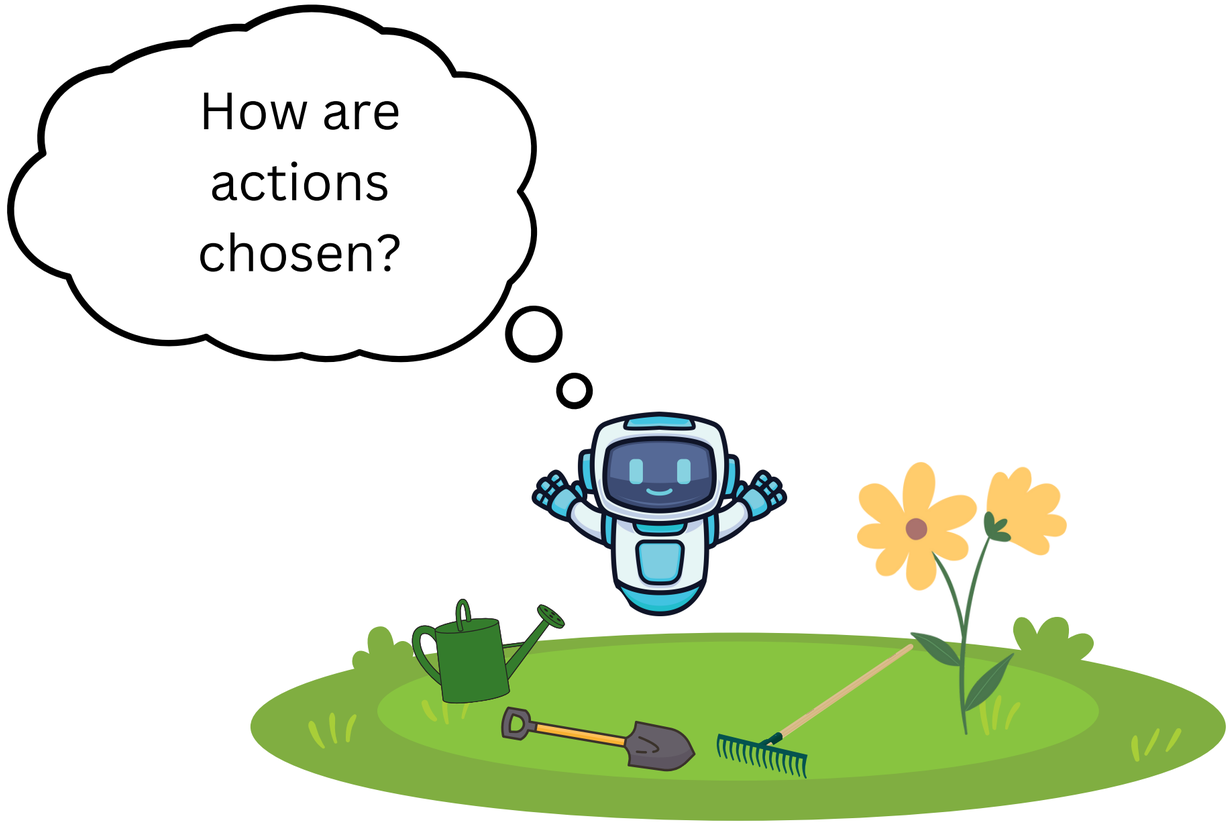
Episode

An episode is a **finished sequence of interactions** forming one complete task run



