ECE 1508: Reinforcement Learning

Chapter 7: Applications and Advancements in DRL

Ali Bereyhi

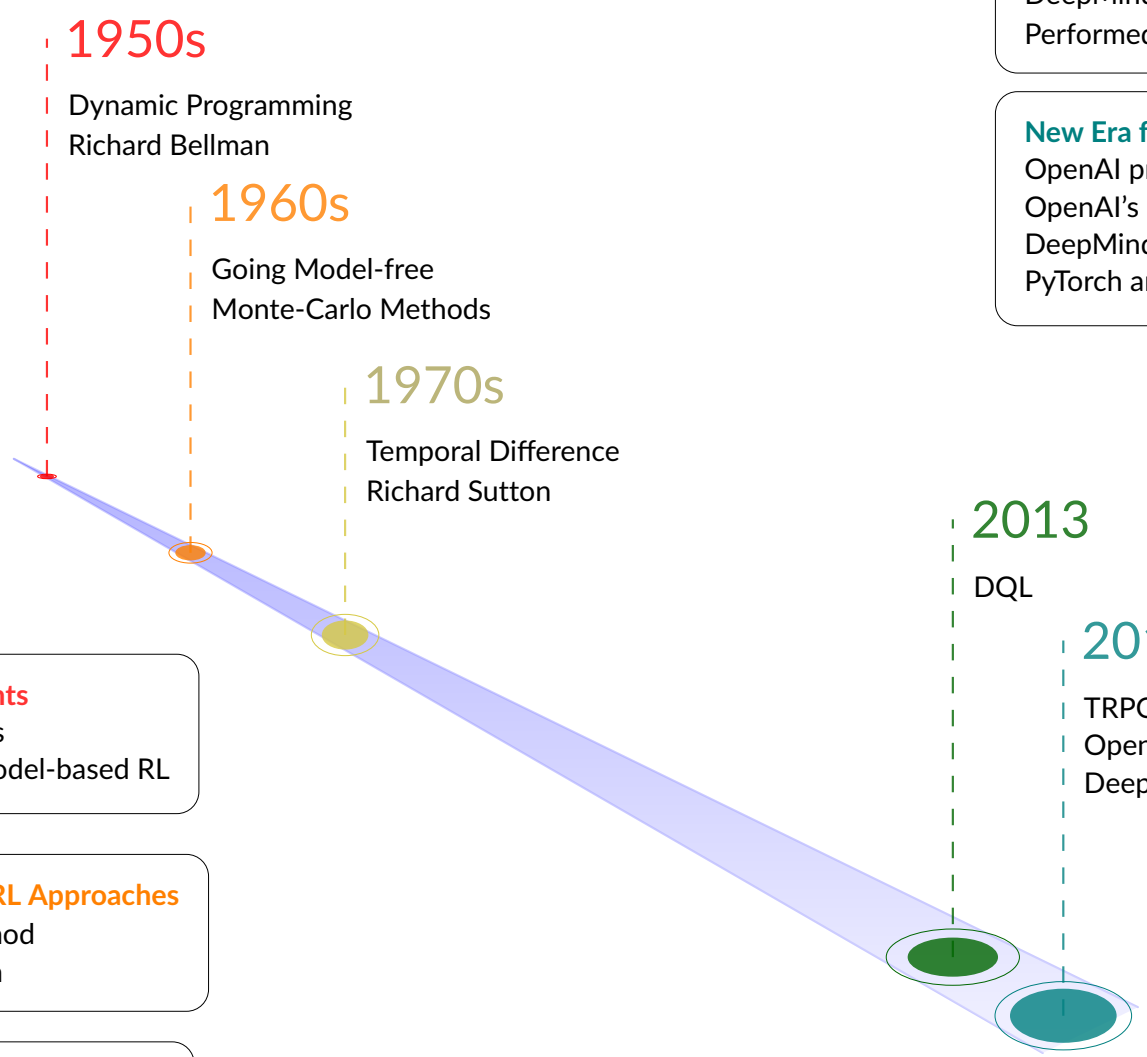
ali.bereyhi@utoronto.ca

Department of Electrical and Computer Engineering

University of Toronto

Summer 2024

Story of RL



DQL
DeepMind played Atari games
Performed better than human

New Era for Deep RL
OpenAI proposed TRPO and PPO
OpenAI's Gym made implementation easy
DeepMind proposed AlphaZero
PyTorch and TensorFlow developed RL packages

Early Developments
Bellman Equations
Understanding Model-based RL

First Model-free RL Approaches
Monte Carlo method
No use of Bellman

Bootstrapping
Connecting DP to Monte-Carlo

DeepMind: *Inspiring RL Group*

DeepMind doubtlessly is a key actor in Deep RL

- *Founded in September 2010*
- *Used DQL to play Atari games in 2013*
  - ↳ *A breakthrough in the area of RL: introducing deep RL*
- *Google buys DeepMind in 2015*
- *Introducing AlphaGo in 2015*
  - ↳ *Beats European master player*
- *Introducing AlphaGo Zero as an improved version of Alpha Go in 2017*
- *Introducing AlphaZero as a general algorithm in 2017*
  - ↳ *It can play superhuman level Go, Chess and Shogi after 24 hours of training*

DeepMind: *Inspiring RL Group*

Playing Atari with Deep Reinforcement Learning

Volodymyr Mnih Koray Kavukcuoglu David Silver Alex Graves Ioannis Antonoglou
Daan Wierstra Martin Riedmiller
DeepMind Technologies

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

[1]Google DeepMind, 5 New Street Square, London EC4A 3TW, UK.
*These authors contributed equally to this work.

