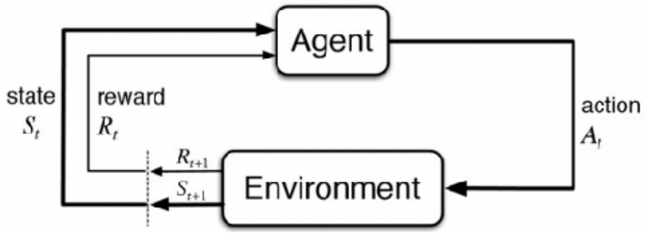
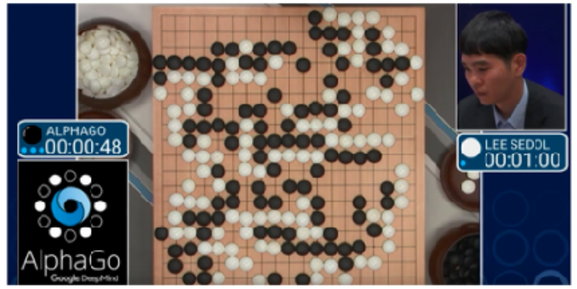
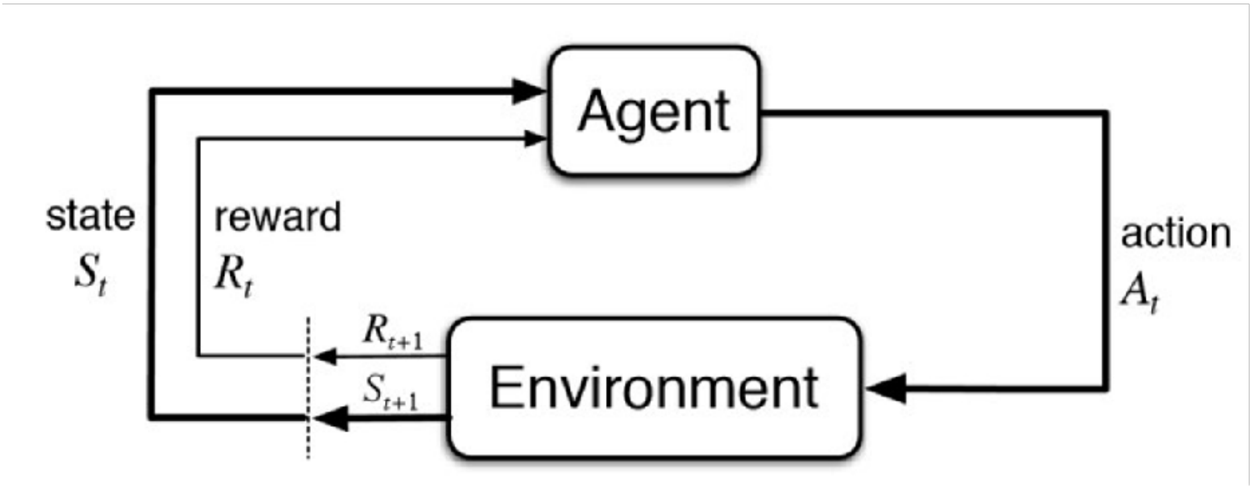
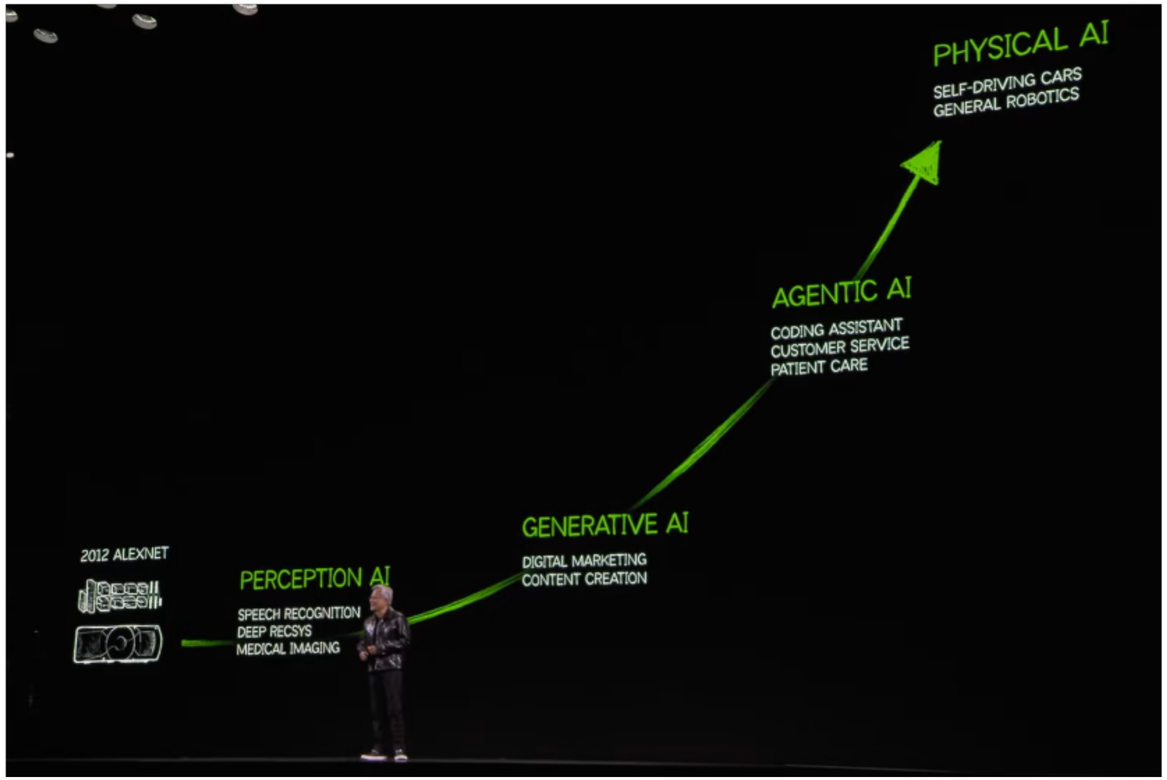




CS260R Reinforcement Learning Lecture 1: Course Overview

Bolei Zhou, UCLA Computer Science

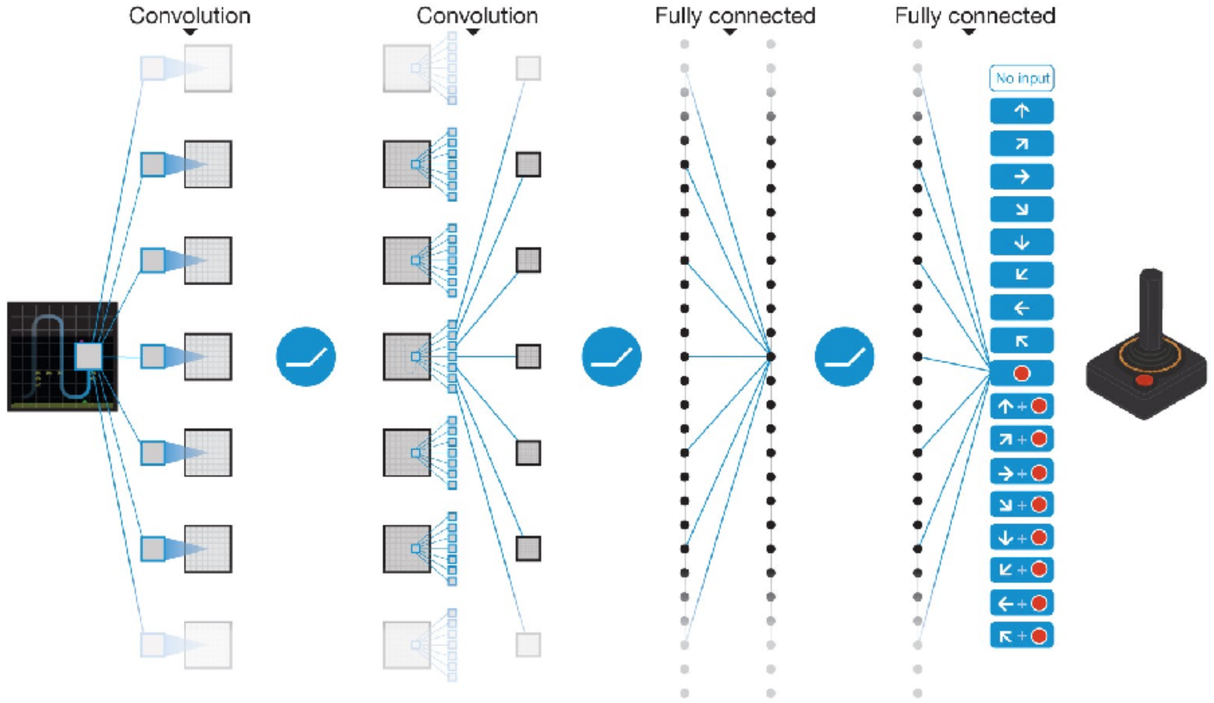




Jetson Huang at GTC

Landmark of Deep RL in Early 2014: DQN

Human-level control through deep reinforcement learning, Nature

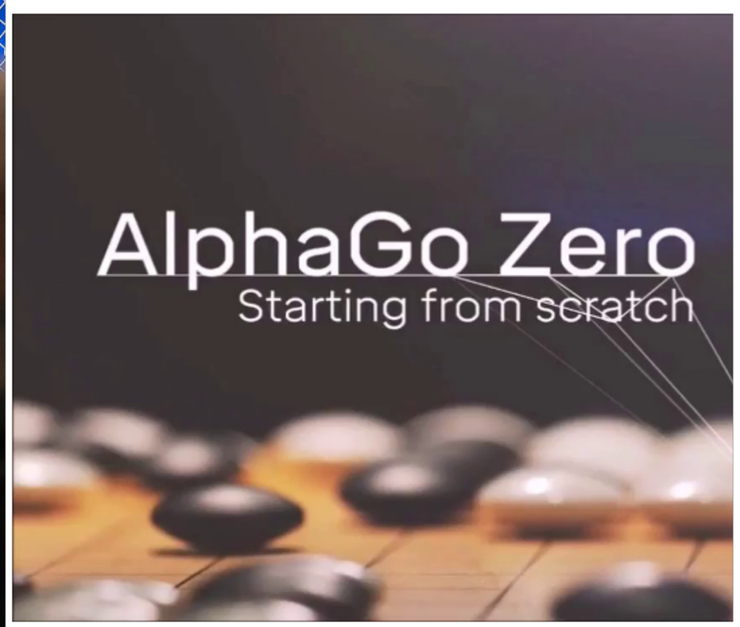


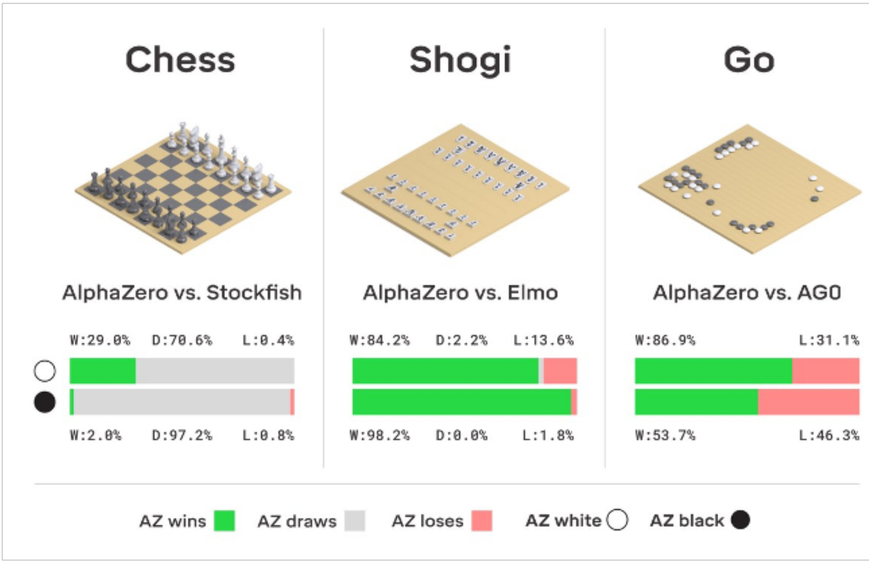
Deep Q network (DQN) was one of the key pieces of work that helped convince Google to buy DeepMind

