	Week	11-Part	. Intro	, +0	Reinforce	ement	Learning
■ What	t is "	Reinforceme	ent les	urning	(RL)	5	
4	Learning	g through	experi	ence/d	uta to	make	good decisions
	under	uncertainty					
■ Dec	noizi	making	is an	essent	ial part	of in	telizence
* This	part	includes	Some	Slides	from	Emma	Brunskill

Week 11-Part 1: Intro to Reinforcement Learning

So far in the Course, we have studied techniques to identify things

Also in real life, we saw a lot of progress on what is called "perceptual machine learning", e.g., to perceive faces, cuts and dogs, ...

12 e.g., to perceive faces, cuts and dogs, digits, ...

Et perceptual machine lowning tries to identify something.

In reality, what we are trying to do is to make decision based on our Perception/information we receive.

Its So, it's critical to think about how to make "good" decisions, when it comes to intelligence.

How to Make Good Decision from limited Experience/Doda These Sort of questions, particularly when faced by uncertainty, had been studied in depth at least since 1950. RL builds strongly from theory and ideas starting in the 1950s with Richard Bellman So, why should we study RL? Because, understanding how to make good decisions from limited experience when faced by uncertainty is essential for any (artificial) intelligent entity

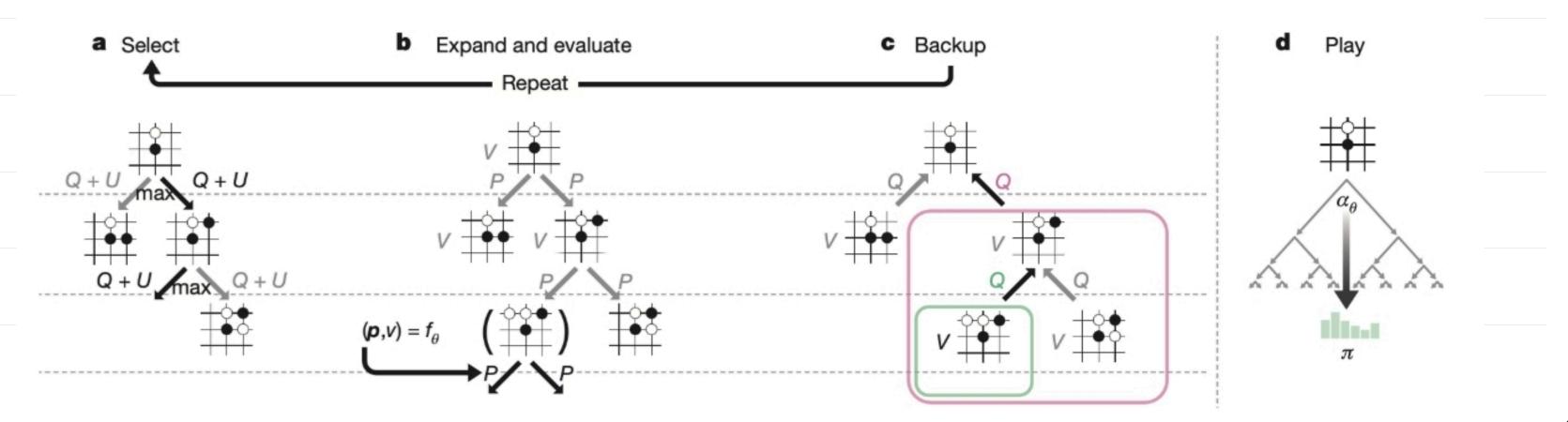
LB Also, because it's Cool. It's practical.

Some impressive successes in the last decade

Board game Go.

The An extremely hard board game.

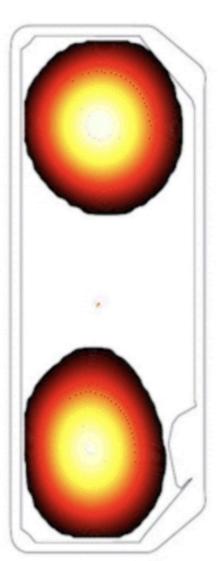
The In 2016, a team called "deepmind", by combining Rl and Monte-carlo-tree-search, they built an agent that could defeat the world champion.



Some impressive successes in the last decade.