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

ARTICLE

doi:10.1038/nature24270

Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver[1,2]*†, Thomas Hubert[1]*, Julian Schrittwieser[1]*, Ioannis Antonoglou[1], Matthew Lai[1], Arthur Guez[1], Marc Lanctot[1], Laurent Sifre[1], Dharshan Kumaran[1], Thore Graepel[1], Timothy Lillicrap[1], Karen Simonyan[1], Demis Hassabis[1]†

Open AI: *A New Player*

↑ "OpenAI Gym Beta". OpenAI Blog. March 20, 2017. Retrieved March 2, 2018.

↑ "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free". WIRED. April 27, 2016. Retrieved March 2, 2018. "This morning, OpenAI will release its first batch of AI software, a toolkit for building artificially intelligent systems by way of a technology called "reinforcement learning""

↑ Shead, Sam (April 28, 2016). "Elon Musk's $1 billion AI company launches a 'gym' where developers train their computers". Business Insider. Retrieved March 3, 2018.

Trust Region Policy Optimization

John Schulman JOSCHU@EECS.BERKELEY.EDU
Sergey Levine SLEVINE@EECS.BERKELEY.EDU
Philipp Moritz PCMORITZ@EECS.BERKELEY.EDU
Michael Jordan JORDAN@CS.BERKELEY.EDU
Pieter Abbeel PABBEEL@CS.BERKELEY.EDU
University of California, Berkeley, Department of Electrical Engineering and Computer Sciences

Proximal Policy Optimization Algorithms

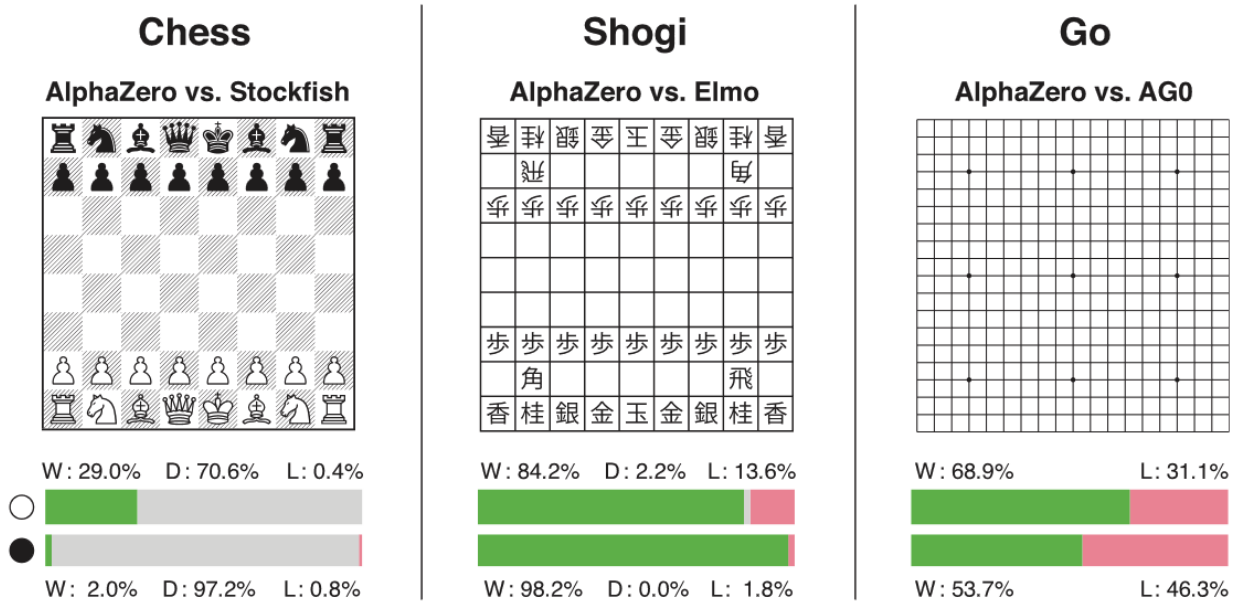
John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov
OpenAI
{joschu, filip, prafulla, alec, oleg}@openai.com

Reward learning from human preferences and demonstrations in Atari

---

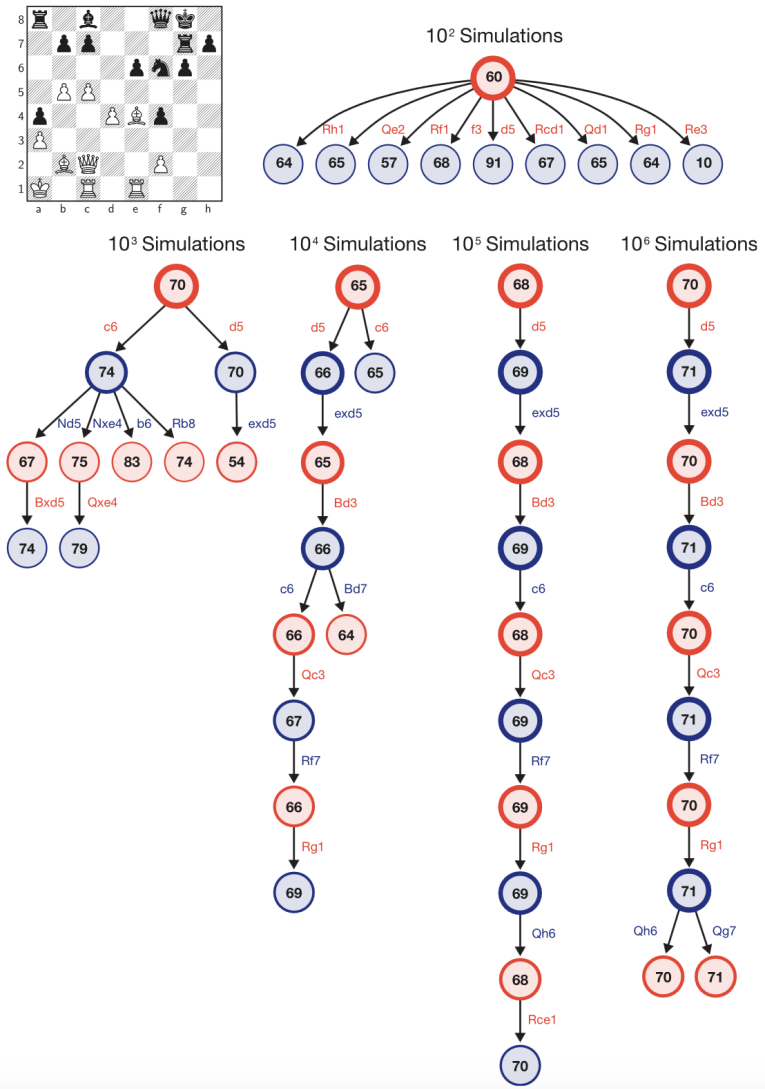
Borja Ibarz DeepMind bibarz@google.com
Jan Leike DeepMind leike@google.com
Tobias Pohlen DeepMind pohlen@google.com
Geoffrey Irving OpenAI irving@openai.com
Shane Legg DeepMind legg@google.com
Dario Amodei OpenAI damodei@openai.com