https://www.nature.com/articles/nature14236

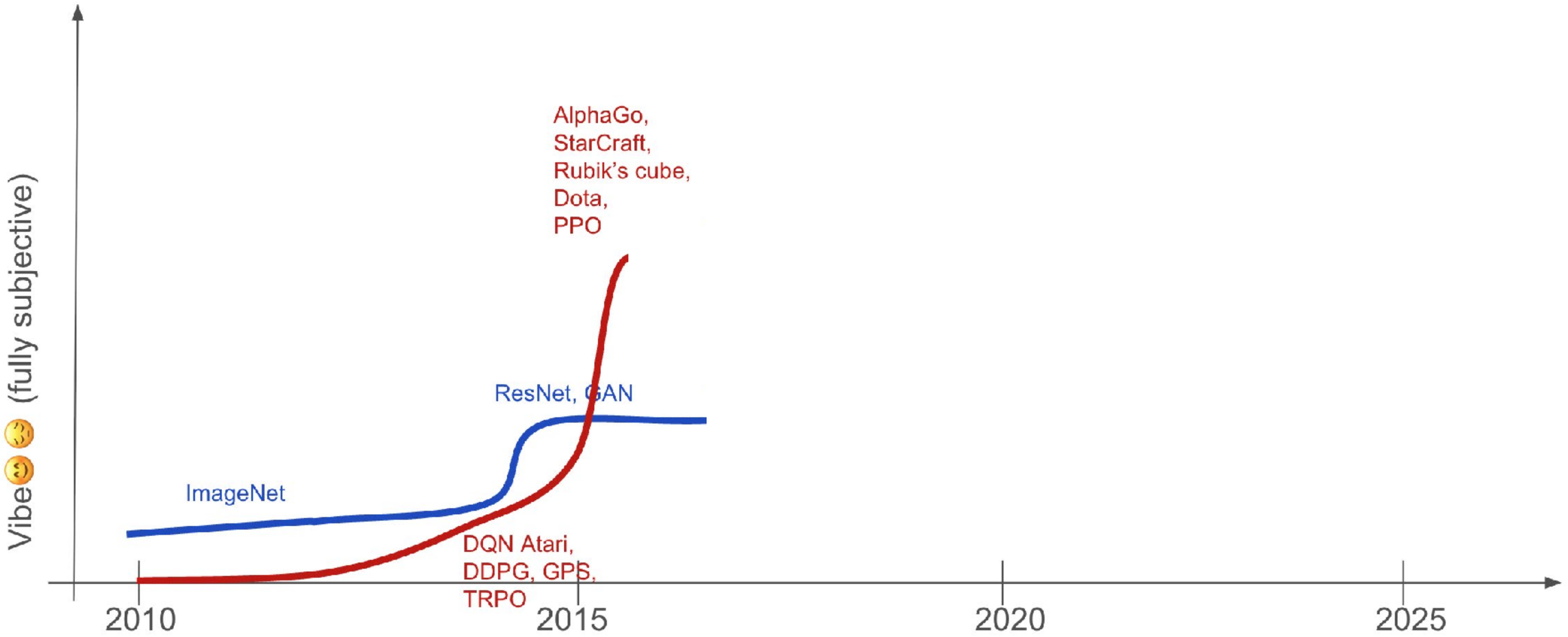
Milestone of Deep RL in March of 2016

AlphaGo series: AlphaGo, AlphaGo Zero, AlphaZero, MuZero



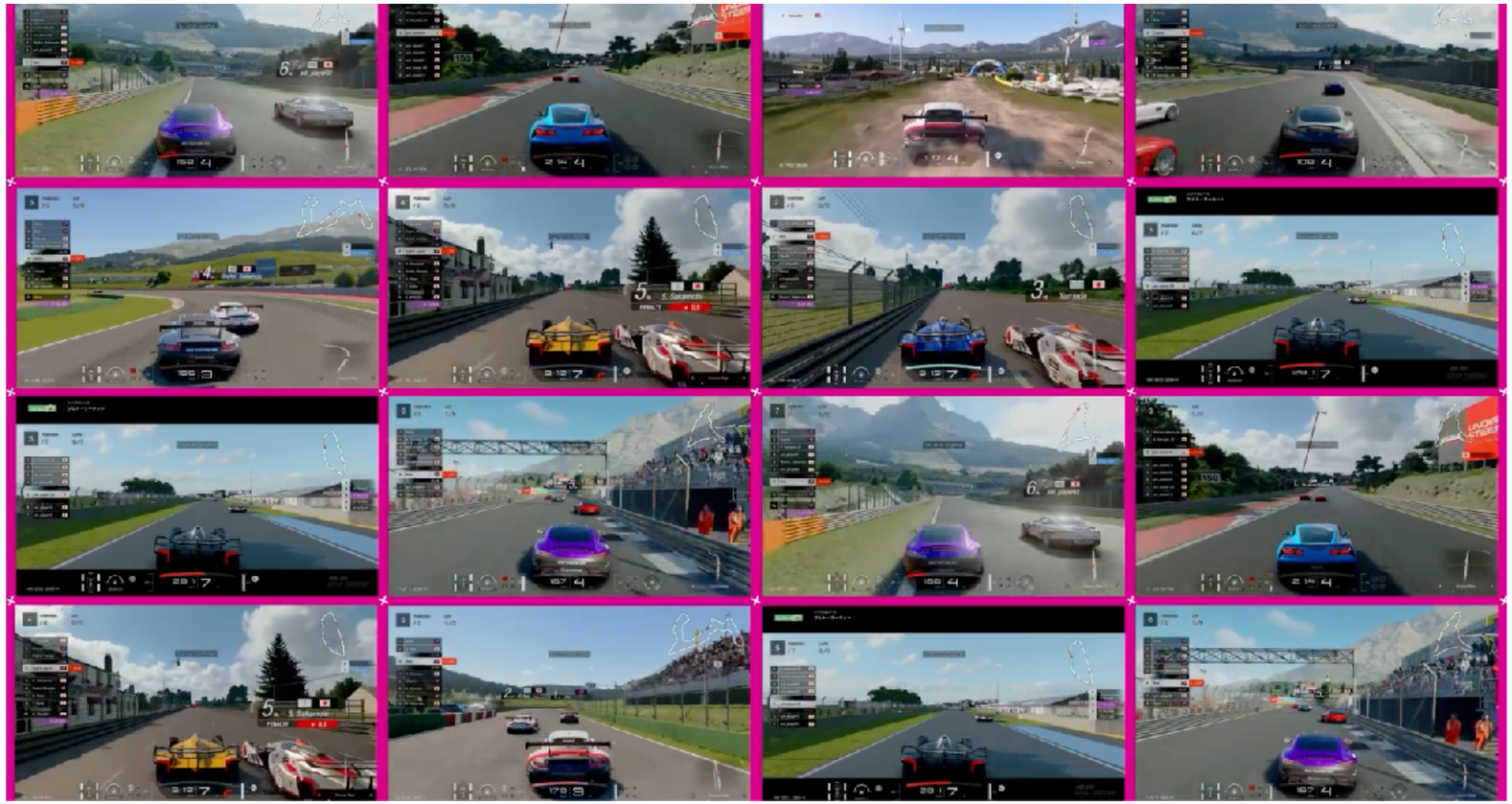


RL hype curve



Exciting progress of RL for robotics: Car Racing

Outracing champion Gran Turismo drivers with deep reinforcement learning (Nature, 2022)



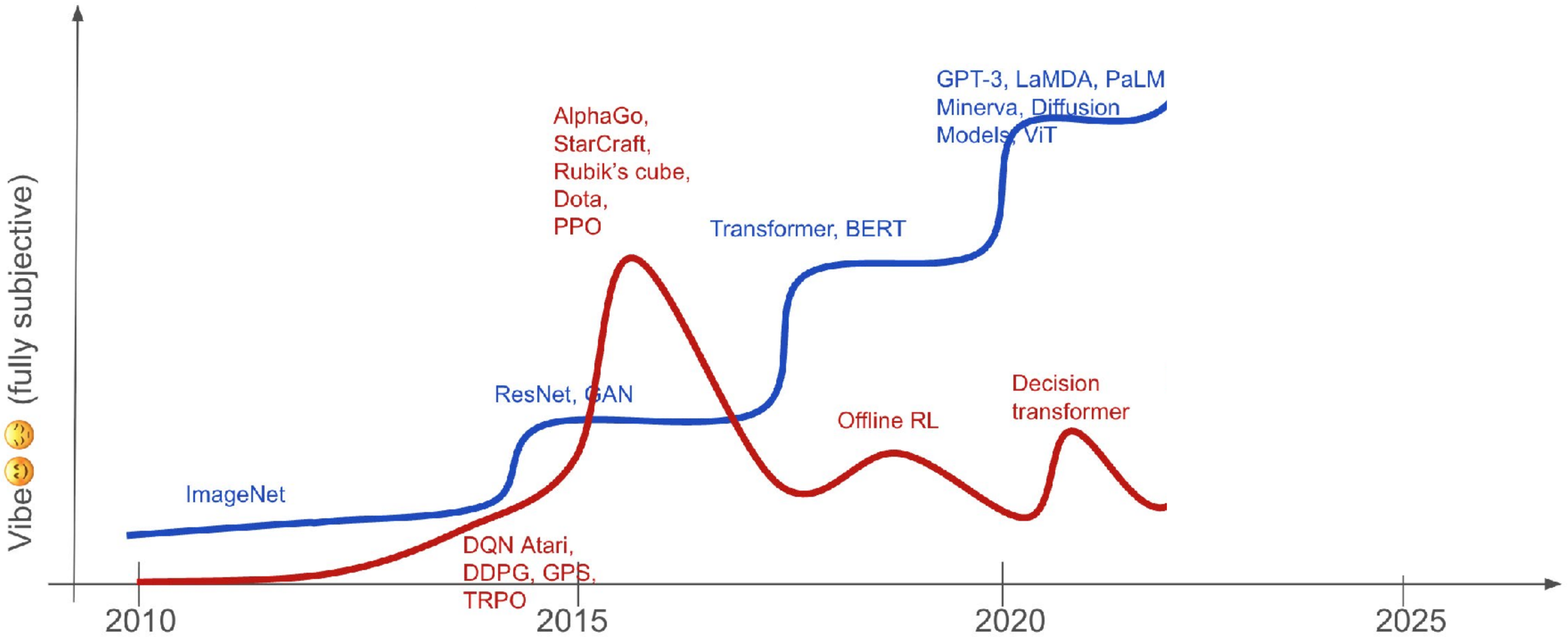
https://www.nature.com/articles/s41586-021-04357-7

Exciting progress of RL for robotics: Drone Racing

Champion-level Drone Racing using Deep Reinforcement Learning (Nature, 2023)

https://www.nature.com/articles/s41586-023-06419-4

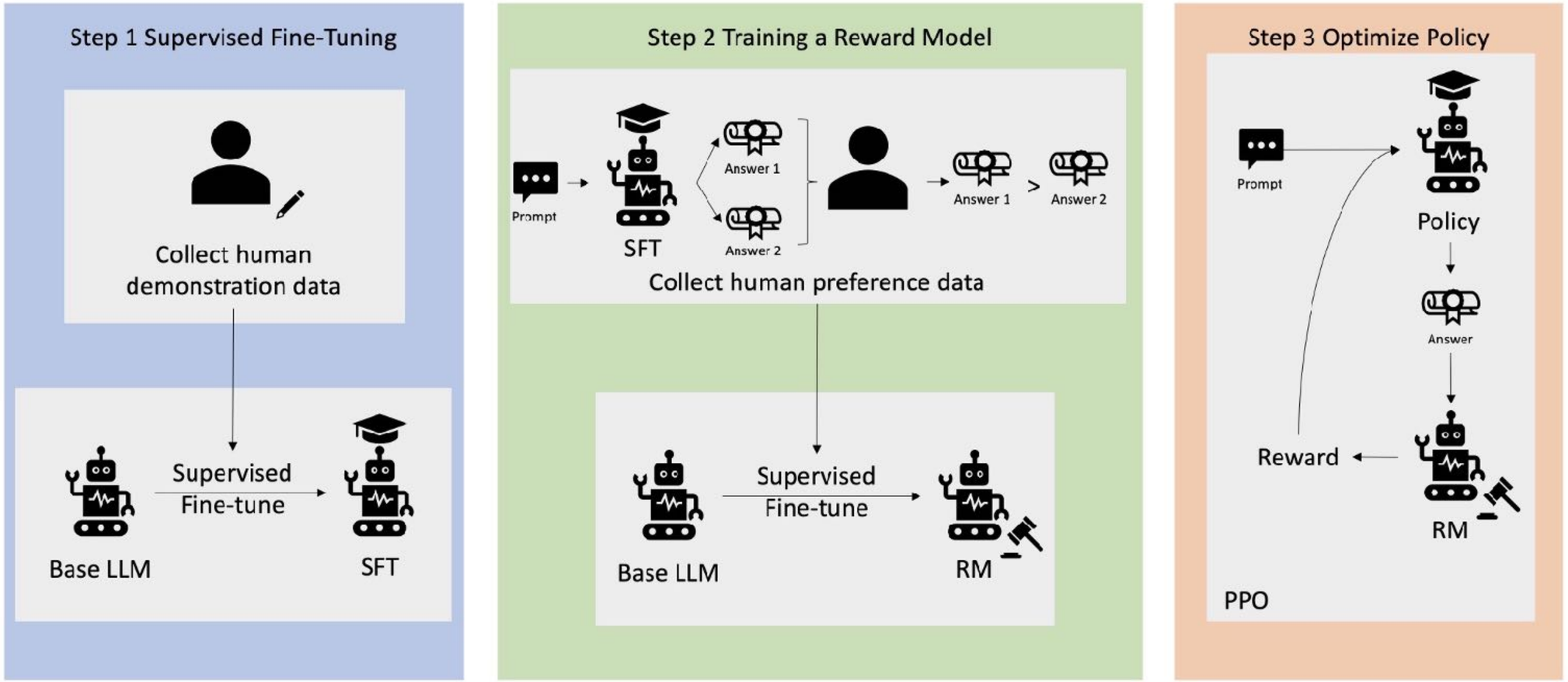
RL hype curve



https://x.com/shaneguML/status/1988008264219697444

Exciting progress of RL for LLM

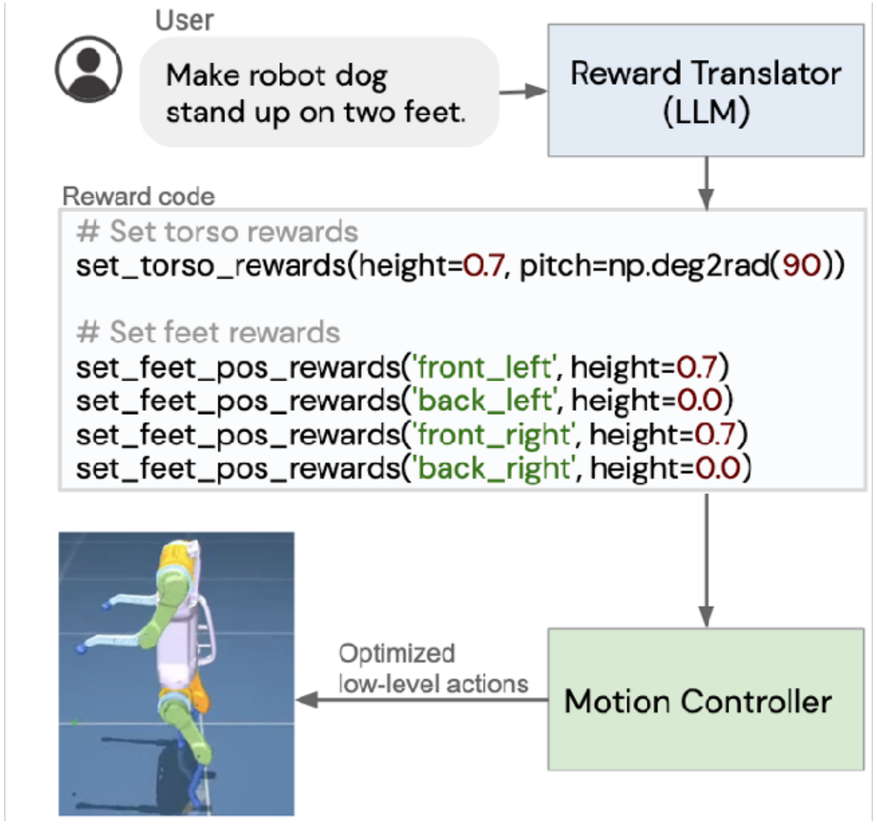
Reinforcement learning from human feedback (RLHF)



https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/

Exciting progress of LLM for RL

Leveraging language models for reward engineering



https://language-to-reward.github.io

Exciting progress of Vision-Language-Action Model

Pi-0.6

https://www.pi.website/blog/pistar06

Gemini Robotics

https://deepmind.google/models/gemini-robotics/

Exciting progress of RL for enabling LLM's reasoning

deepseek

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

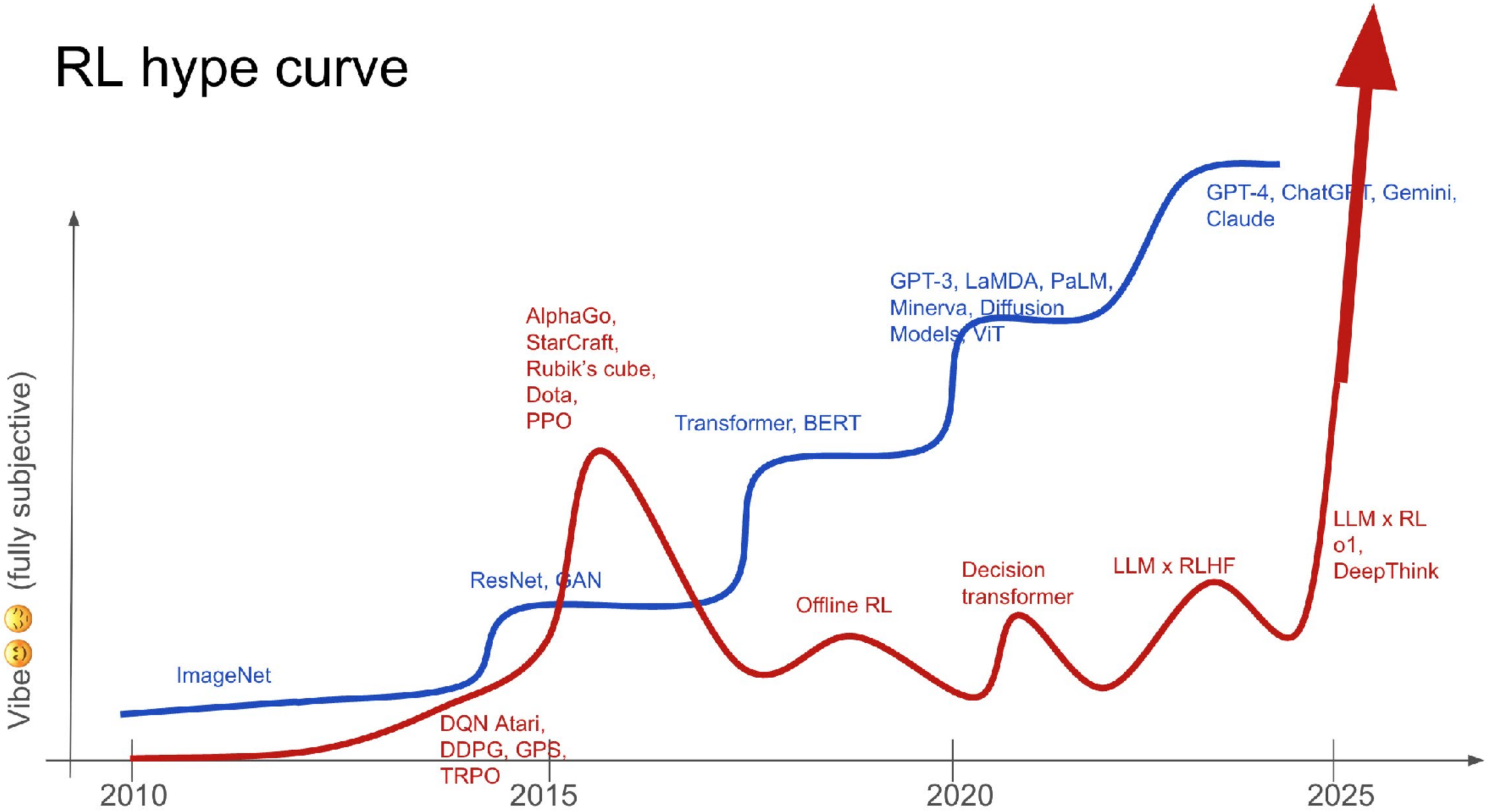
research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

https://arxiv.org/pdf/2501.12948

RL hype curve



https://x.com/shaneguML/status/1988008264219697444

Does RL really incentivize the reasoning capability of LLM?