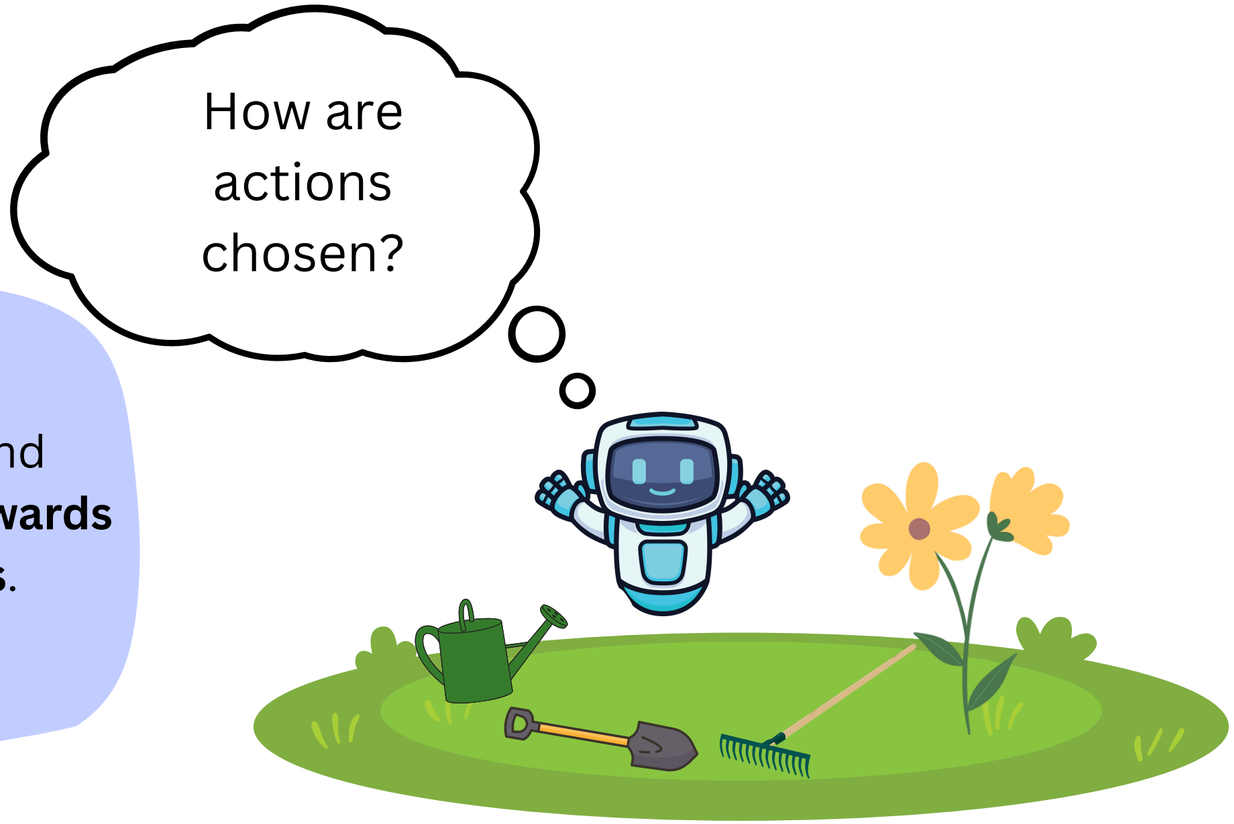
How does the agent evaluate situations? |