Alpha Zero



- *Conventional chess engines used to do depth search via DFS*
- *AlphaZero collect samples via tree search*
  - ↳ *It uses these samples as dataset*
  - ↳ *It learns using an actor-critic method*

Alpha Zero: *Monte Carlo Search Trees*



Legged Robot: *Winner of DARPA Subterranean Challenge*

Deep RL used by team CERBERUS to train a legged robot in 2022

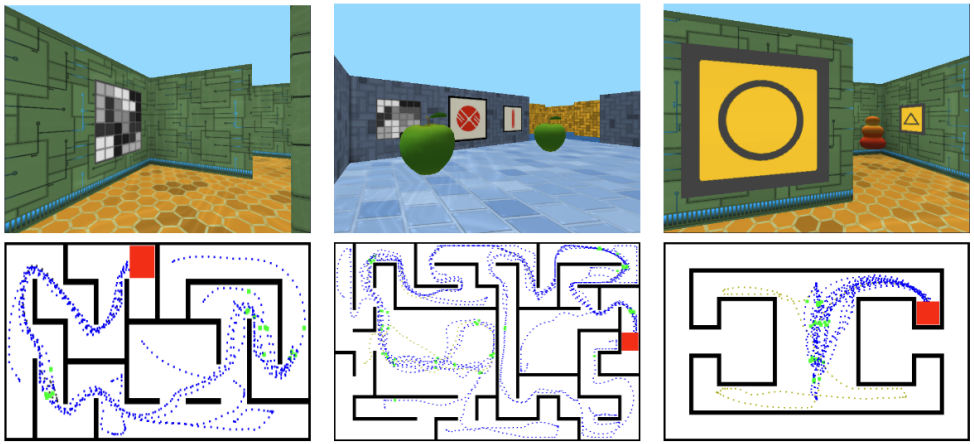
You may check out the details at this link

DeepMind: *Learning to Navigate Complex Environments*

LEARNING TO NAVIGATE
IN COMPLEX ENVIRONMENTS

Piotr Mirowski*, Razvan Pascanu*, Fabio Viola, Hubert Soyer, Andrew J. Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, Raia Hadsell

DeepMind
London, UK



DeepMind developed an algorithm for navigation in 2017

- *It was presented in ICLR 2017: see the paper*
- *A detailed presentation can be followed on YouTube*

DeepMind: *Reducing Google Energy Bill by 40%*

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

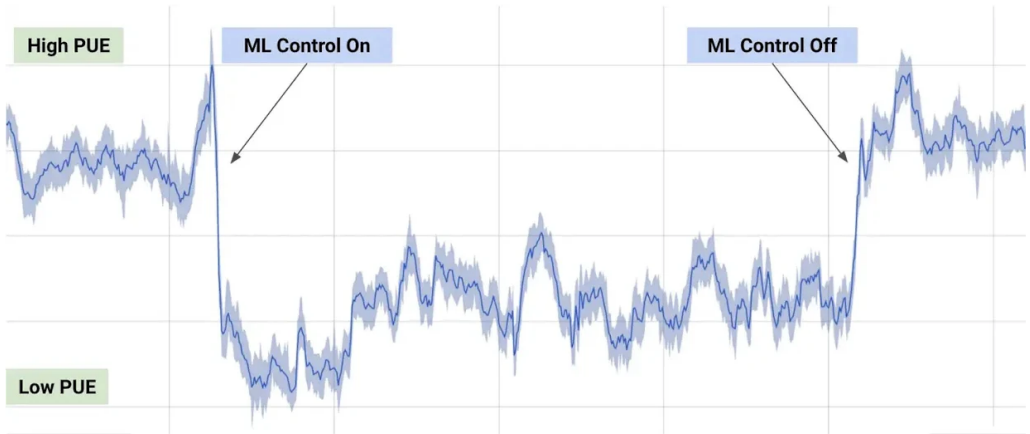
20 JULY 2016

Richard Evans, Jim Gao

Data center cooling using model-predictive control

Nevena Lazic, Tyler Lu, Craig Boutilier, Moonkyung Ryu
Google Research
{nevena, tylerlu, cboutilier, mkryu}@google.com

Eehern Wong, Binz Roy, Greg Imwalle
Google Cloud
{ejwong, binzroy, gregi}@google.com



DeepMind Blog: *Learn About Cool Applications*

Check about other cool applications at the DeepMind Blog

All categories

RESEARCH

AI achieves silver-medal standard solving International Mathematical Olympiad problems

Breakthrough models AlphaProof and AlphaGeometry 2 solve advanced reasoning...