arXiv put on 22 Jan 2025

deepseek

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

https://arxiv.org/pdf/2501.12948

NeurIPS'25 Best paper run-up

Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

Yang Yue [1*†] Zhiqi Chen [1*] Rui Lu [1] Andrew Zhao [1] Zhaokai Wang [2] Yang Yue [1]
Shiji Song [1] Gao Huang [1✉]

[1] LeapLab, Tsinghua University [2] Shanghai Jiao Tong University
{le-y22, zq-chen23}@mails.tsinghua.edu.cn {gaohuang}@tsinghua.edu.cn

https://limit-of-rlvr.github.io

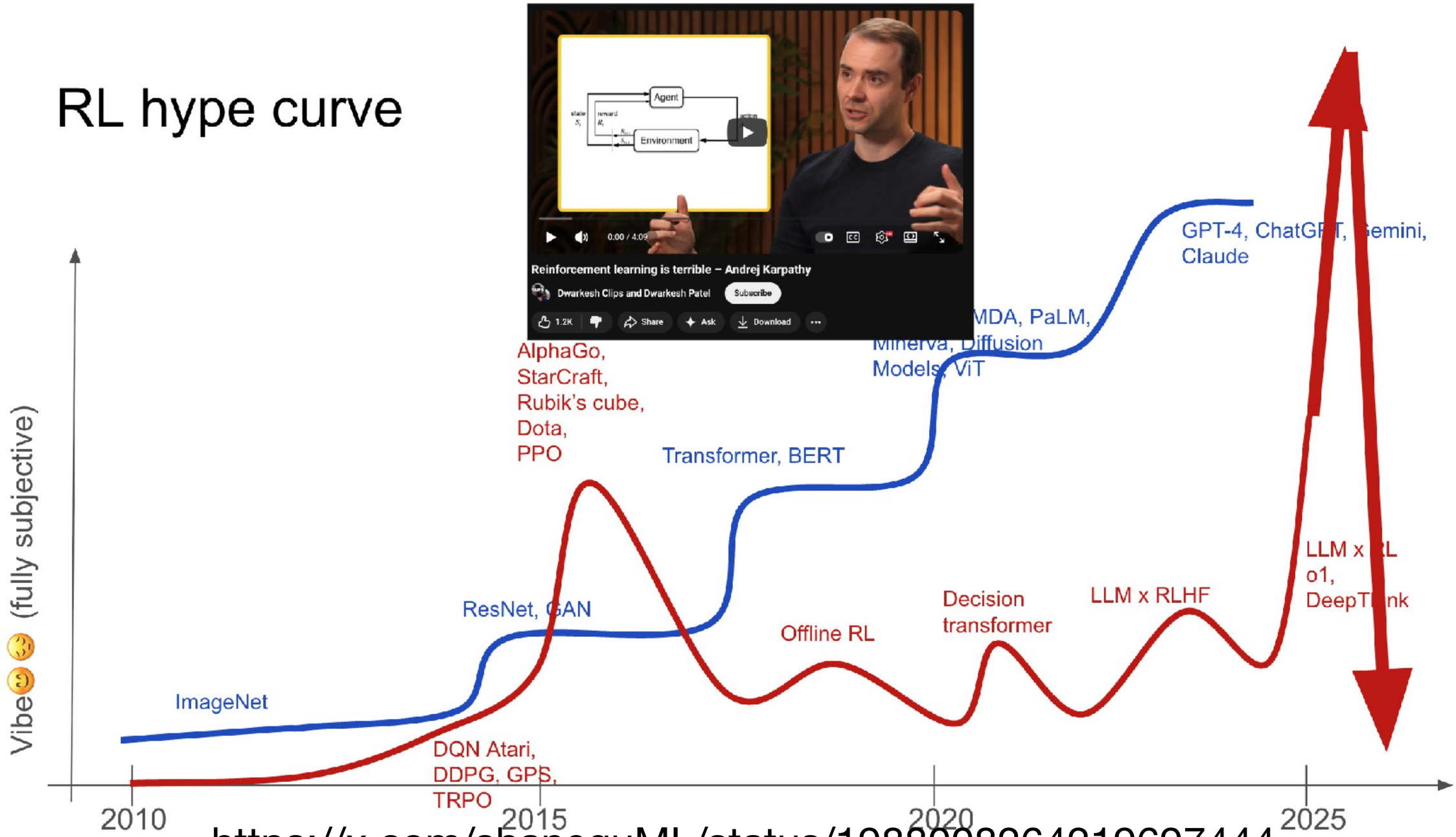
Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has recently demonstrated notable success in enhancing the reasoning performance of large language models (LLMs), particularly in mathematics and programming tasks. It is widely believed that, similar to how traditional RL helps agents to explore and learn new strategies, RLVR enables LLMs to continuously self-improve, thus acquiring novel reasoning abilities that exceed the capacity of the corresponding base models. In this study, we take a critical look at *the current state of RLVR* by systematically

RL improves efficiency at finding correct answers, not the fundamental reasoning capacity of the model.

https://openreview.net/pdf?id=4OsgYD7em5

Deep RL is a roller coaster—only for the strong-hearted



https://x.com/shaneguML/status/1988008264219697444

Richard Sutton – Father of RL thinks LLMs are a dead end

A recent podcast by Turing Awardee Richard Sutton

https://www.youtube.com/watch?v=21EYKqUsPfg

Yann LeCun’s Cake Analogy

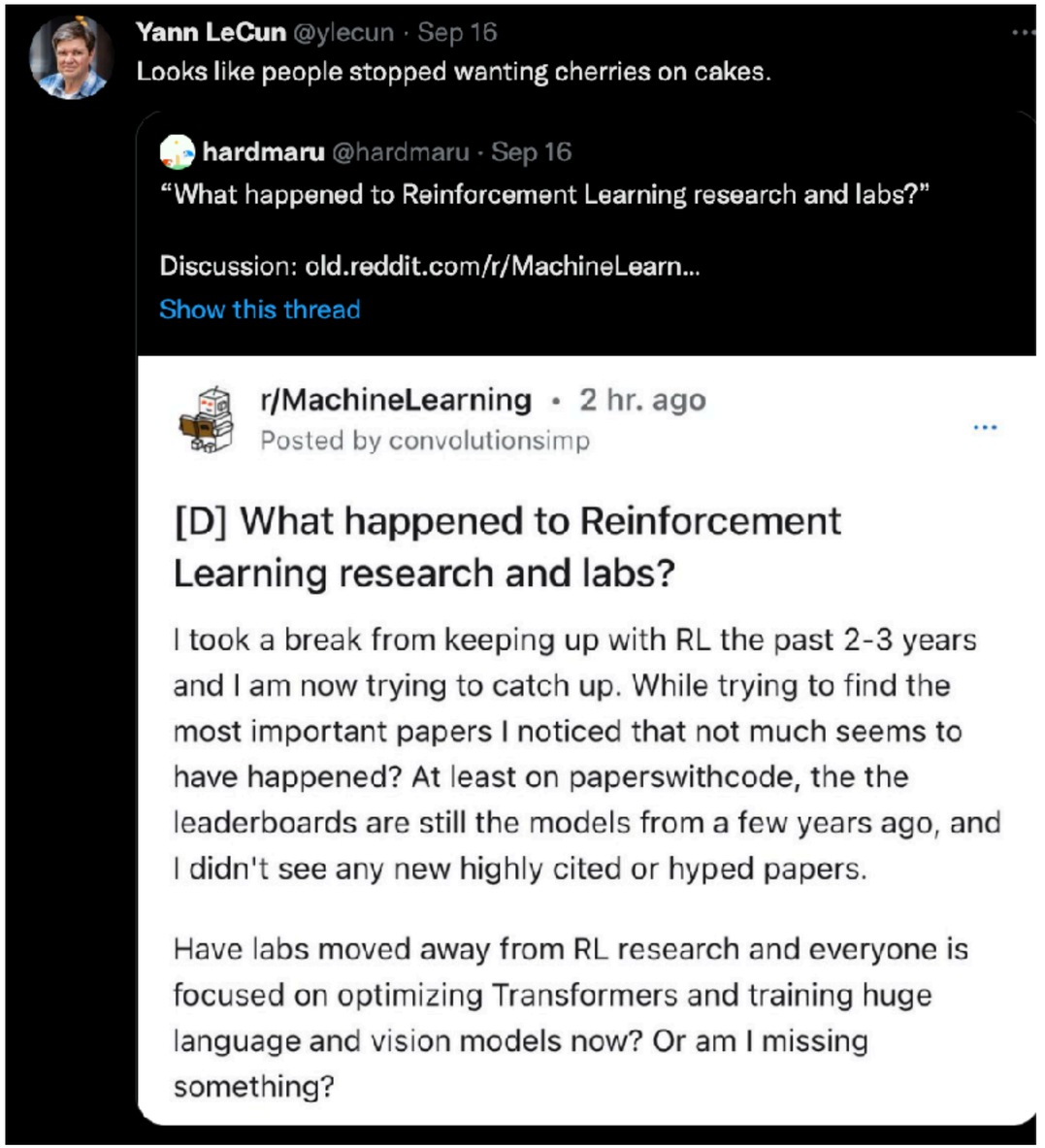
Yann LeCun’s cake analogy

- **“Pure” Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning (cake)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- **(Yes, I know, this picture is slightly offensive to RL folks. But I’ll make it up)**

Interesting discussion on the progress on RL

https://old.reddit.com/r/MachineLearning/comments/xfmqny/d_what_happened_to_reinforcement_learning/



Demystifying RL: one objective of this course

- Know the strengths and weaknesses of the RL methods and how to use it properly in your work

Be cautious about AI snake oil

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Learning objectives of this course

• Students should understand the reinforcement learning foundations such as value-based RL, policy-based RL, actor-critic.

• Students should understand the commonly used RL algorithms such as PPO, SAC, and apply them into some application

• Students should know the strength and weakness of RL and their real-world applications

.

What we will talk about in this course

- Key elements of RL: state, reward, action, Markov decision process, exploration and exploitation, etc.
- RL algorithms: Q-learning, policy gradients, actor-critic.
- Others advanced topics: imitation learning, model-based RL, human-in-the-loop RL.
- Case studies: Projects in DeepMind, OpenAI, and others.
- Hands-on experiences on RL through 4 programming assignments
- 4^{th} programming assignment will be an open-ended tournament, as a mini course project

Course syllabus

Week 1: Overview
Lecture 1: Course introduction
Lecture 2: RL basics and coding with RL
Action: Assignment 0 out
Discussion Session: N/A

Week 2: RL basics
Lecture 1: Markov decision process
Lecture 2: Policy iteration and value iteration
Action: Assignment 1 out
Discussion Session: Policy Evaluation/Iteration Example, Q&A

Week 3: Tabular methods
Lecture 1: Model-free prediction
Lecture 2: Model-free control
Discussion Session: Value Iteration, Q&A

Week 4: Value function approximation and deep Q learning
Lecture 1: value function approximation
Lecture 2: Deep Q Learning
Action: Assignment 1 due, Assignment 2 out
Discussion Session: Q&A

Week 5: Policy-based RL: basics
Lecture 1: Policy Optimization 1
Lecture 2: Policy Optimization 2
Discussion Session: Assignment 1 Review,

Week 6: Policy-based RL: state-of-the-art
Lecture 1: Policy Optimization 3
Lecture 2: Policy Optimization 4
Action: Assignment 2 due, Assignment 3 out
Discussion Session: Policy Gradient, Q&A

Week 7: Model-based RL
Lecture 1: Model-based RL
Lecture 2: Connection to optimal control
Discussion: N/A

Week 8: Advanced topics
Lecture 1: Imitation learning
Lecture 2: Distributed computing and RL system design
Action: Assignment 3 due, Assignment 4 out
Discussion: Assignment 2 Review, Q&A

Week 9: Advanced topics
Lecture 1: Offline RL and real-world RL
Lecture 2: Human-in-the-loop RL
Discussion: N/A

Week 10: Advanced topics
Lecture 1: RL for LLM and LLM for RL
Lecture 2: Course summary
Action: Assignment 4 due
Discussion: Review

Course Logistics

Instructor: Bolei Zhou <bolei@cs.ucla.edu>

TAs:
- Haoyuan Cai <haoyuan@cs.ucla.edu >
- Matthew Leng <matthewleng@cs.ucla.edu >

Lecture Time: Tuesday/Thursday 10:00 am - 11:40 am

TA's Discussion Session and Office Hours (you can go to either one as long as space allows):
- DIS 1A, Friday 12 pm - 1:50 pm, DODD 170, Matthew Leng
- DIS 1B, Friday 2 pm - 3:50 pm, BOLTER 2444, Haoyuan Cai

Prof. Zhou's Office Hours: Eng VI, 295D, Thursday 5:00 pm - 5:45 pm

Piazza Forum Link: https://piazza.com/ucla/winter2026/cs260r
- Expected reply time: 36 hours. Go to TA office hour/discussion if urgent or possible

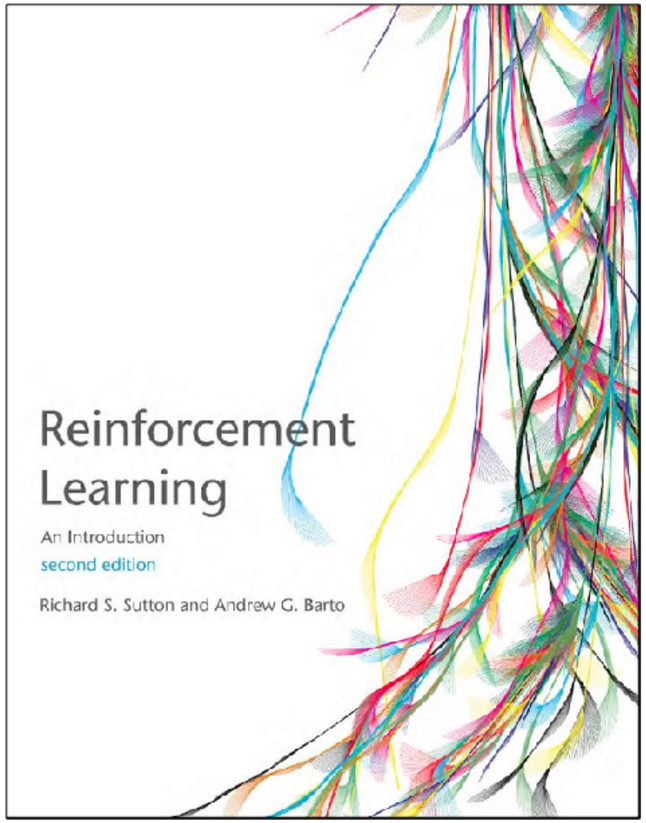
Slides and assignments will be made available at this main google Doc

https://docs.google.com/document/d/1KdCCszXqyBQWa7r3Rm0EN6Ci5PLSo0s3Ul75siO8k_k/edit?usp=sharing