kisz@hpi.de
hpi.de/kisz



How does the agent evaluate situations? ||

kisz@hpi.de
hpi.de/kisz

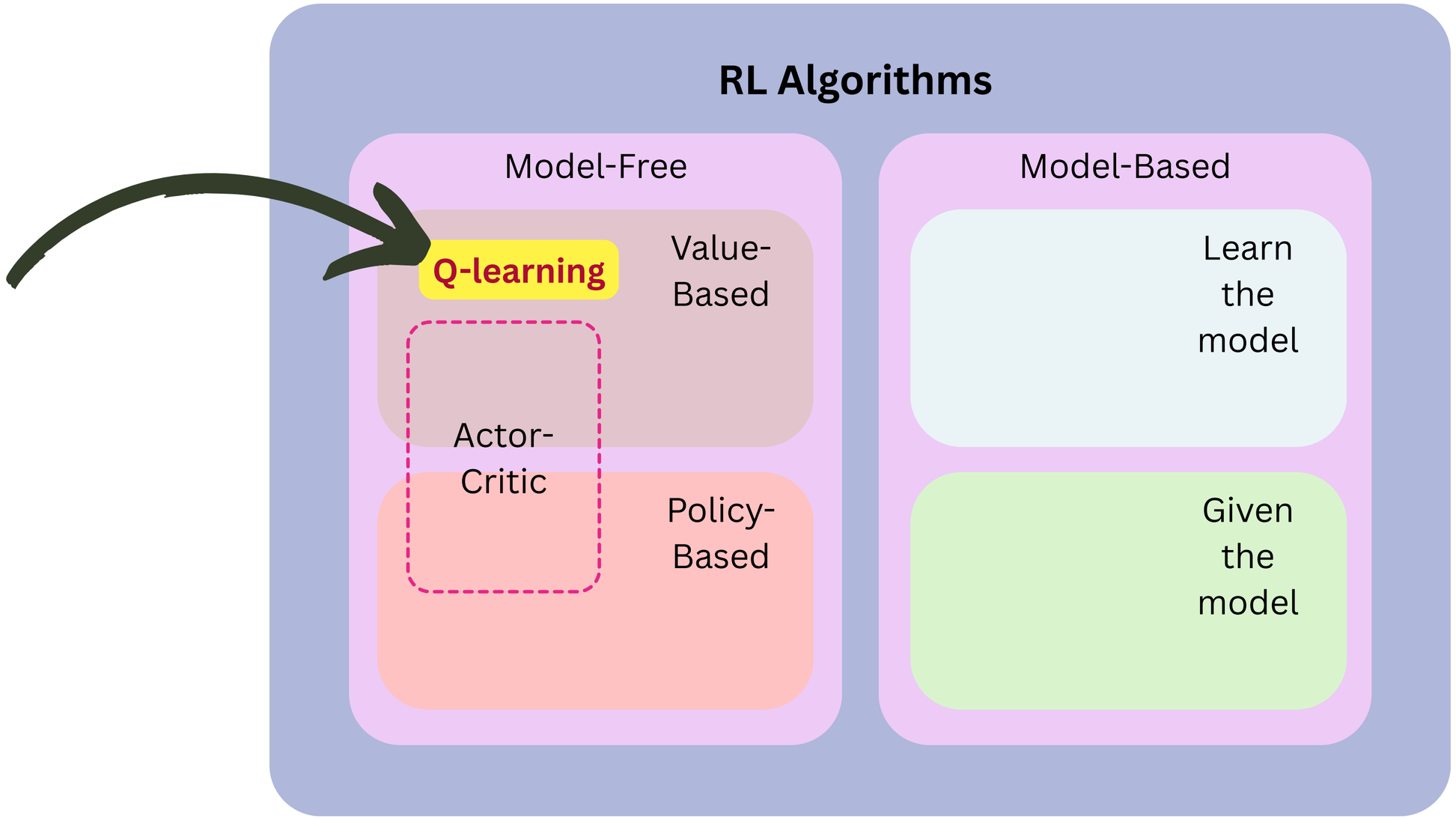


The policy chooses and adjusts actions **using rewards and value estimates**.

Algorithm

KI Service Zentrum by Hasso-Plattner-Institut HPI

kisz@hpi.de
hpi.de/kisz

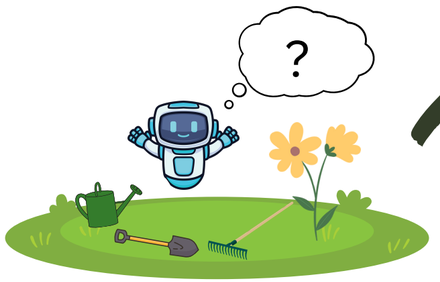


What are Q-values?