25 JULY 2024

RESEARCH

Google DeepMind at ICML 2024

Exploring AGI, the challenges of scaling and the future of multimodal generative AI

19 JULY 2024

TECHNOLOGIES

Generating audio for video

Video-to-audio research uses video pixels and text prompts to generate rich...

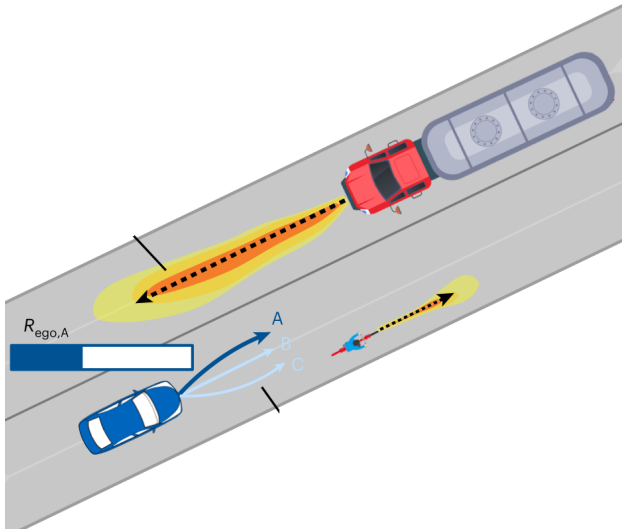
17 JUNE 2024

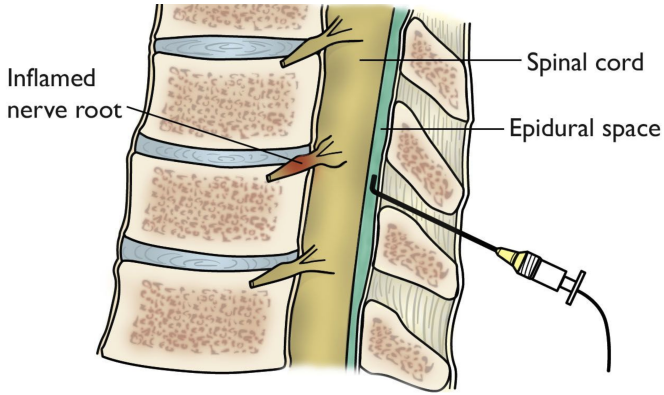
Challenge of *Sparse Rewarding*

In many RL problems, we deal with sparse rewards

- *In Frozen Lake game, we only get the reward at the end!*

Can this be tolerated in many practical problems?





Waiting for those sparse rewards could be very dangerous!

A Solution: *Reward Shaping*

If we could formulate what we want: maybe, we could shape the reward

- *In Assignment 1, we did it for simple Frozen Lake game*
  - ↳ *Just need to add some negative rewards in between*
- *By this approach, we make much denser reward*