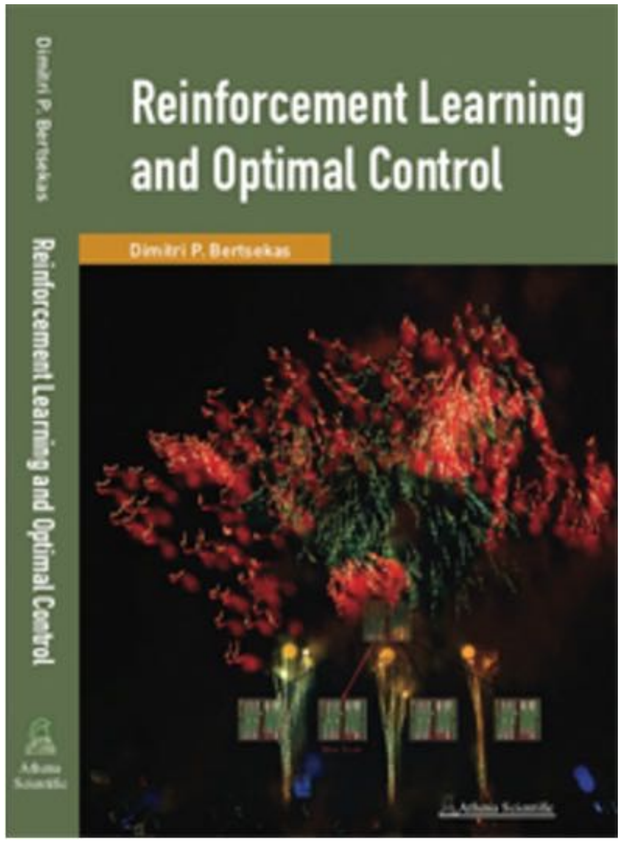
Auditing and sit-in students, you don’t need to email me to be added in Bruinlearn, please go to this google doc link for everything!

Optional Textbooks

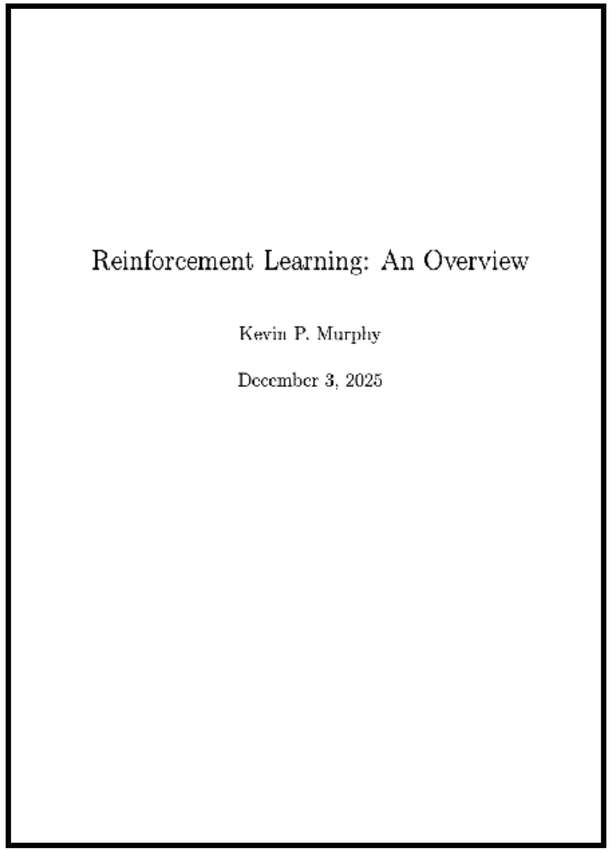
Sutton and Barton
http://incompleteideas.net/book/the-book-2nd.html



Dimitri P. Bertsekas
https://web.mit.edu/dimitrib/www/RLbook.html



Kevin P. Murphy
https://arxiv.org/pdf/2412.05265


Reinforcement Learning: An Overview

Kevin P. Murphy

December 3, 2025

Pre-requisites for Enrollment

- This is CS-2XX course, which means it is a graduate level course
- All enrolled students must have taken the Linear Algebra course and Probability course, and one machine learning relevant course (data mining, pattern recognition, deep learning, etc).
- Coding experience with python and PyTorch

.

Grading

- Attendance: 10%
- Assignments+miniproject: 50%
- Final exam: 40%

4 Assignments + 1 Mini-project

All programming based

Assignment 1: MDP basics, Policy and value iterations

Assignment 2: on-policy, off-policy learning, Q learning

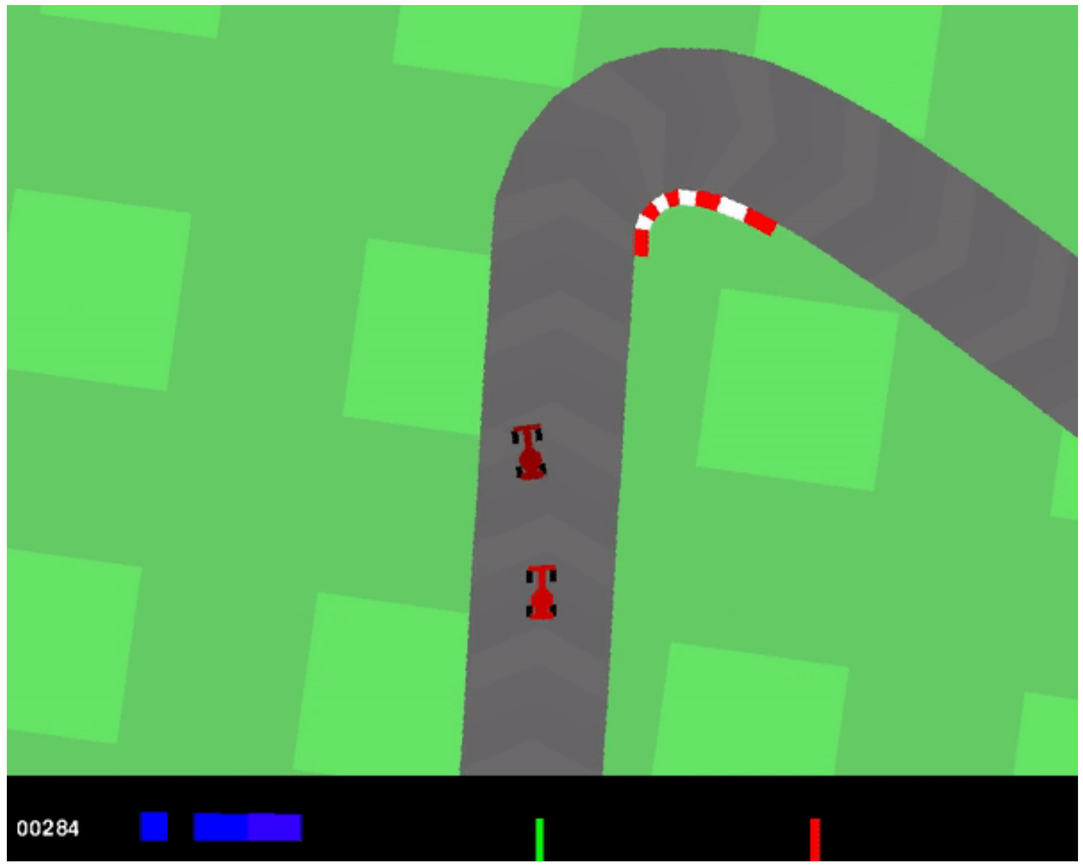
Assignment 3: policy gradient, SOTA methods,

Assignment 4: imitation learning and RLHF

Mini-project: Open-ended environment and tournament

Open-ended environments

Competitive Env: https://github.com/ucla-rlcourse/competitive-rl



Open-ended Environments

MetaDrive

https://metadriverse.github.io/metadrive/

Last year’s miniproject: MetaDrive-Racing Competition

- You can be creative and try best
- Leaderboard: top xx% get some bonus score

A short break, next is RL overview

Supervised Learning

Learn to classify object

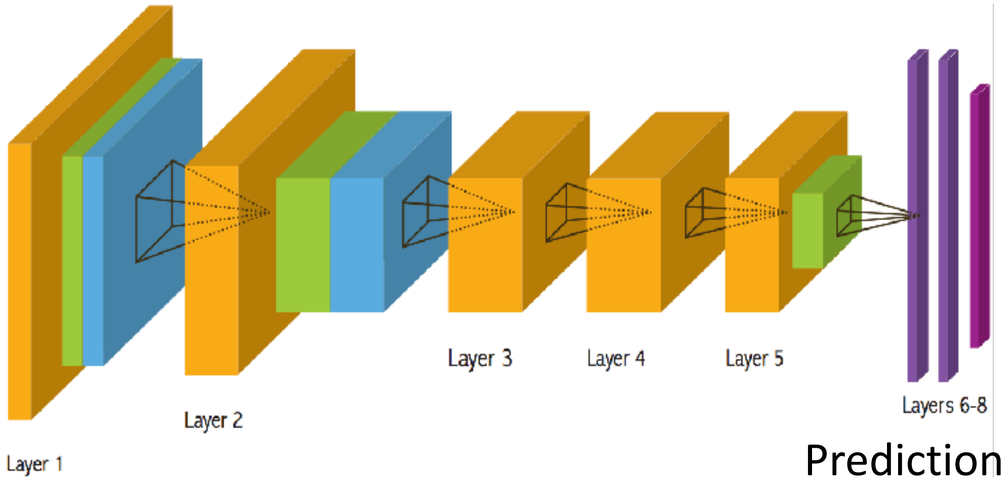
- Annotated images, data follows i.i.d distribution.
- Learners are told what the labels are.

Training annotated data

Car

Airplane

Chair



Prediction: Airplane

Prediction is wrong
Correct label: Car

Back-propagation

Reinforcement Learning

Learn to play Breakout

- Data are not i.i.d. Instead, a correlated time series data
- No instant feedback or label for correct action

Action: Move LEFT or Right

Atari Game: Breakout

Training data

Step1

Step2

Step3

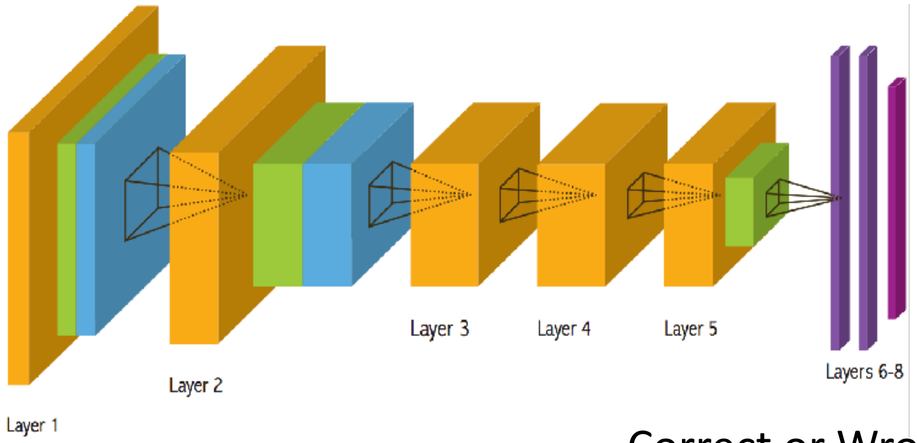
Step4

Step5

. . .

Human playing?

Step3



Action:
Move LEFT

Correct or Wrong???
Don’t know for now, until game is over
Delayed reward

Backpropagation?

Difference between Reinforcement Learning and Supervised Learning

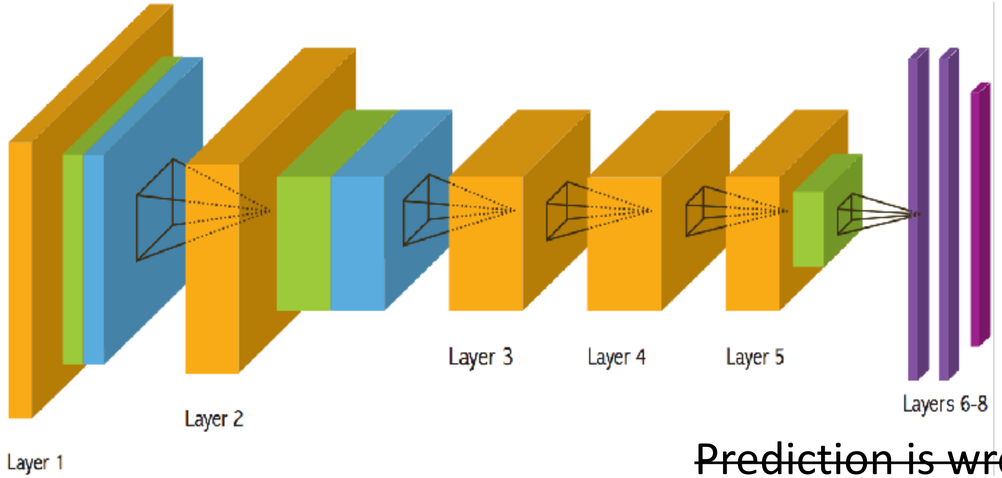
- Sequential data as input (not i.i.d)
- The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.
- Trial-and-error exploration (balance between exploration and exploitation)
- There is no supervisor, only a reward signal, which is also delayed

Properties of reinforcement learning

- Trial-and-error exploration
- Delayed reward
- Time matters (sequential data, non i.i.d data)
- Agent’s actions affect the subsequent data it receives (agent’s action changes the environment)

Reinforcement Learning of Image Classification

- Annotated images, data follows i.i.d distribution.
- Learner is given a positive reward if the prediction is correct



Prediction:
Airplane

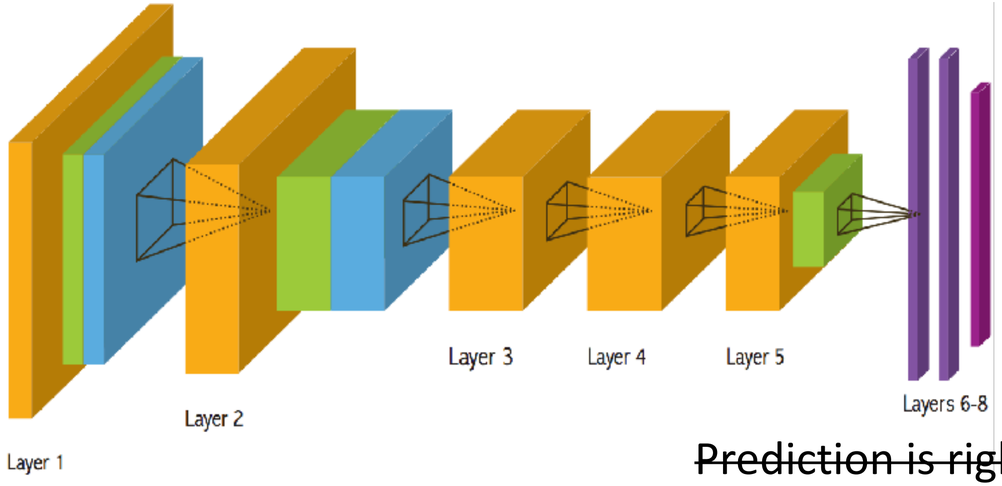
~~Prediction is wrong~~
~~Correct label: Car~~

Reward is 0

Back-propagation

Reinforcement Learning of Image Classification

- Annotated images, data follows i.i.d distribution.
- Learner is given a positive reward if the prediction is correct



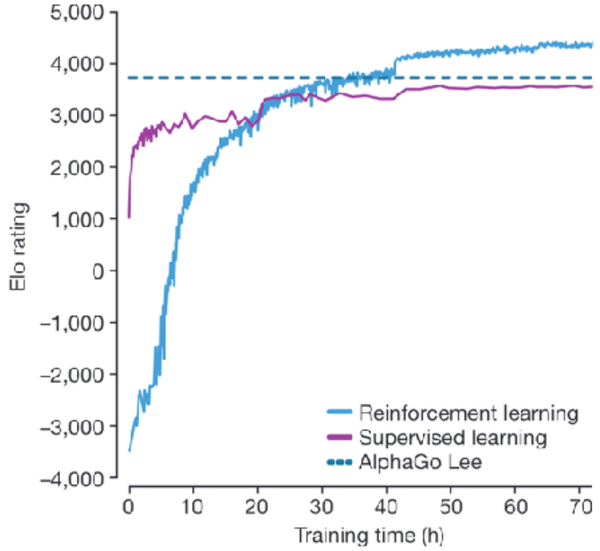
Prediction:
Car

~~Prediction is right~~ Reward is 1

Back-propagation

Why we care about RL

- Learn to control something via trial-and-error in **model-free** setting
- You don’t need to model the whole dynamics explicitly
- The model may achieve super-human performance
  - Upper bound for supervised learning is human-performance.
  - Upper bound for reinforcement learning?