estimates **how good an action is in a given state**

used for **choosing the best action**

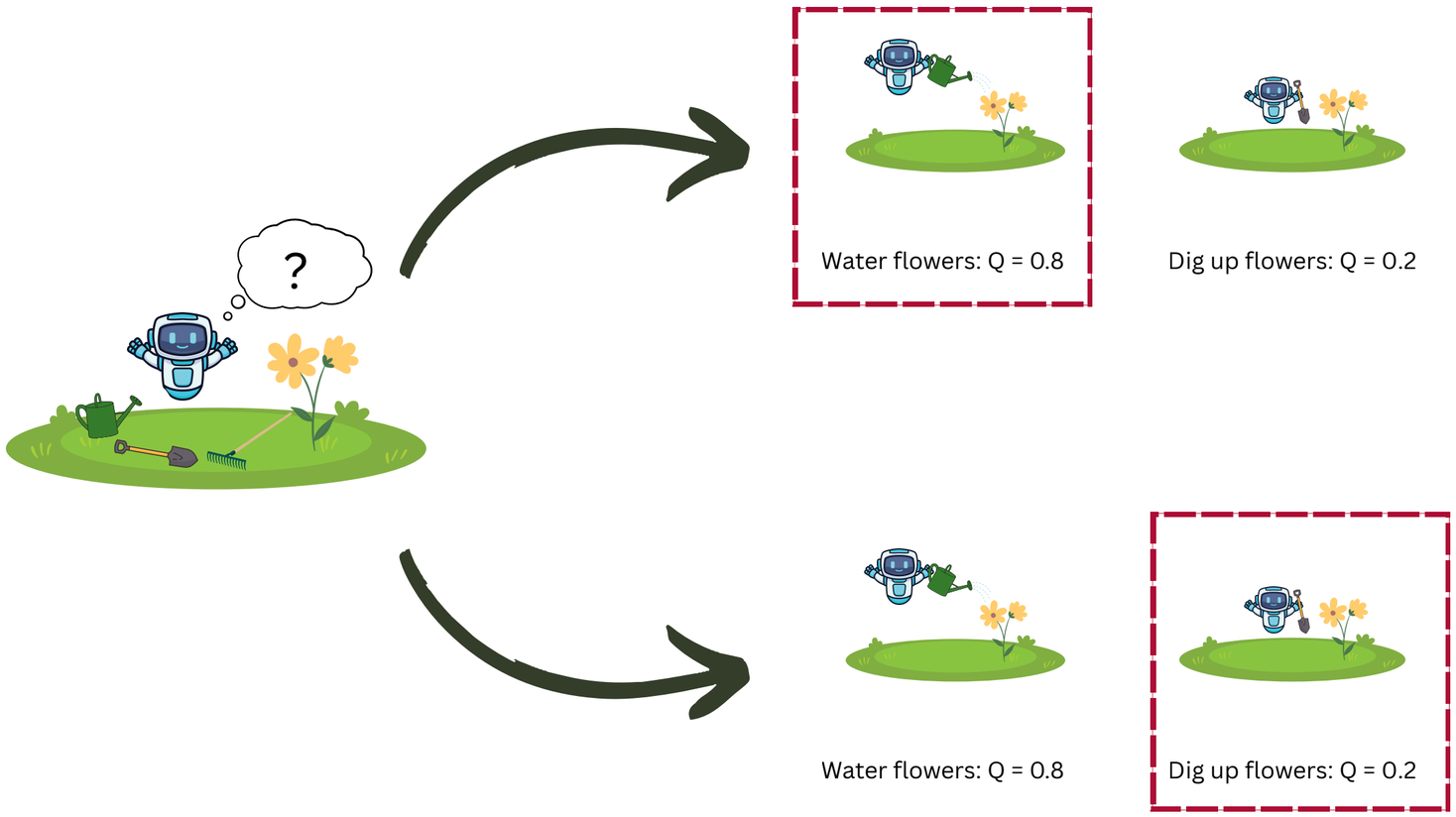
higher Q-values
→ better actions



Water flowers: Q = 0.8

Dig up flowers: Q = 0.2

Exploration vs. Exploitation



Exploitation
Choosing the best-known action to get rewards

Exploration
Trying new actions to discover better options

How are Q-values learned?