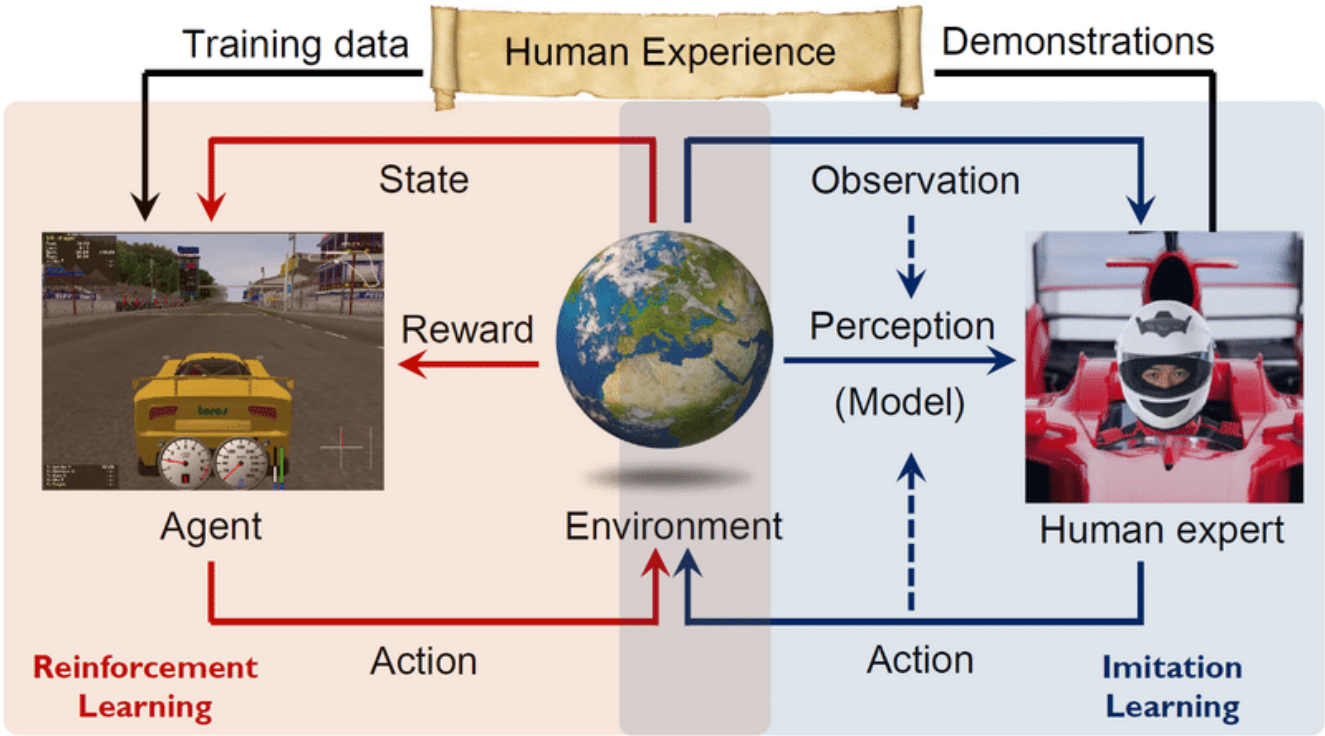
\+ *Can we always do it?*

– *Not really!*

In practice, we cannot necessarily formulate the types of rewards we want

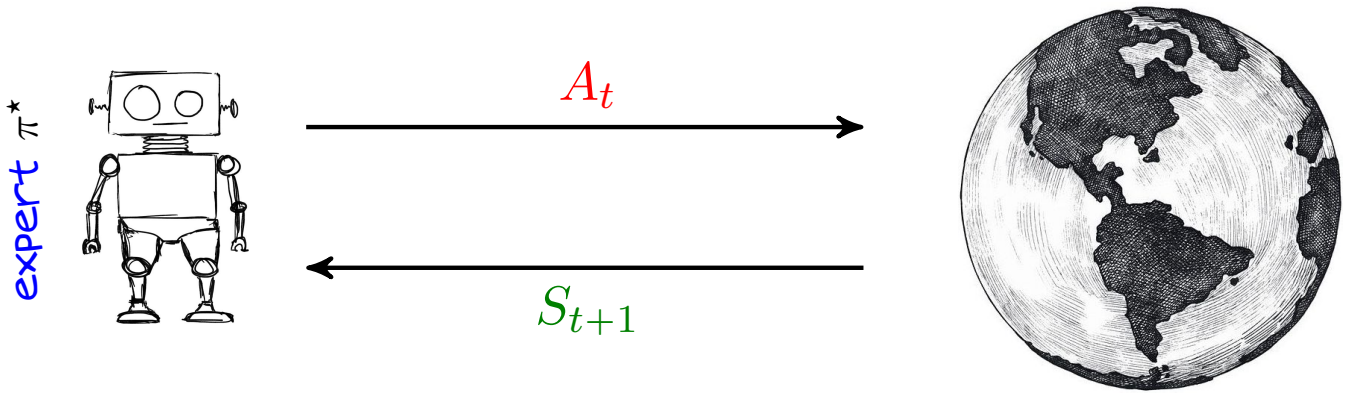
- *We often have access only to a human expert*
  - ↳ *a professional driver or an expert surgeon*

Imitation Learning: *Learning by Expert*



In practice, we want to employ an expert to learn safely and efficiently

Behavior Cloning



We are looking for mimicking the expert

$$\min_{\boldsymbol{\theta}} \mathcal{L}\left(\pi_{\boldsymbol{\theta}}, \pi^\star\right)$$

and of course, we do it by sampling!

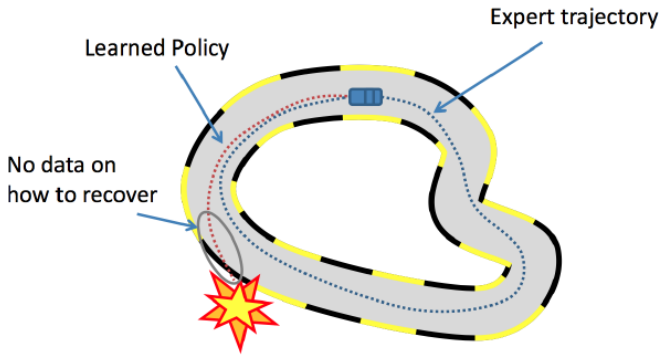
Behavior Cloning: *Not Always Efficient*