AlphaGo vs. AlphaGo

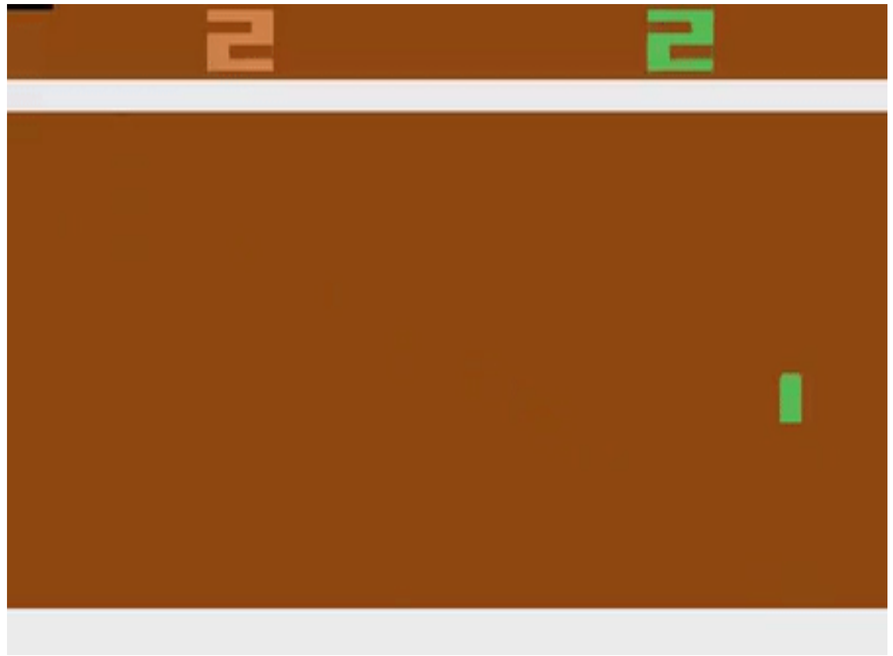
Silver and Schrittwieser, et al. Mastering the game of Go without human knowledge,

Examples of reinforcement learning applications

- A chess player makes a move: the choice is informed both by planning-anticipating possible replies and counterreplies.
- A baby deer struggles to stand, 30 min later it can run 36 kilometers per hour.
- Portfolio management.
- Playing Atari game

Another RL example: Pong

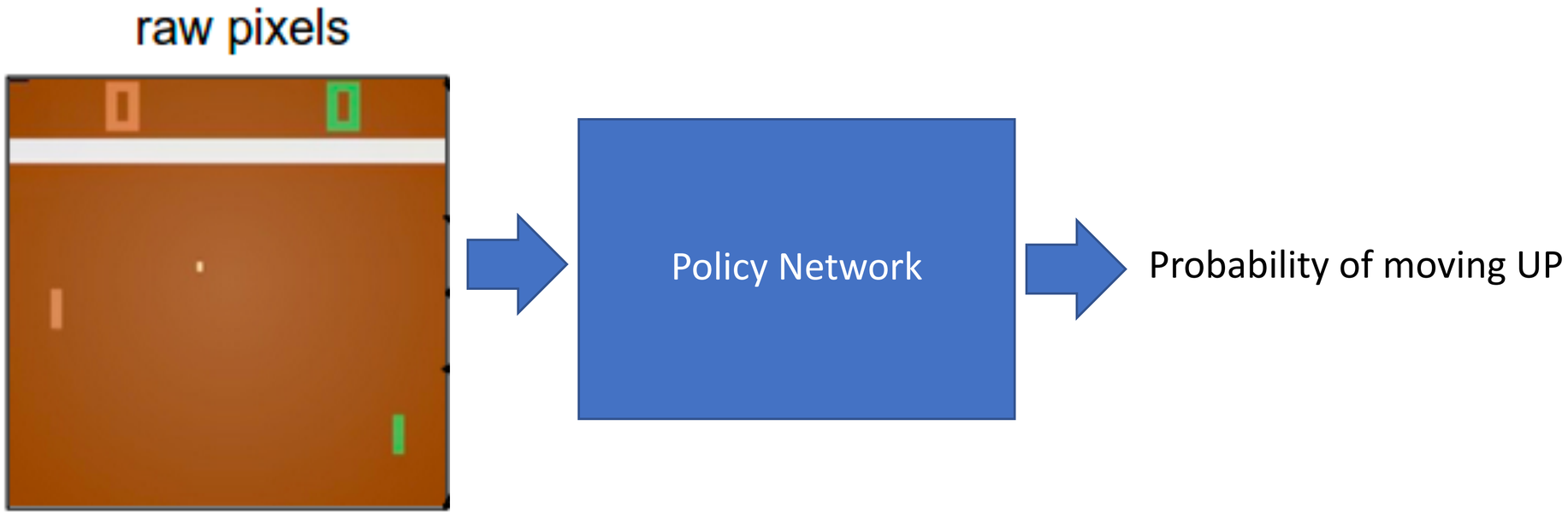
Action: move UP or DOWN



From Andrej Karpathy blog: http://karpathy.github.io/2016/05/31/rl/

Another RL example: Pong

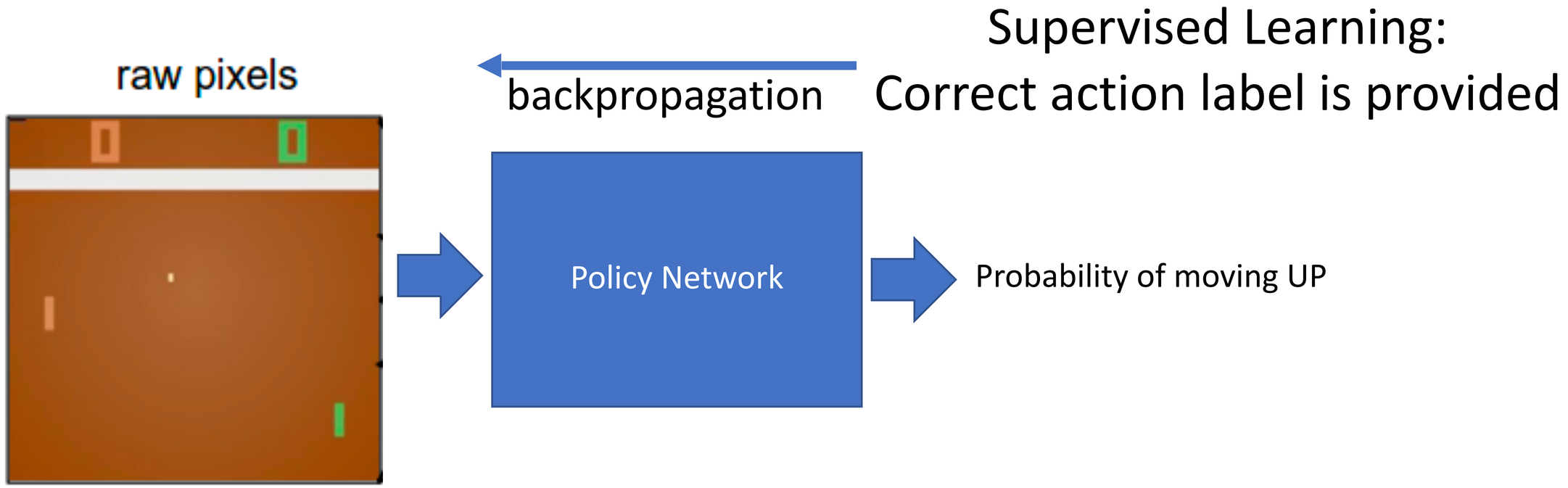
Action: move UP or DOWN



From Andrej Karpathy blog: http://karpathy.github.io/2016/05/31/rl/

Another RL example: Pong

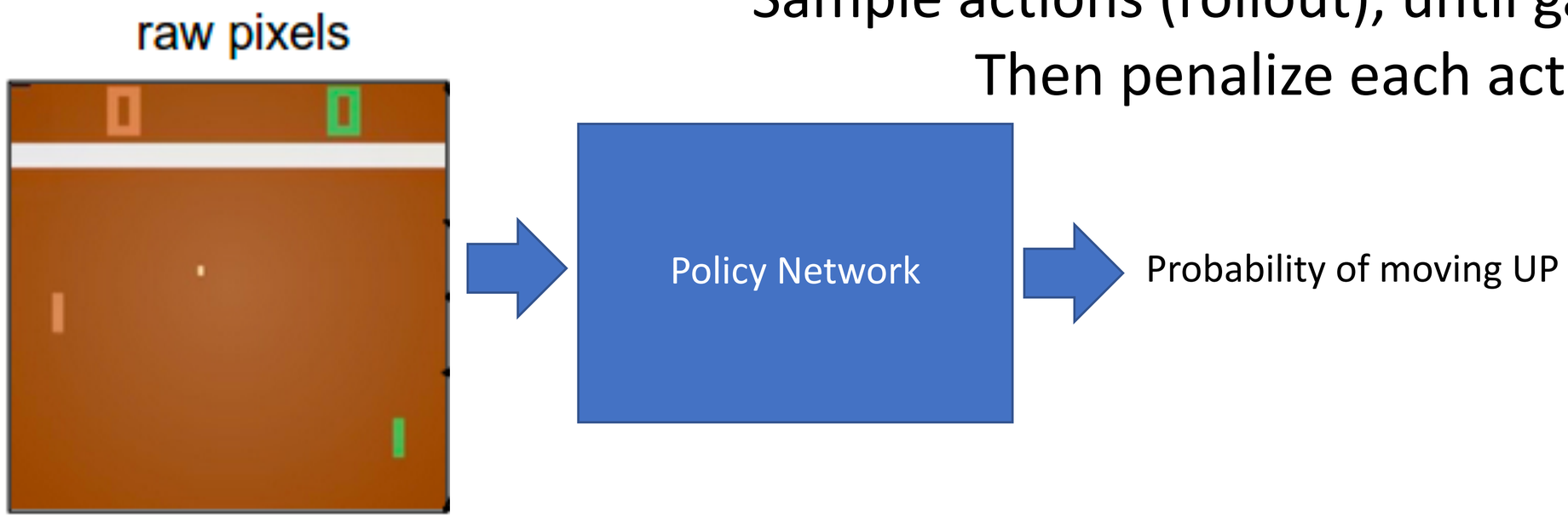
- Action: move UP or DOWN



Another RL example: Pong

- Action: move UP or DOWN

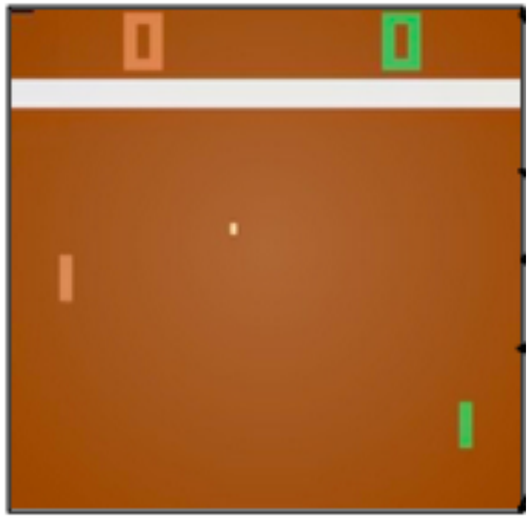
Reinforcement Learning:
Sample actions (rollout), until game is over,
Then penalize each action



Another RL example: Pong

- Action: move UP or DOWN

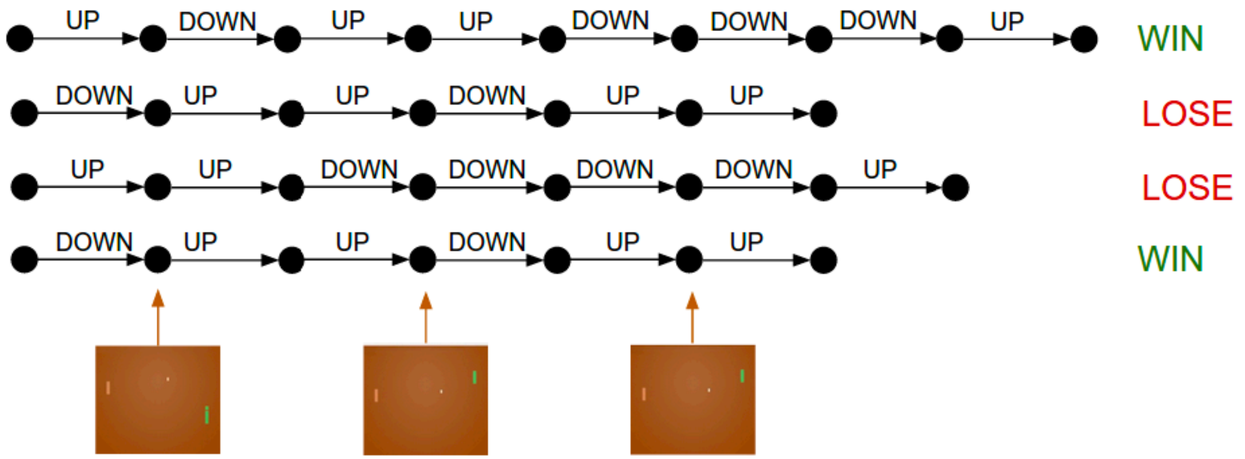
raw pixels



Reinforcement Learning:
Sample actions (rollout), until game is over,
Then penalize each action

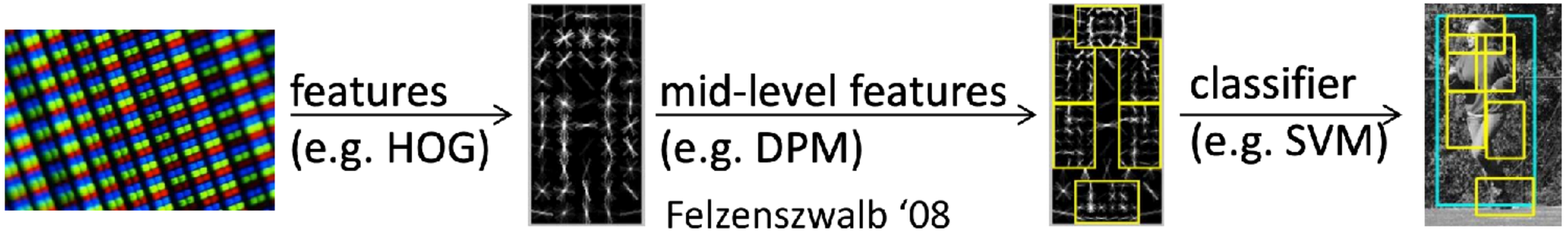
Possible rollout sequence:

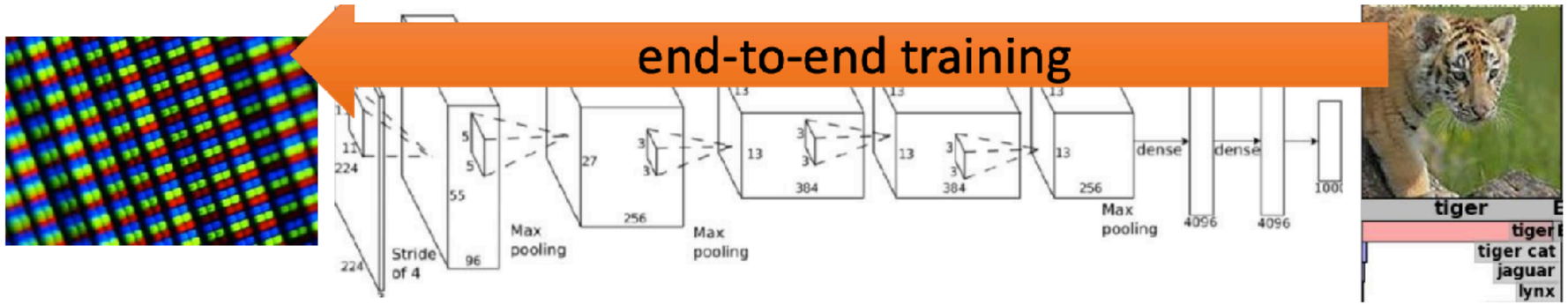
Eventual Reward:



Why deep reinforcement learning?