KI Service Zentrum by Hasso-Plattner-Institut
HPI
kisz@hpi.de
hpi.de/kisz

- Q-values are not known at the beginning

$$Q'(s,a) = Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

new Q-value

current Q-value

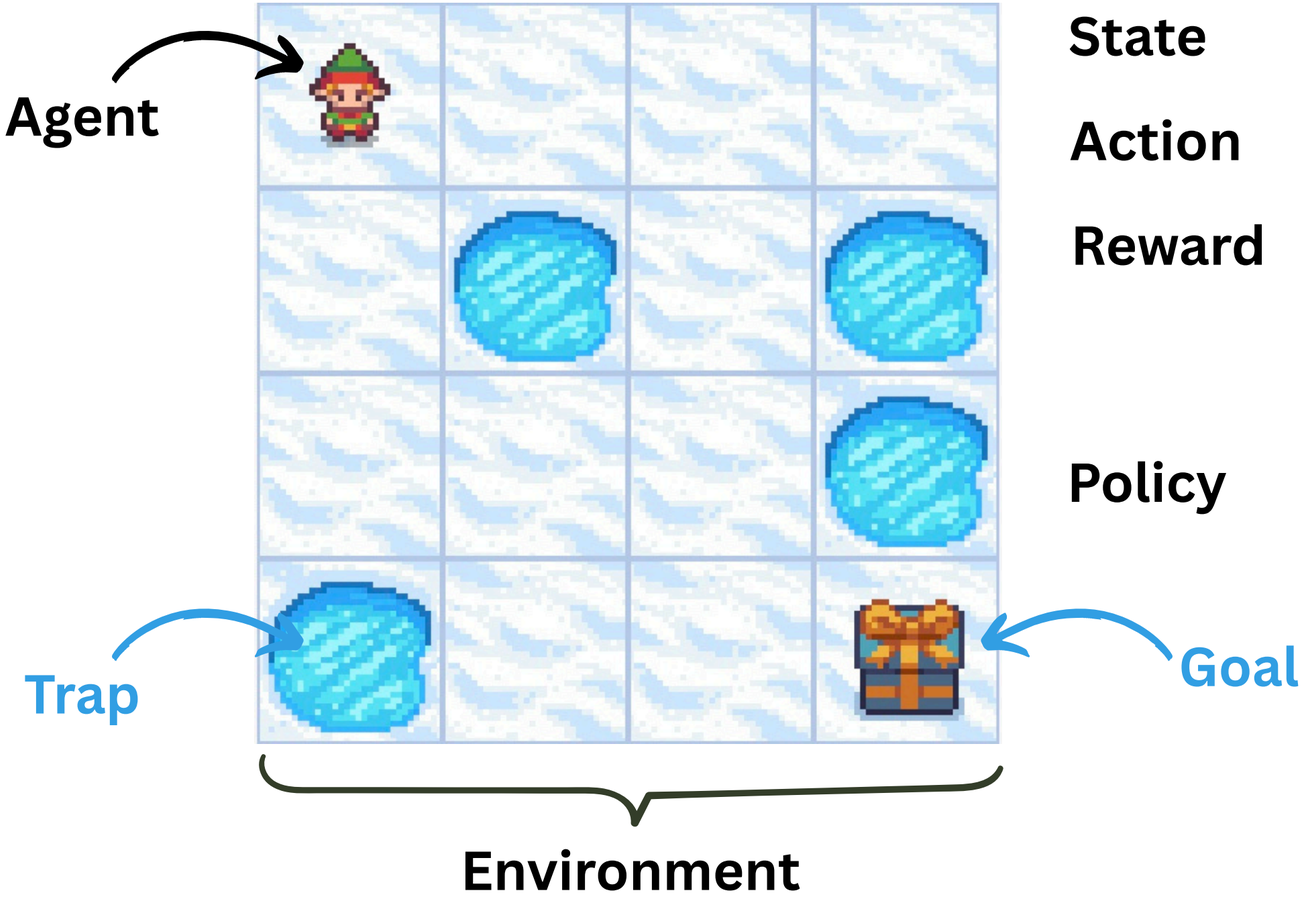
new information

Formula symbols

$Q'(s,a)$	updated value
$Q(s,a)$	current value estimate
α	learning rate
r	reward
γ	discount factor
$\max_{a'} Q(s',a')$	best next-state value
s	state
a	action
s'	next state

Example: Frozen Lake

KI Service Zentrum by Hasso-Plattner-Institut HPI
kisz@hpi.de
hpi.de/kisz



State = Position in the grid

Action = if possible, left, right, up, down

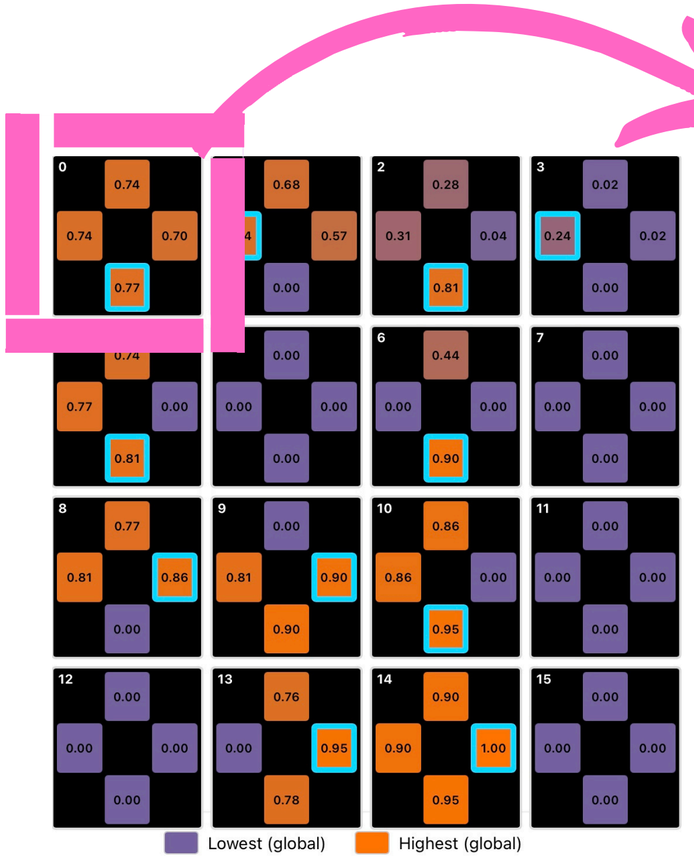
Reward = +1 for reaching the goal
0 for every step on a frozen tile
0 for falling into a hole