Algorithm 1: DAgger: Dataset Aggregation

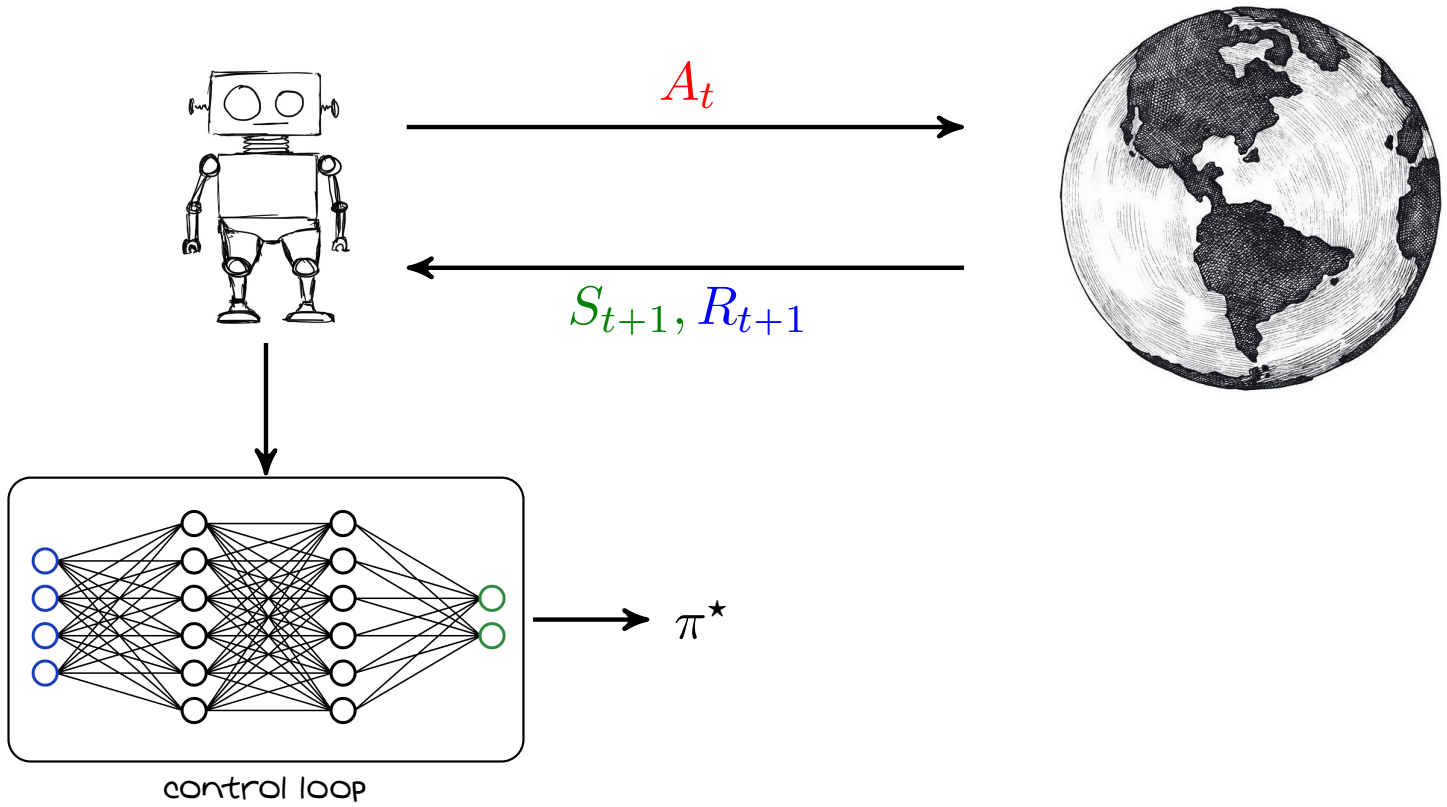
Data: π^*
Result: $\hat{\pi}^*$
$\mathcal{D} \leftarrow 0$
Initialize $\hat{\pi}$
for $i = 1$ **to** N **do**
  $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}$
  Rollout policy π_i to sample trajectory $\tau = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots\}$
  Query expert to generate dataset $\mathcal{D}_i = \{(\boldsymbol{x}_0, \pi^*(\boldsymbol{x}_0)), (\boldsymbol{x}_1, \pi^*(\boldsymbol{x}_1)), \dots\}$
  Aggregate datsets, $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$
  Retrain policy $\hat{\pi}$ using aggregated dataset $\mathcal{D}$
return $\hat{\pi}$