- Analogy to traditional CV and deep CV

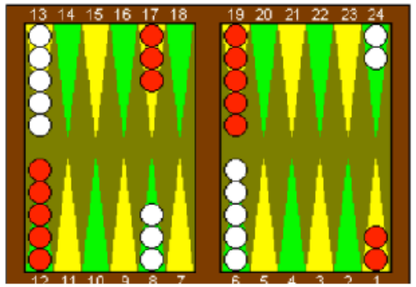




Why deep reinforcement learning?

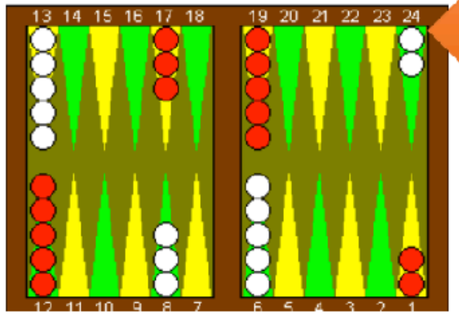
- Standard RL and deep RL

TD-Gammon, 1995

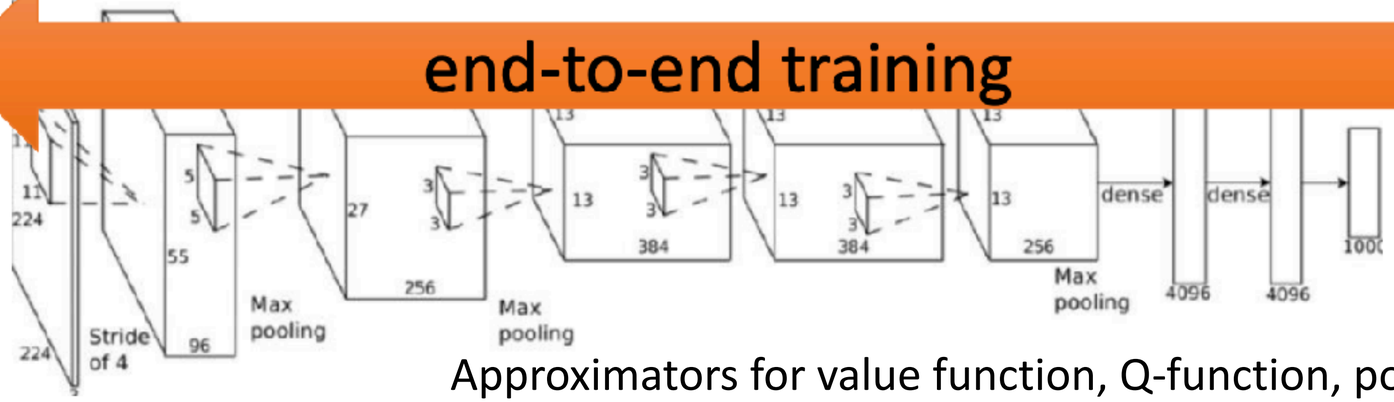


game of backgammon

features → ? → more features → ? → linear policy or value func. → action



end-to-end training



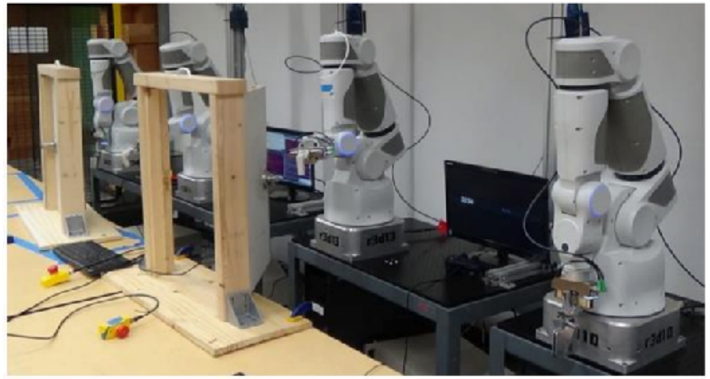
action

Approximators for value function, Q-function, policy networks

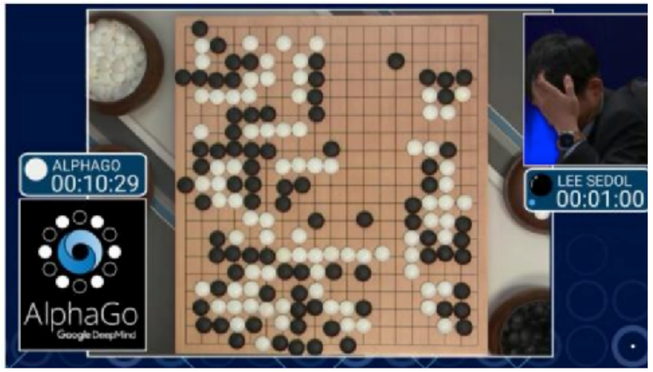
Why RL works now?

- One of the most exciting areas in machine learning

Game playing



Robotics

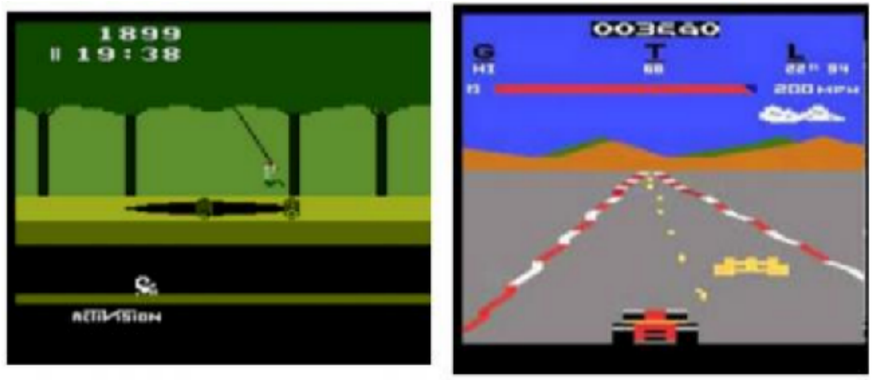


Beating best human player

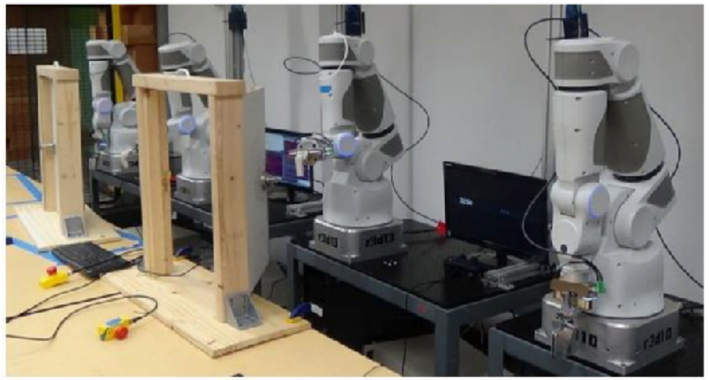
[Playing Atari with Deep Reinforcement Learning]
[Mastering the game of Go without Human Knowledge]

Why RL works now?

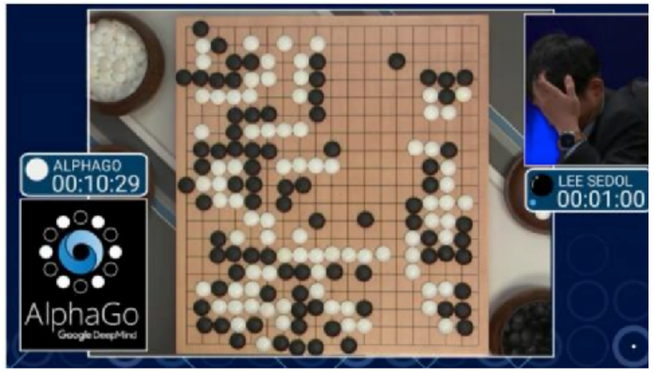
- Computation power: many GPUs to do trial-and-error rollout
- End-to-end training, features and policy are jointly optimized toward the end goal
  - Acquire the high degree of proficiency in domains governed by simple, known rules and reward function



Game playing



Robotics



Beating best human player

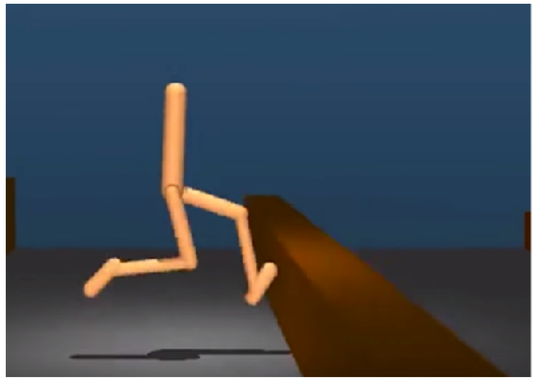
What are the applications of RL?

- Robot learning (CoRL'25 accepted papers: https://openreview.net/group?id=robot-learning.org/CoRL/2025/Conference#tab-your-consoles)
- Computer graphics and physical simulation
- Large-scale machine learning systems
- Human-in-the-loop systems
- Model-free control

What are the applications of RL?

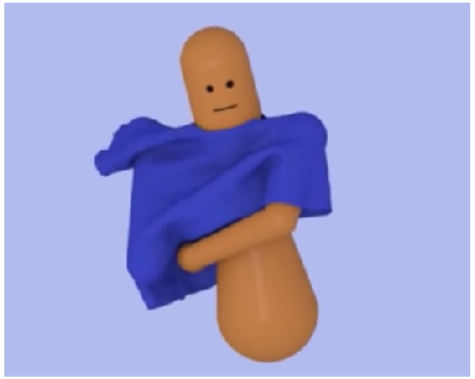
Some interesting examples:

Learning to walk



https://www.youtube.com/watch?v=gn4nRCC9TwQ

Learning to dress

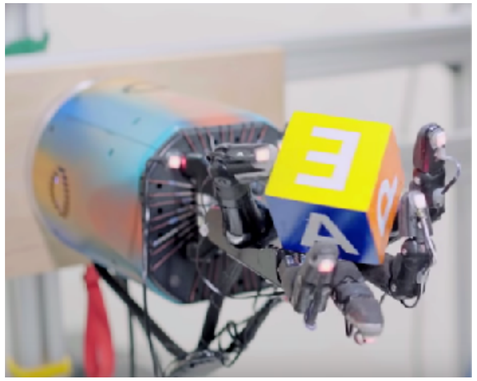


https://www.youtube.com/watch?v=ixmE5nt2o88

Learning to grasp

https://ai.googleblog.com/2016/03/deep-learning-for-robots-learning-from.html

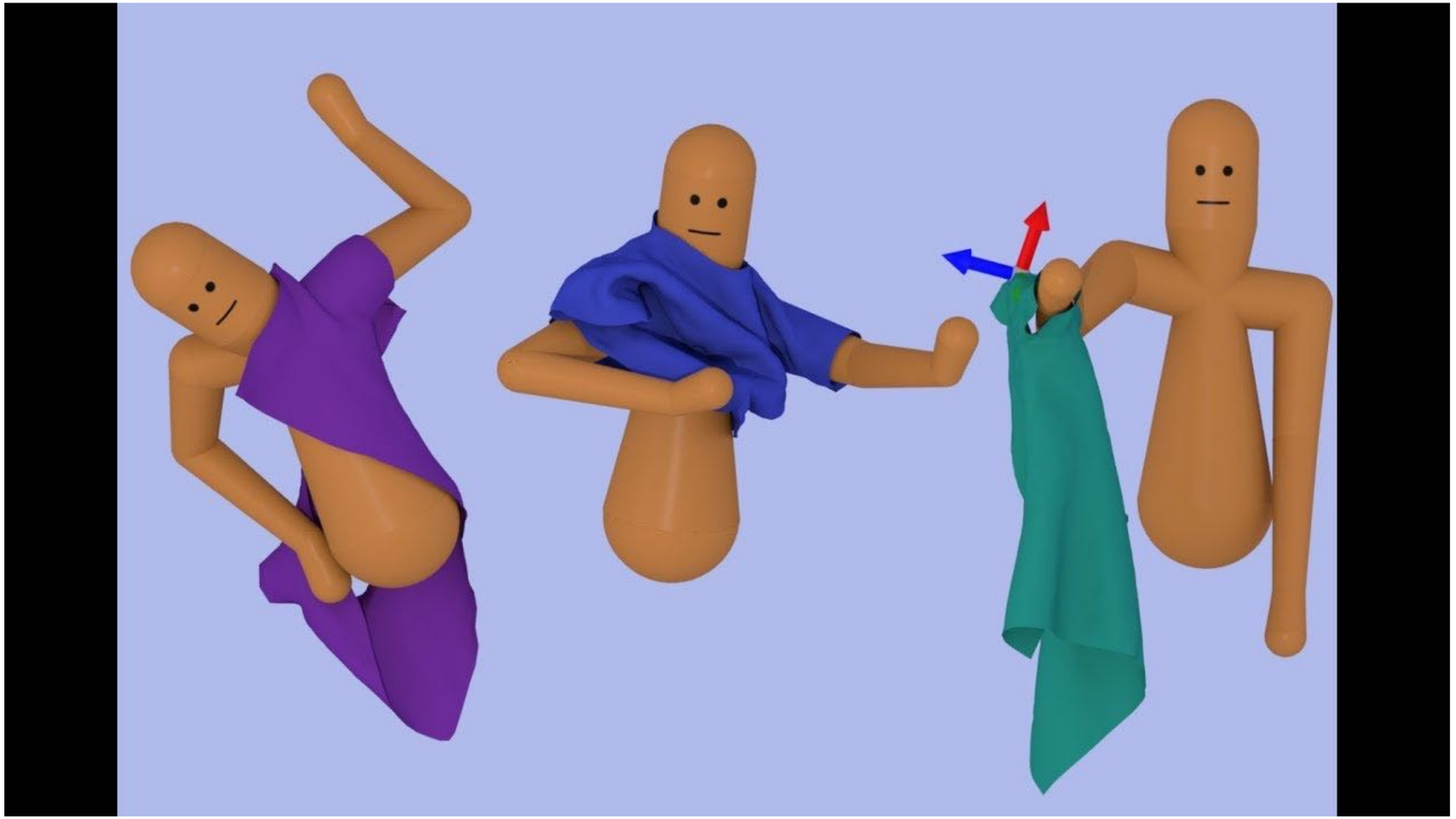
Learning to manipulate



https://www.youtube.com/watch?v=jwSbzNHGflM

Learning to walk

Learning to dress



Learning to dress: synthesizing human dressing motion via deep reinforcement learning. https://ckllab.stanford.edu/learning-dress-synthesizing-human-dressing-motion-deep-reinforcement-learning