Policy = choose the action with the best Q-value

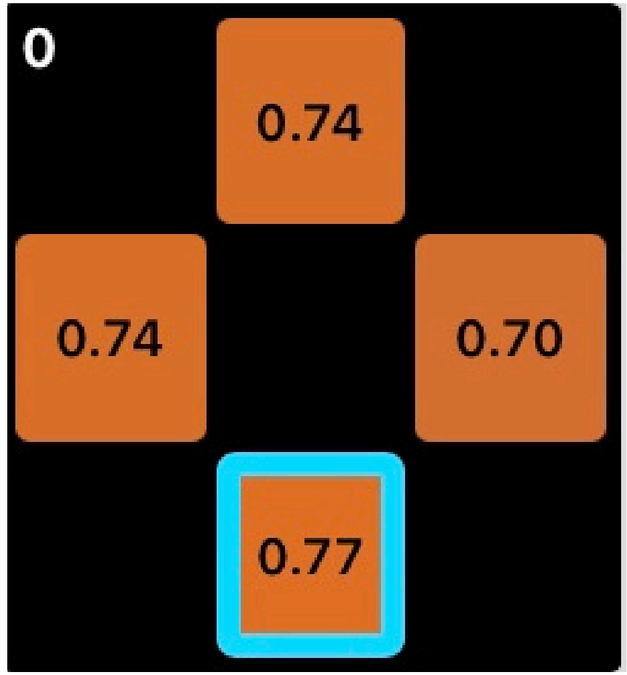
Example: Frozen Lake

KI Service Zentrum by Hasso-Plattner-Institut
HPI
kisz@hpi.de
hpi.de/kisz

Environment



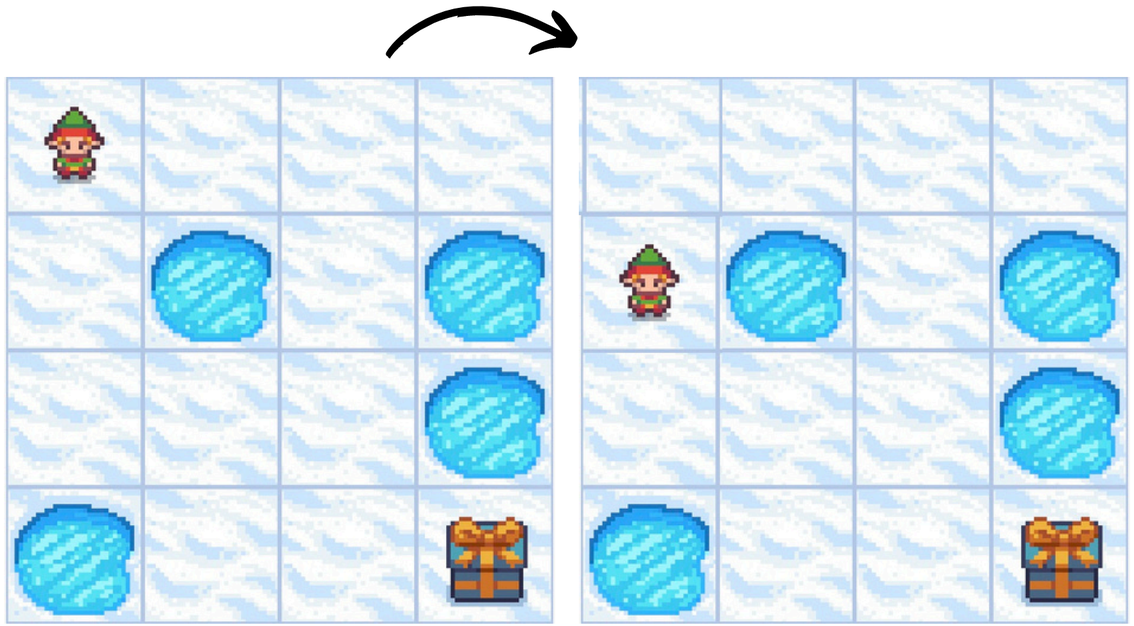
Q-Table

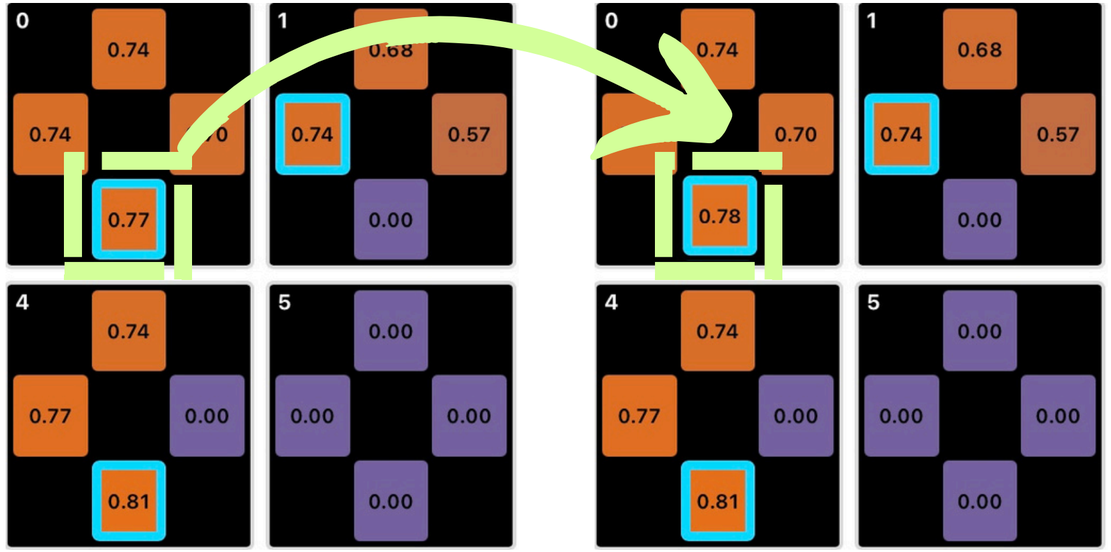


Q-Table for