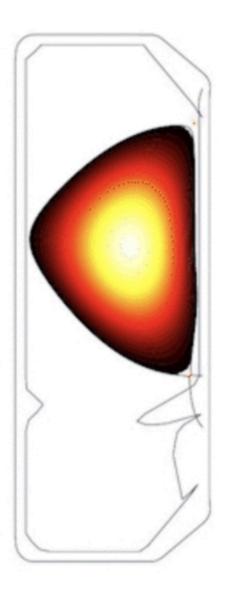
Plasma Control for Fusion Science

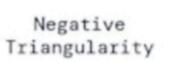


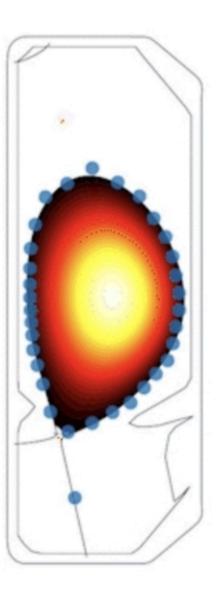


Droplets









ITER-like shape

Some impressive successes in the Efficient and targeted Covid-19 testing 4 Bastani et al. Noture 2021, "Efficient and targeted Gvid-19 border testing via reinforcement learning" Eva uses prior testing results to: optimize testing allocation Pseudonymized, produce risk estimates aggregated PLF form for grey-listing Central database Exit No test Eva Laboratory logs results in Test 24-48 h Submitted by visitors 24 h

Passenger tested

at port of entry.

Sample sent to laboratory

before entry

- Chal GPT

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

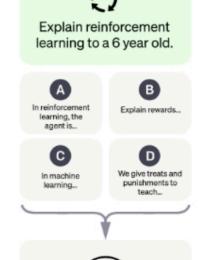
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

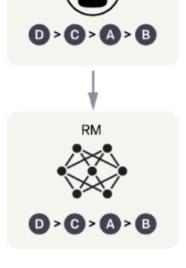
Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

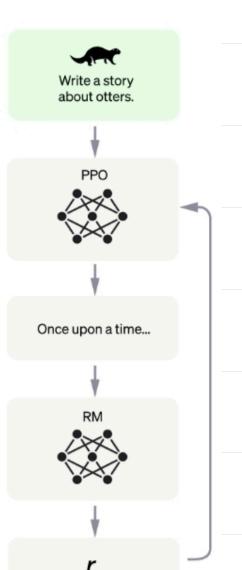
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



RL Generally Involves ...

- Optimi-zution
- Delayed Consequences
- E Exploration
- Generalization

Oplimization To find the best way to make decision

Figur decisions must yield to best outcomes or at least very good outcomes Best outomes? How can we specify if an outcome is the best outcome? How to compare outcomes?

We do it with an explicit notion of decision utility

/9

Delayed Consequences Decisions now Can impact things much later 43 e.g., Consider Saving for retirement Delayed Consequences introduces two challenges 43 When planning: Even when we know how the world works, decisions involves reasoning about not just immediate benefit of a decision, but also its long-term consequences. 41 When learning:

Delayed Consequences Decisions now Can impact things much later 43 e.g., Consider Saving for retirement Delayed Consequences introduces two challenges 43 When planning 4 When learning: When learning, we don't know how the world works. We want to learn it through direct world experience. But, temporal credit assignment is hard. · You take some action now, and later on you receive a good/bad outcome. How do you figure out which of your actions caused that good or bad letter result.

Exploration

We learn from direct experience from interacting with environment you only learn about what you try out.

To Don't know what would have happened for other decisions.

That's why it's important to sometimes explore alternative actions cause it may give you valvable information.

Generalization

Goud decisions are learnt trom post experience. 다 We need a mapping from possible states to decisions Why not just preprogram a decision Policy/mapping? 43 Because the number of possible states of the environment on be hyge. Atari Game 43 from a small set of states that we have seen we must learn a mapping that generalizes well to the states that we have not seen.

RL Vs. AI planning Vs. (UN) supervised learning

	Supervised Learning	Unsupervised Learning	AI planning	RL
Optimization (over actions)				
Learns from experience/data				√
Generalization				
Delayed Consequences				
Exploration				