Learning to manipulate amorphous materials

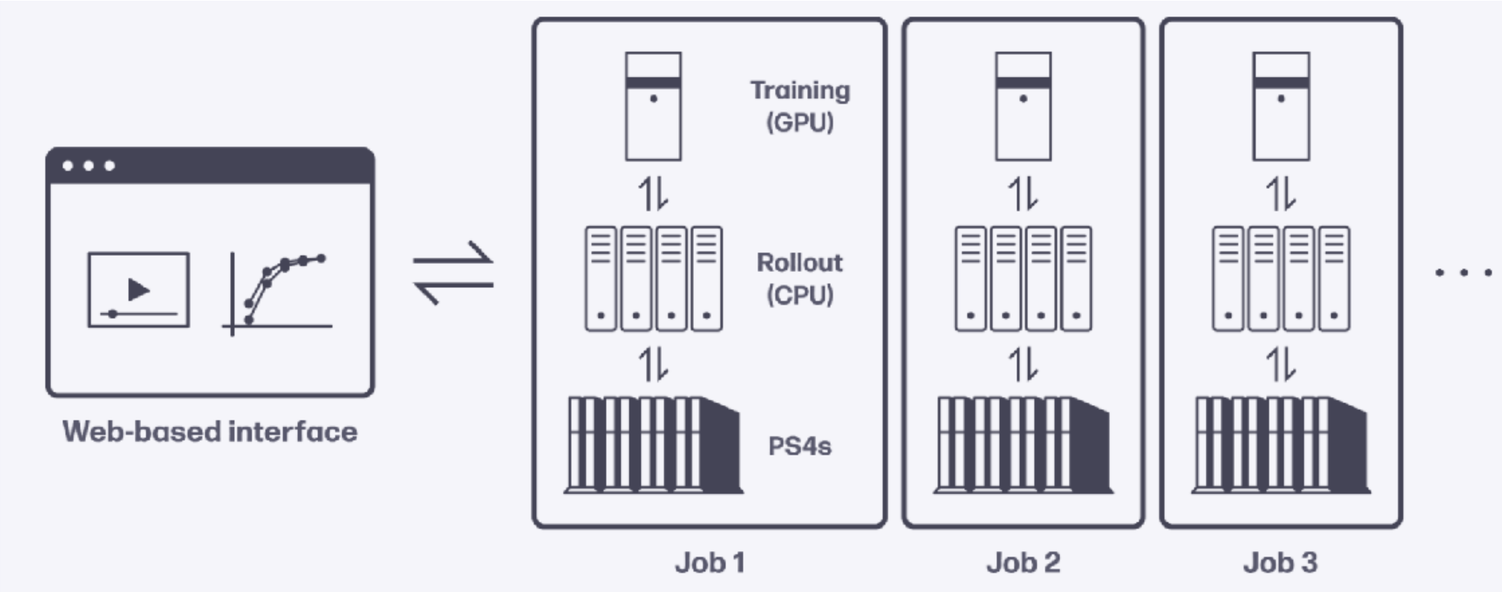
Yunbo Zhang, Wenhao Yu, C. Karen Liu, Charles C. Kemp, Greg Turk. SIGGRAPH ASIA (2020)
https://www.cc.gatech.edu/~yzhang3027/publication/learning-to-manipulate-amorphous-materials/

Learning to Race

Super-Human Performance in Gran Turismo Sport Using Deep Reinforcement Learning. https://arxiv.org/pdf/2008.07971.pdf

Learning to Race (GT Sophy, Feb. 2022)

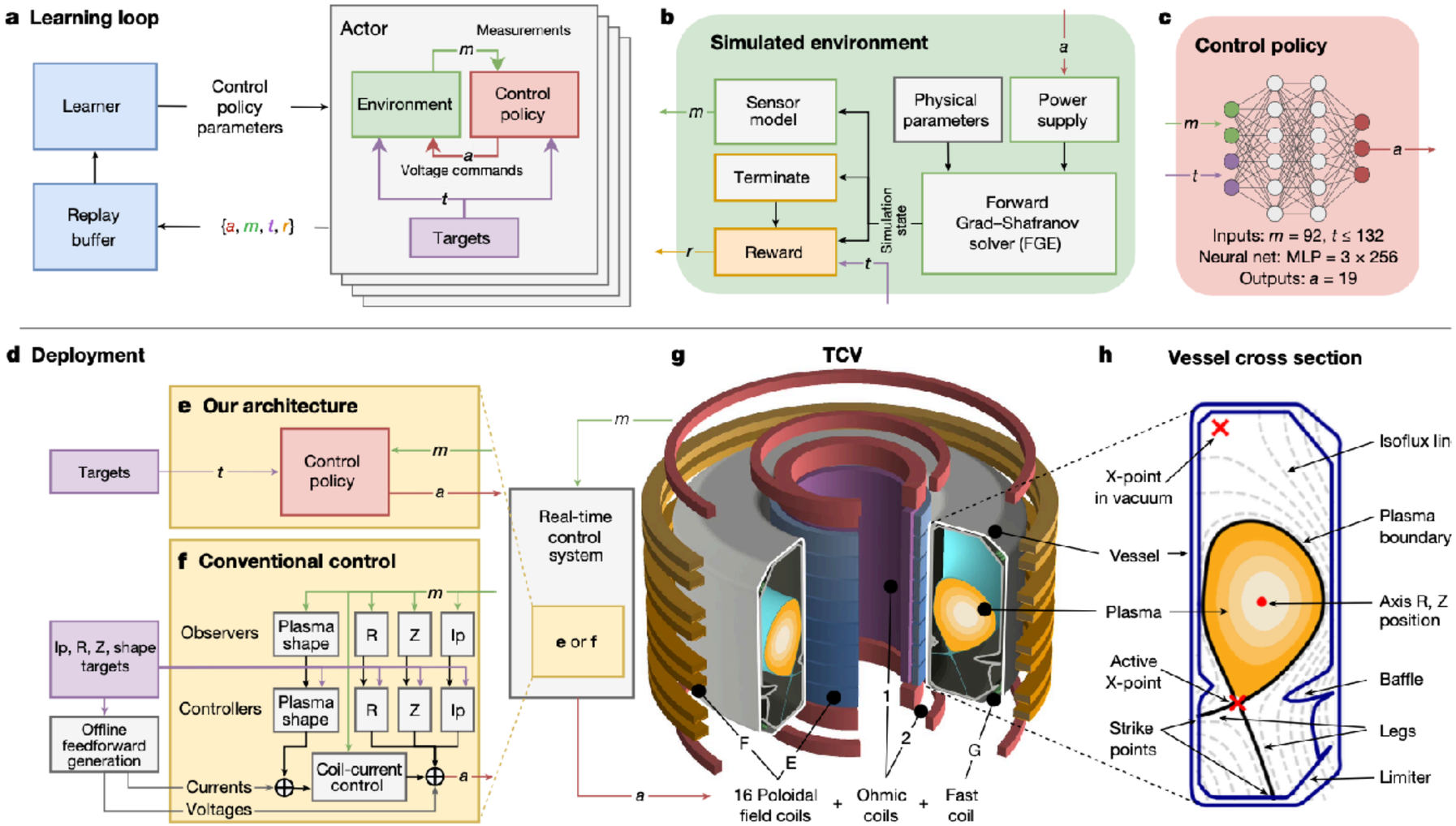
Distributed, Asynchronous Rollouts and Training (DART)



https://www.nature.com/articles/s41586-021-04357-7

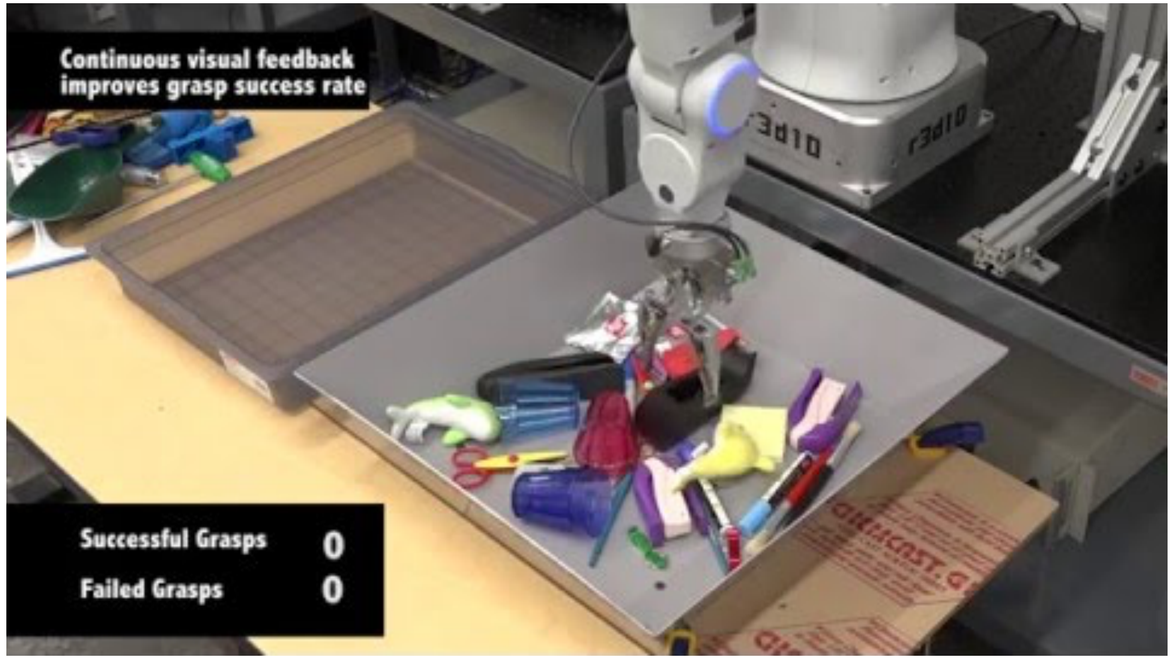
https://www.gran-turismo.com/us/gran-turismo-sophy/technology/

Learning to control tokamak plasmas in nuclear fusion (Feb. 2022)



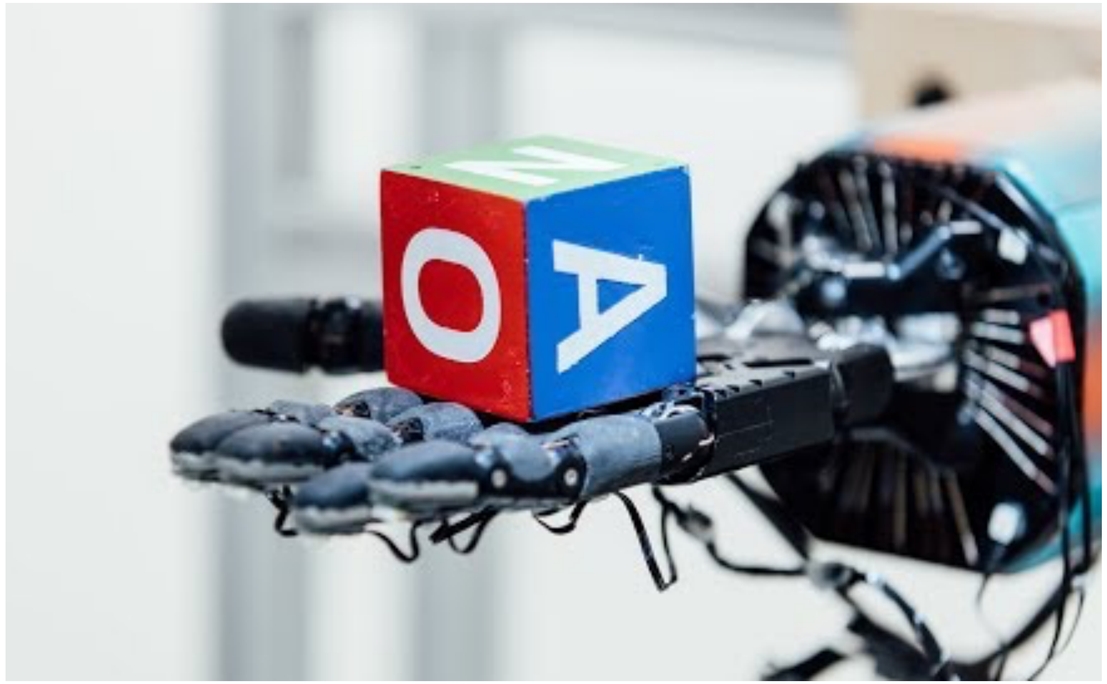
https://www.wired.com/story/deepmind-ai-nuclear-fusion/ https://www.nature.com/articles/s41586-021-04301-9

Learning to grasp

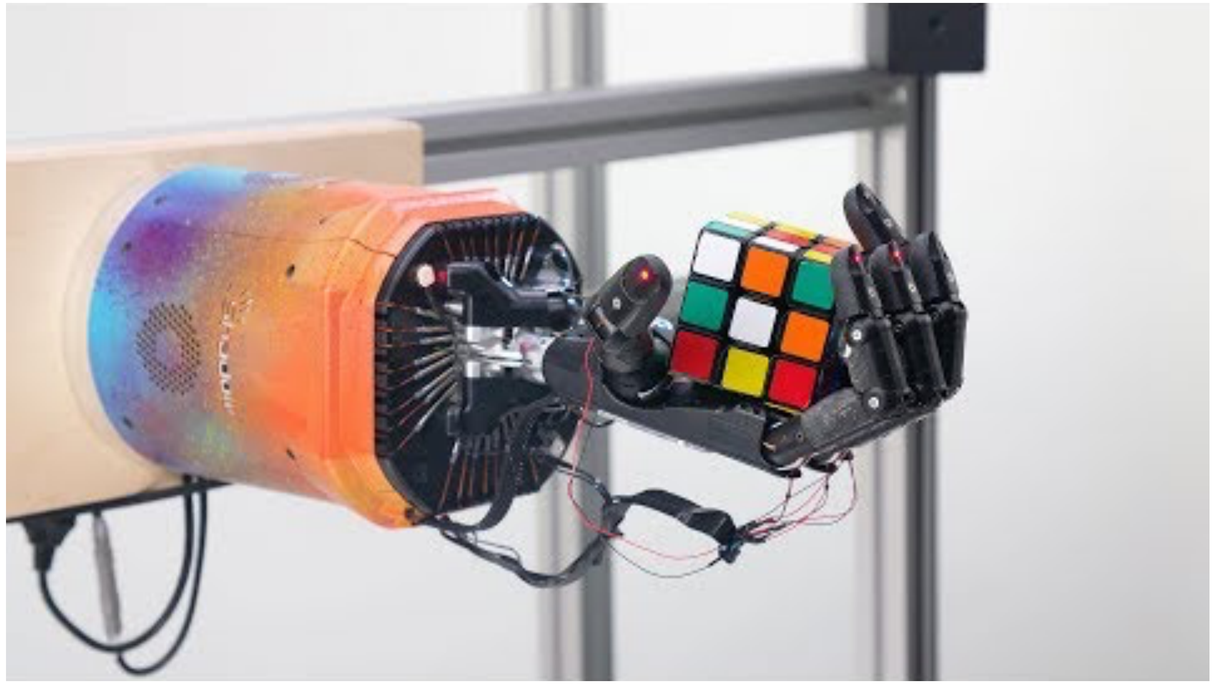


Learning to manipulate cubic using fingers

2018



2019



OpenAI

Sim2Real transfer learning: How to apply the agent trained in simulator to real world?