DAgger: *famous efficient algorithm for behavior cloning*

But, we are always limited in collecting expert samples



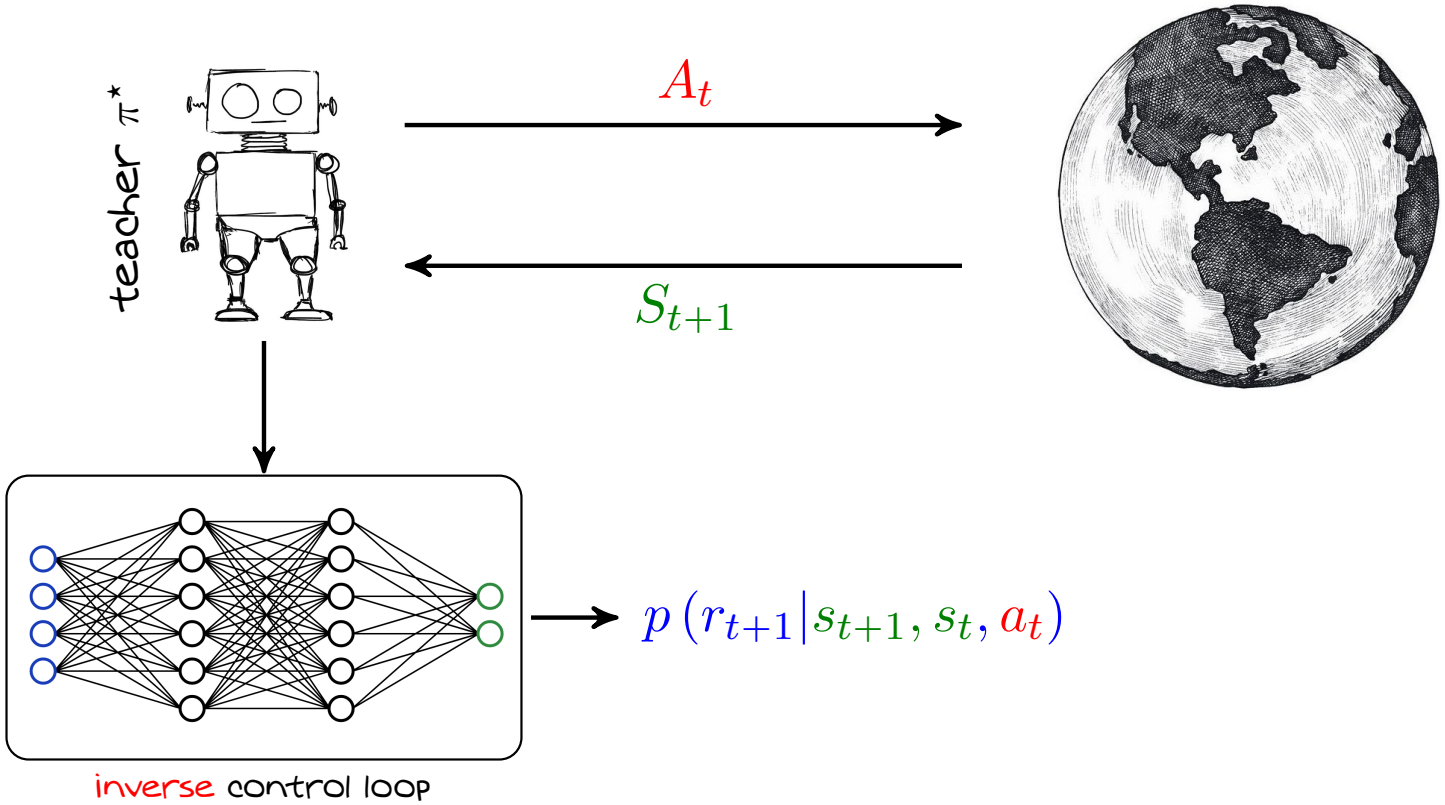
Inverse Reinforcement Learning



In classical RL setting

we collect rewards $\rightsquigarrow$ *we try to learn optimal policy*

Inverse Reinforcement Learning: *Learning Reward Function*



In inverse RL setting

a teacher plays $\rightsquigarrow$ we try to learn rewarding system

Inverse Reinforcement Learning: *Some Notes*

This problem can be cast using same formulation: we can write the optimal value function again

$$
\begin{aligned}
v_{\star}(s) &= \mathbb{E}_{\pi^{\star}}\left\{G_t | S_t = s\right\} \\
&= \mathbb{E}_{\pi^{\star}}\left\{\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} | S_t = s\right\}
\end{aligned}
$$

In RL, we know R_{t+i+1} and look for $\pi^{\star}$

- *We use function approximation, e.g., $\pi^{\star}(\cdot) \equiv f_{\mathbf{w}}(\cdot)$*

In inverse RL, we know $\pi^{\star}$ and look for R_{t+i+1}

- *We can again use function approximation, e.g., $p(r_{t+1}|\cdot) \equiv f_{\mathbf{w}}(\cdot)$*

Pieter Abbeel has a nice lecture on inverse RL: take a look at this link

Also, you may find this imitation learning lecture-note useful to learn a bit more!

Using Expert Demonstration

Not all experts can formulate their policy

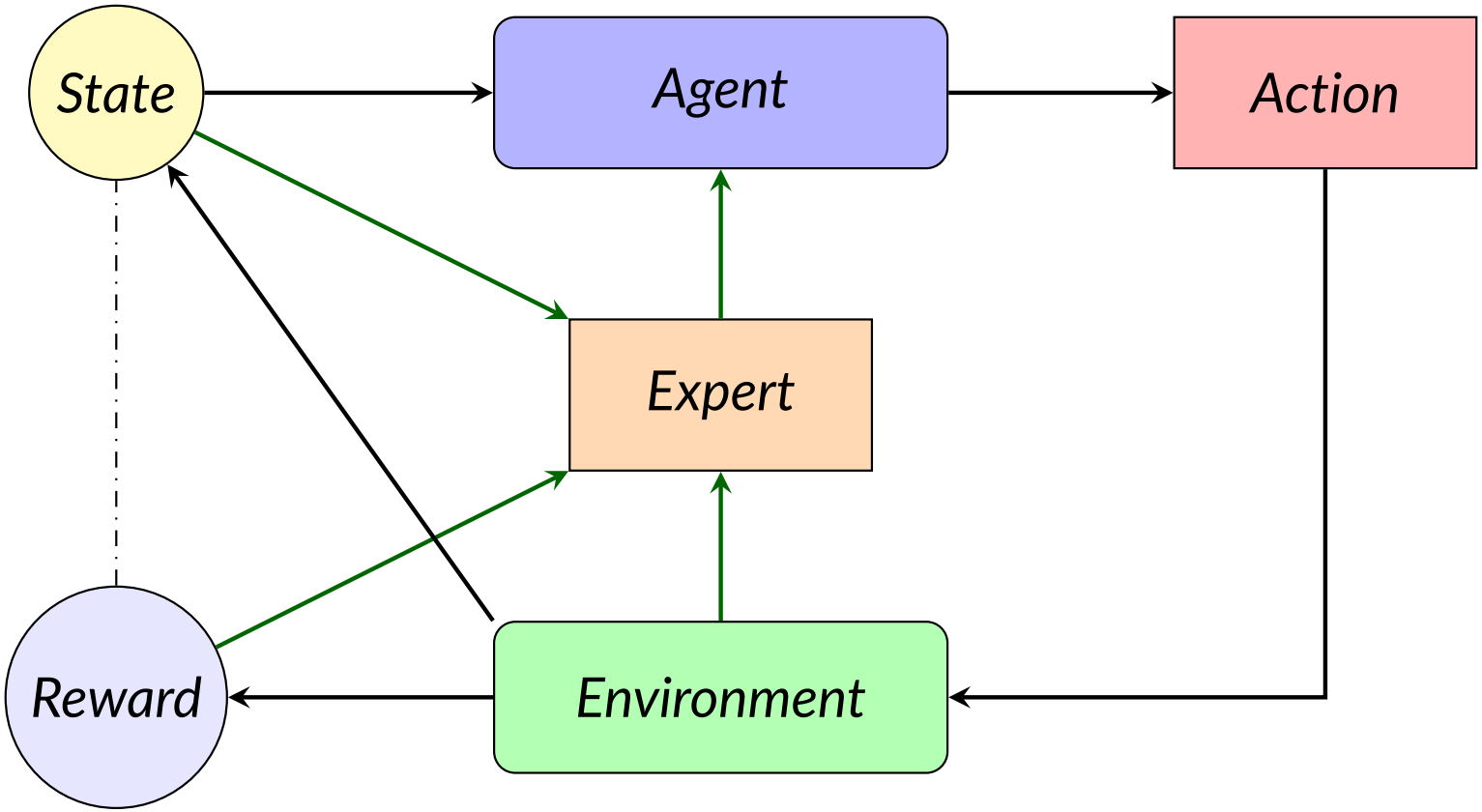
- *It might be not really easy to formulate it*
  - ↳ *A board-game player my act based on experience and intuition*
- *It might be much easier to demonstrate the policy*
  - ↳ *A professional driver could explain what they do in various conditions*