the starting position

Example: Frozen Lake

KI Service Zentrum by Hasso-Plattner-Institut
HPI
kisz@hpi.de
hpi.de/kisz





$$Q(s,a) = 0.77$$

$$R = 0$$

$$\max Q(s',a') = 0.81$$

$$\alpha = 0.25$$

$$\gamma = 1.0$$

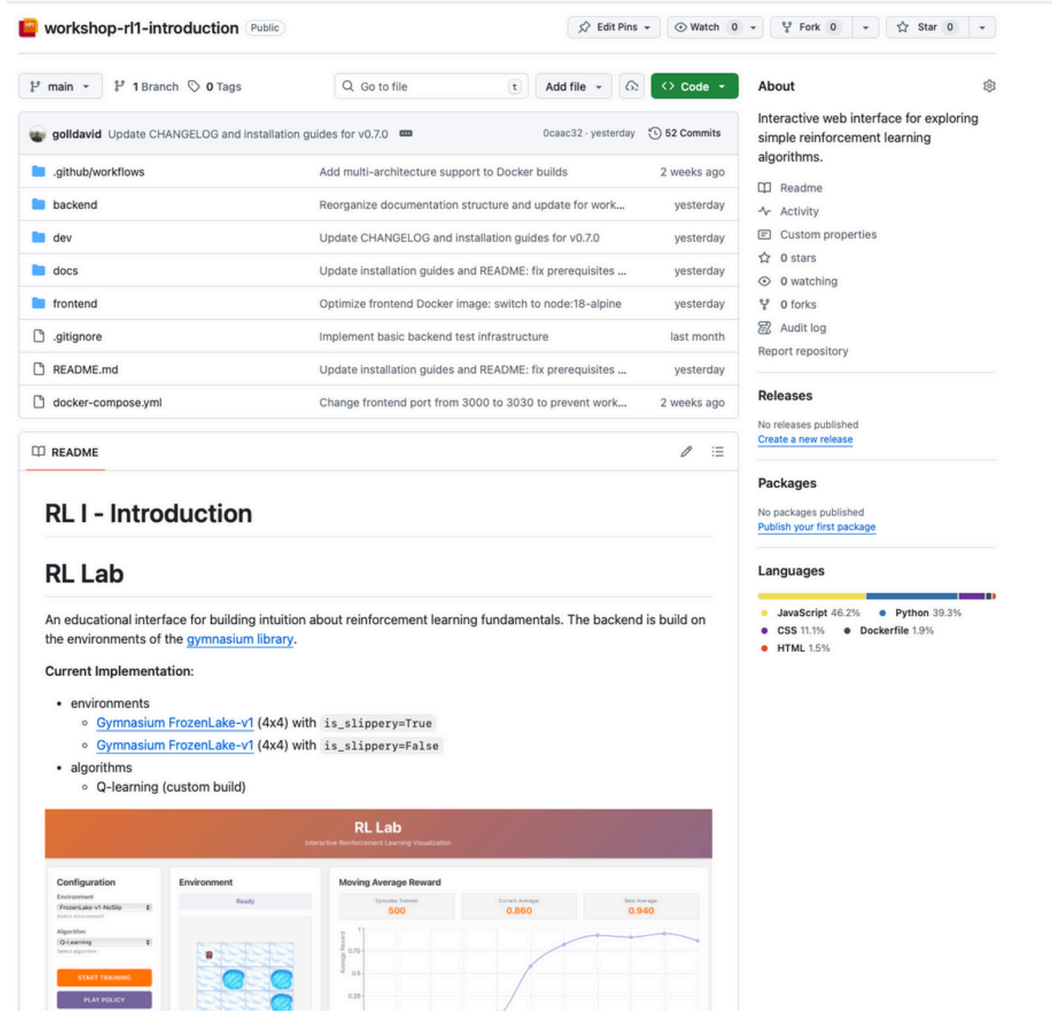
$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\big)$$