Sim2Real transfer learning: How to apply the agent trained in simulator to real world?

Build more realistic and efficient simulator

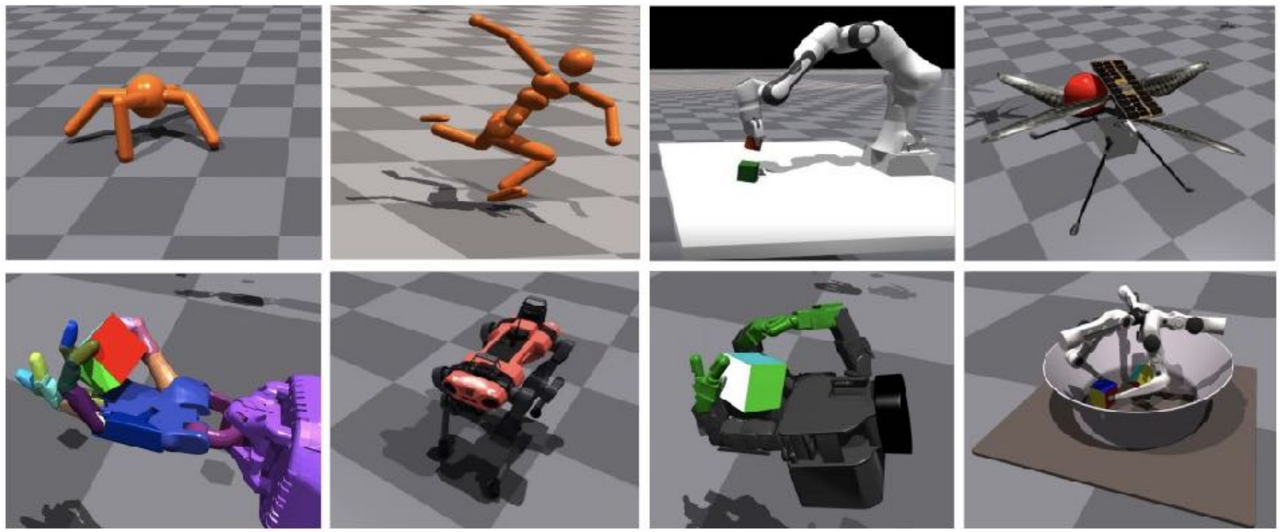


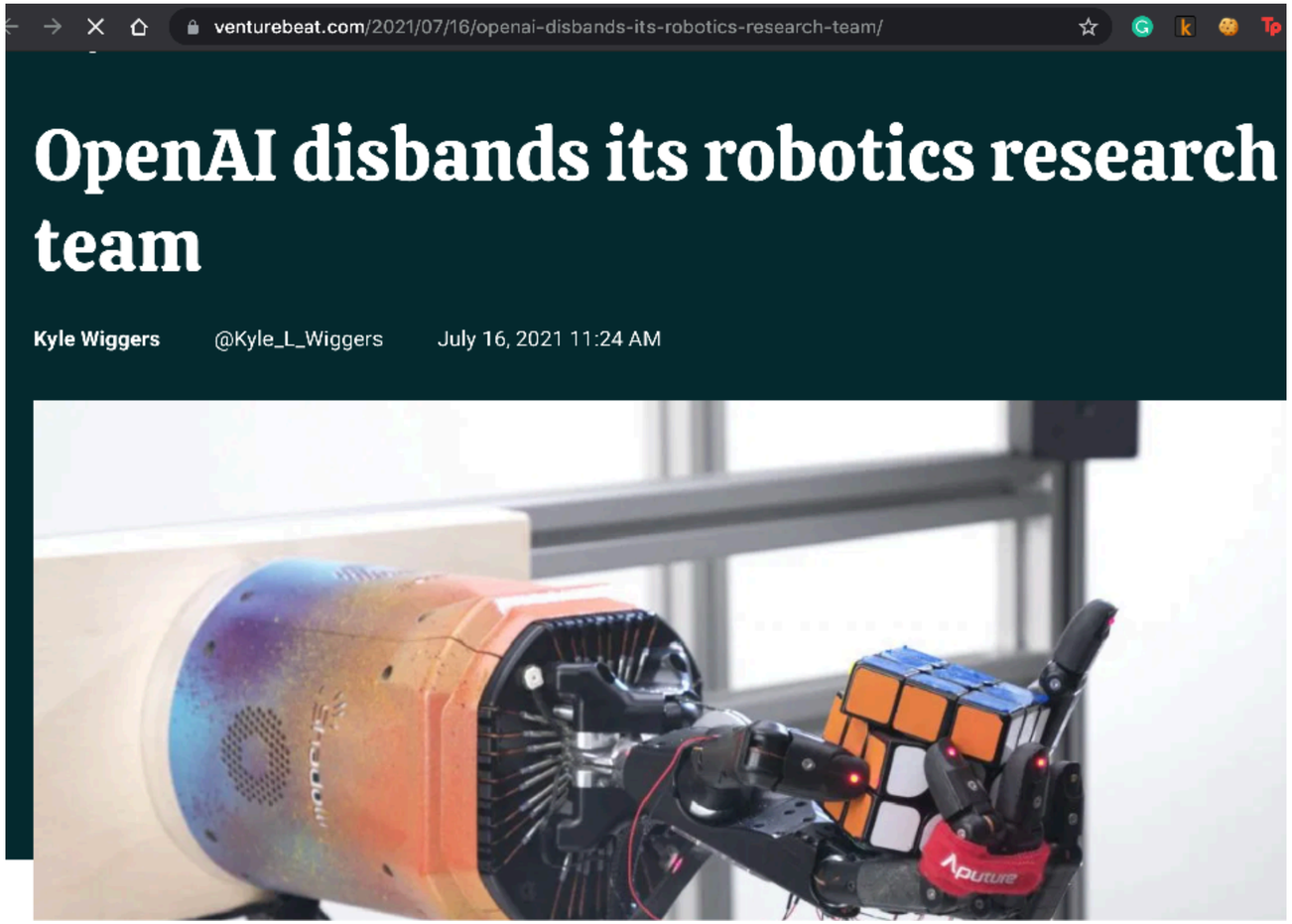
Figure 1: Isaac Gym allows high performance training on a variety of robotics environments. We benchmark on 8 different environments that offer a wide range of complexity and show the strengths of the simulator in blazing fast policy training on a single GPU. *Top*: Ant, Humanoid, Franka-cube-stack, Ingenuity. *Bottom*: Shadow Hand, ANYmal, Allegro, TriFinger.

Nvidia Issac Gym: https://sites.google.com/view/isaacgym-nvidia

Quick adaptation algorithms

https://ashish-kmr.github.io/rma-legged-robots/

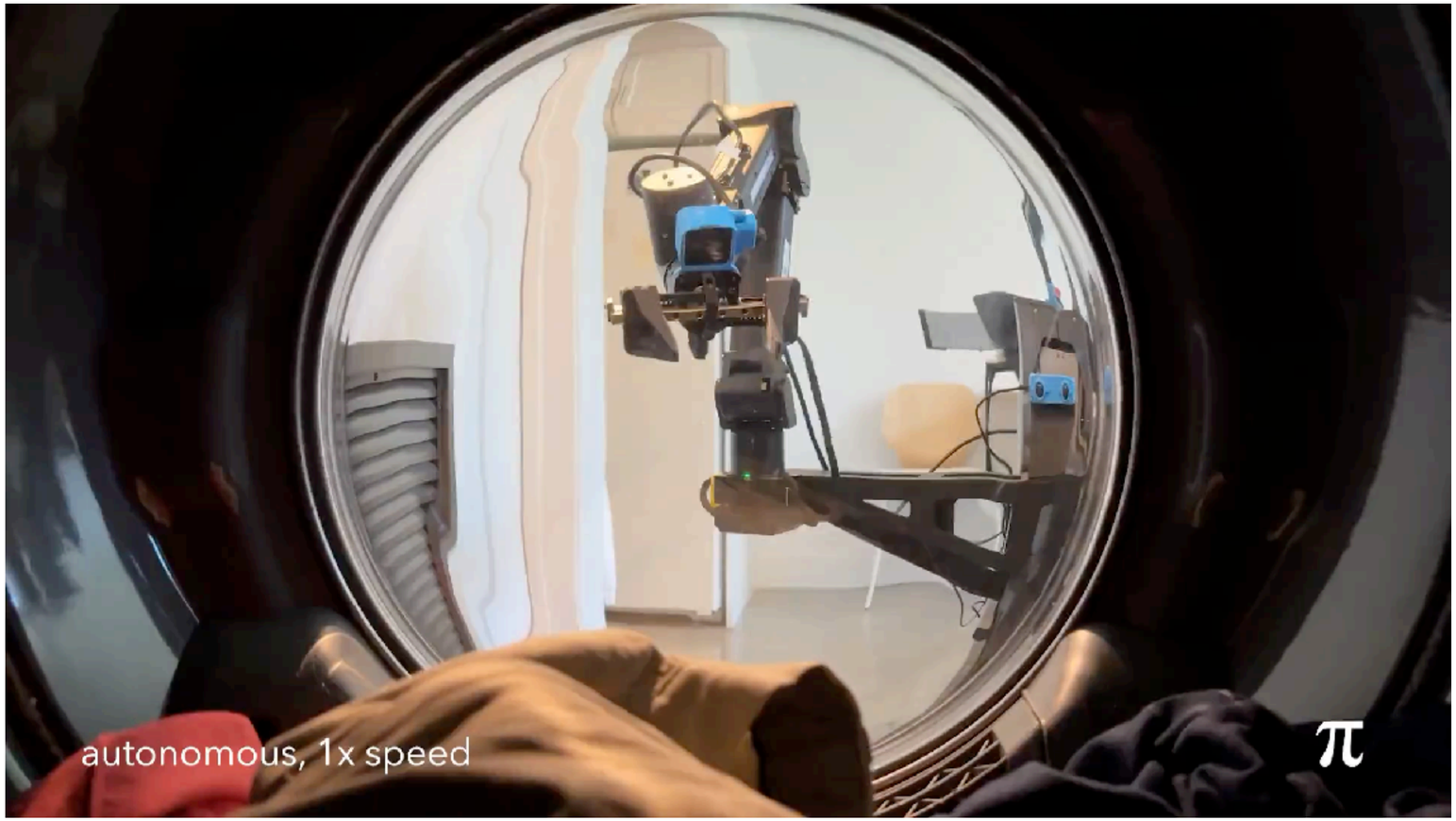
A strategic change.



In a statement, an OpenAI spokesperson told VentureBeat: “After advancing the state of the art in reinforcement learning through our Rubik’s Cube project and other initiatives, last October we decided not to pursue further robotics research and instead refocus the team on other projects. Because of the rapid progress in AI and its capabilities, we’ve found that other approaches, such as reinforcement learning with human feedback, lead to faster progress in our reinforcement learning research.”

But a new wave of AGI robots is happening

- Physical Intelligence (Pi) company founded by Sergey Levine and Chelsea Finn



https://www.physicalintelligence.company /

“A number of recent hires suggest that OpenAI’s robot efforts are now accelerating.”

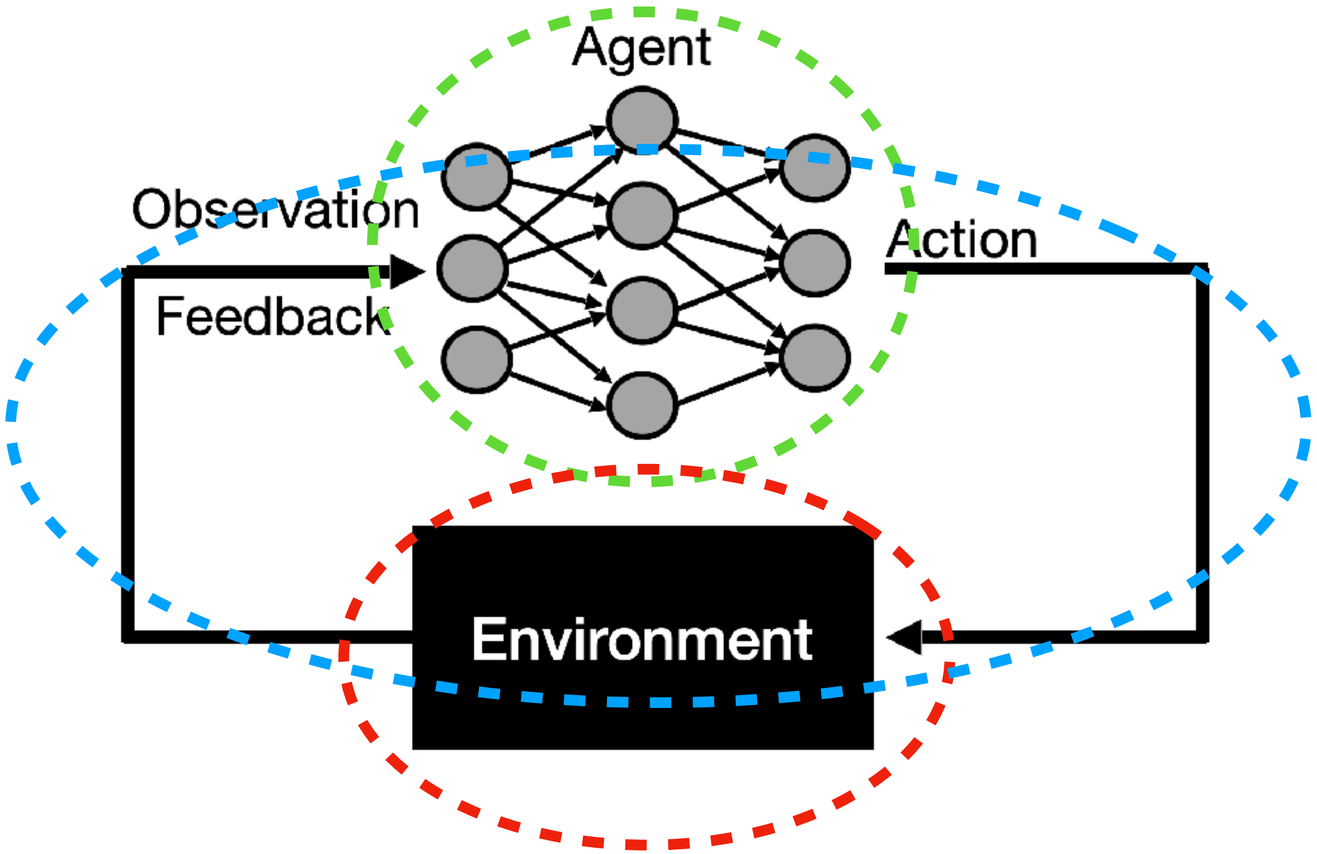
WILL KNIGHT BUSINESS SEP 15, 2025 6:00 AM

OpenAI Ramps Up Robotics Work in Race Toward AGI

The company behind ChatGPT is putting together a team capable of developing algorithms to control robots and appears to be hiring roboticists who work specifically on humanoids.

https://www.wired.com/story/openai-ramps-up-robotics-work-in-race-toward-agi/

What are the potential issues with RL?



Agent:

- Representation learning
- Interpretability
- Policy pre-training

Environment:

- Diversity of the environment
- Overfitting issue: training and testing in the same environment
- Reward engineering

Learning pipeline:

- No safety guarantee
- Very low sample efficiency

Assignment 0 and What's Next

- Assignment 0 is out (an easy warm-up) at https://github.com/ucla-rlcourse/cs260r-assignment-2026winter/
- RL example code: https://github.com/ucla-rlcourse/RLexample
- Next lecture: RL basics and coding with RL
- Please read **Sutton and Barton: Chapter 1 and Chapter 3**