In this case learning the rewarding system is not really feasible

- *We don't exactly know optimal policy*
- *We can only describe it using expert demonstrations*

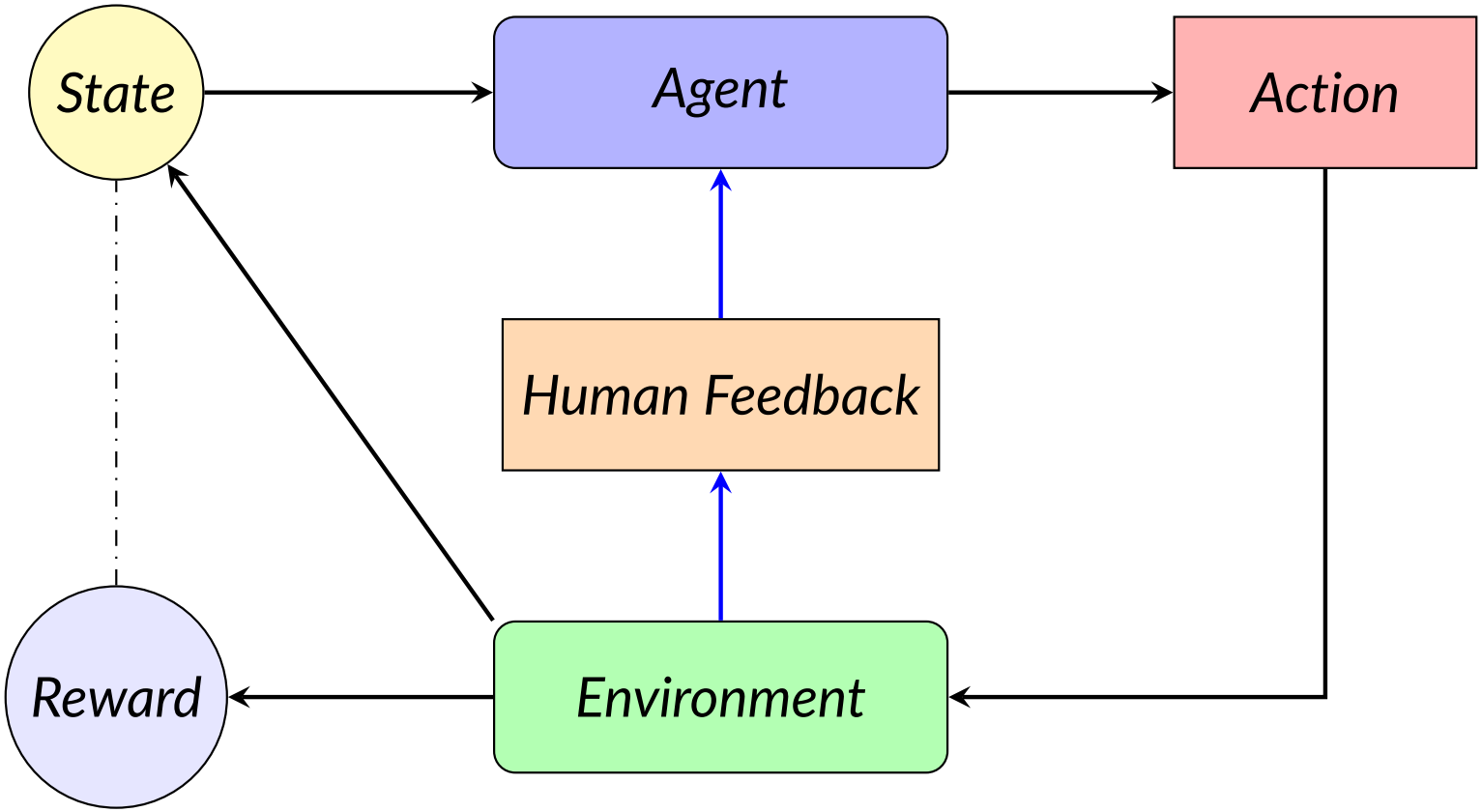
RL via *Expert Demonstration*

We could include expert demonstrations into our control loop



RLHF: *RL from Human Feedback*

This can be further extended to the use of human feedback: that human does not need to play optimal!



ChatGPT also extensively uses RLHF: see this lecture

Multi-Agent RL

Up to now, we considered other agents as a part of environment

- *But, they can have their own policies*
  - ↳ *They could collaborate to play jointly optimal*
  - ↳ *There might be adversary agents trying to impact negatively*

Take a look at this manuscript to learn about multi-agent RL

The End

Many thanks for attending the lectures . . .

- *Be confident and trust on what you have learned* 👍
  - ↳ *You are now experts in Deep RL!*
- *Feel free to reach out*
  - ↳ *Would be more than happy to help!*
  - ↳ *I would appreciate your feedback on Quercus* 💬
- *Always track the progress in the field*
  - ↳ *It's fast growing, so you should keep yourself posted* ⏭
  - ↳ *You are ready to follow any advanced topics in RL* ⤢

Next Summer, there will be a Generative AI course

↳ *We may use some Deep RL there as well* ☺