$$Q'(s,a) = 0.77 + 0.25(0.81 - 0.77) = 0.78$$

Try it out

https://github.com/aihpi/workshop-rl1-introduction

kisz@hpi.de
hpi.de/kisz



What learning parameters are there?

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz

$$Q'(s,a) = Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

$$a = \begin{cases} \arg\max_{a'} Q(s,a') & \text{with probability } 1-\varepsilon, \\ \text{random action} & \text{with probability } \varepsilon. \end{cases}$$

Parameter	Symbol	Meaning	Impact
learning rate	α	determines how strongly new information influences the previous Q-value	high values → fast but unstable learning low values → slow but stable learning
discount factor	γ	weights future rewards relative to immediate rewards	high values → long-term focus low values → short-term focus
exploration rate	ϵ	indicates how often random actions are chosen instead of the current best action	high values → explore many different actions low values → exploit the learned action more
number of episodes	-	number of learning cycles the model goes through	few episodes → poor learning many episodes → risk of overfitting, training time
initialization of Q-values	-	initial values of the Q-table	influences how quickly the model arrives at good decisions

Challenges in RL

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz

Exploration vs. Exploitation

- Balance between exploring new actions and exploiting known knowledge
- Too much exploration → inefficient
- Too little → gets stuck in poor strategies

Reward Shaping

- Rewards must encourage the desired behavior
- Poor or overly simple reward designs can lead to unexpected strategies (CAREFUL!)

Learning Stability & Convergence

- Updates can become unstable or diverge

Sample Efficiency

- RL often requires many interactions
- In real scenarios (e.g., robotics) this is costly or slow

Partial Observability & Uncertainty

- The agent often cannot observe the full state of the environment
- Must still make robust decisions

Scalability & Complexity

- Large state or action spaces make learning computationally expensive/infeasible
- Often requires function approximation (e.g., neural networks)

What have we learned so far?

KI Service Zentrum by Hasso-Plattner-Institut HPI

kisz@hpi.de
hpi.de/kisz

Basics

- **Agent-Environment interaction**
- **Policy learns to maximize** long-term cumulative **reward**
- Learning through **trial and error**
- Parameters: α (learning rate), γ (discount factor), ε (exploration rate)

Specifically

- **Q-Learning**
- **Q-Value function:** measures the expected future reward of an action in a specific state
- **Gym library** (FrozenLake)

Potential and Limitations

- **Potential:** Provides a framework for continual learning via interaction (core idea of RL is closely related to biological learning)
- **Challenges:** Sample inefficiency, sim2real, reward design, instability, safety,

Outlook

KI Service Zentrum by Hasso-Plattner-Institut

kisz@hpi.de
hpi.de/kisz

Deep Reinforcement Learning

Procedure in an RL project

Choice of algorithm

Current application areas

KI Service Zentrum
by Hasso-Plattner-Institut
HPI

kisz@hpi.de

hpi.de/kisz

Your opinion is relevant!

QR code to feedback form

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt