

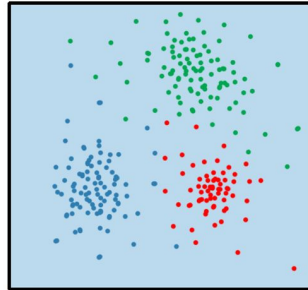
RL, 2025

Классификация областей ML

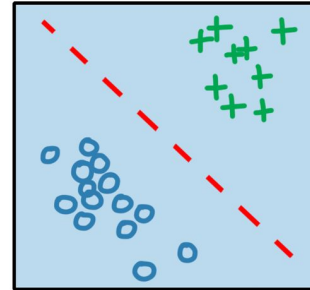
Обучение с подкреплением -
набор методов решение
задач принятия решений
методом проб и ошибок

machine learning

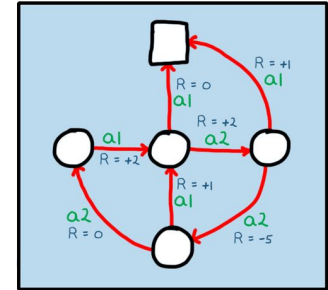
unsupervised
learning



supervised
learning



reinforcement
learning



Основные понятия

Агент (Agent) — это система, которая принимает решения и выполняет действия в окружающей среде.

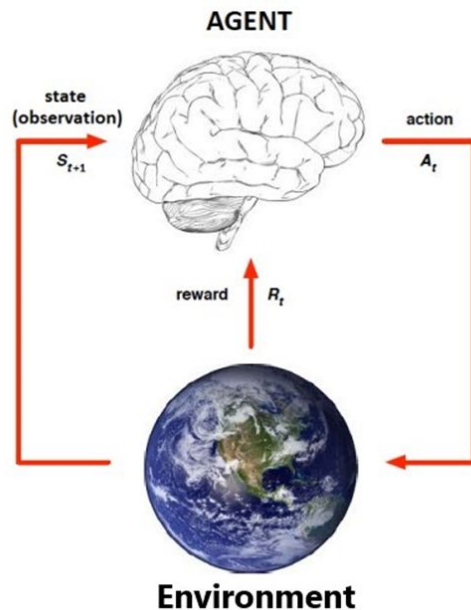
Среда (Environment) — это внешний мир, с которым агент взаимодействует.

Состояние (State) — это описание текущего положения агента в среде.

Действие (Action) — это выбор агента в каждом шаге взаимодействия с окружающей средой.

Политика (Policy) — это стратегия или правило, по которому агент выбирает действия в зависимости от состояния.

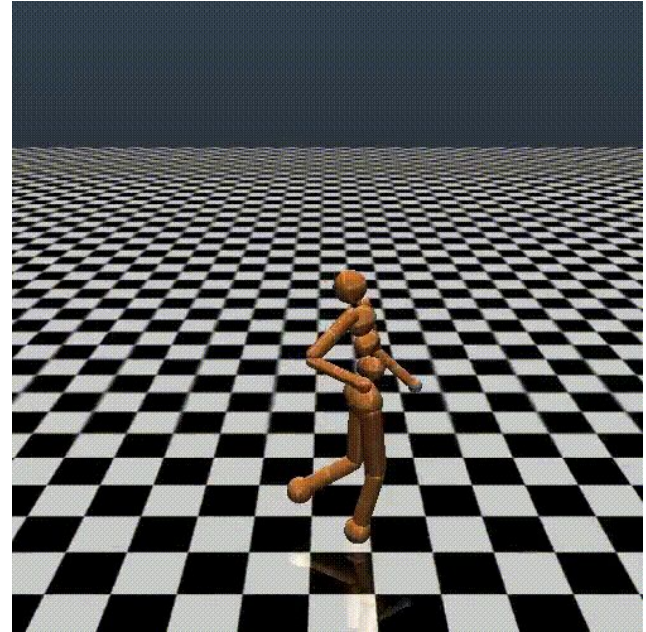
Награда (Reward) — это числовое значение, которое агент получает после выполнения действия в определенном состоянии. Награда отражает, насколько успешным было действие агента в конкретной ситуации. Агент стремится максимизировать суммарную награду.



Robotics (Locomotion)

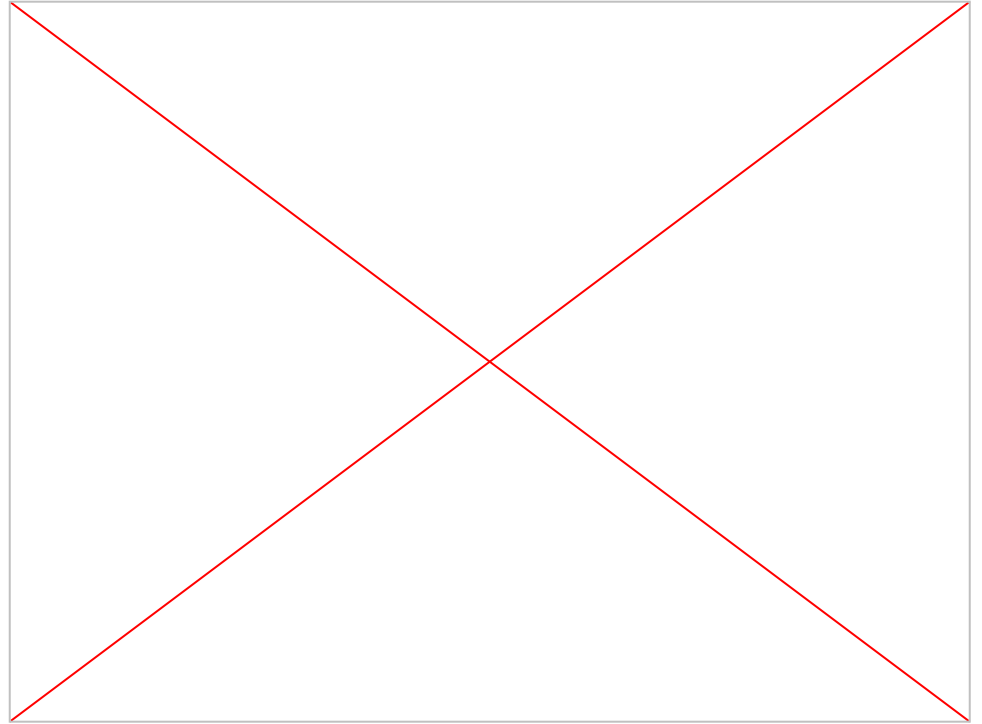
RL агент обучается в симуляции, а
затем переносится на железо.

Cartwheel A



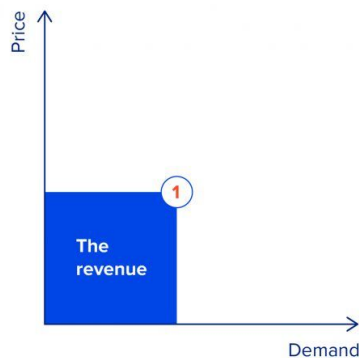
Robotics (Manipulation)

По экспертным демонстрациям работы манипулятора научиться повторять поведение (behavioral cloning)

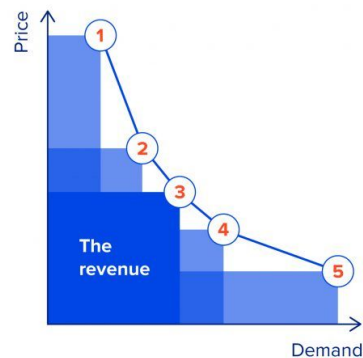


Табличный ML

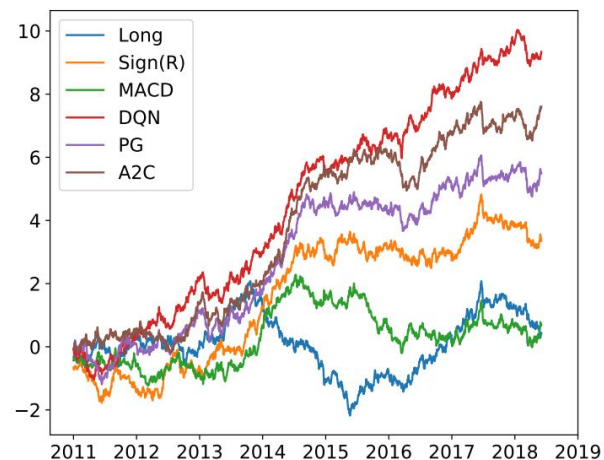
RL применяется в динамическом ценообразовании. Реже в трейдинге и рекомендательных системах.



Static pricing (single price point)

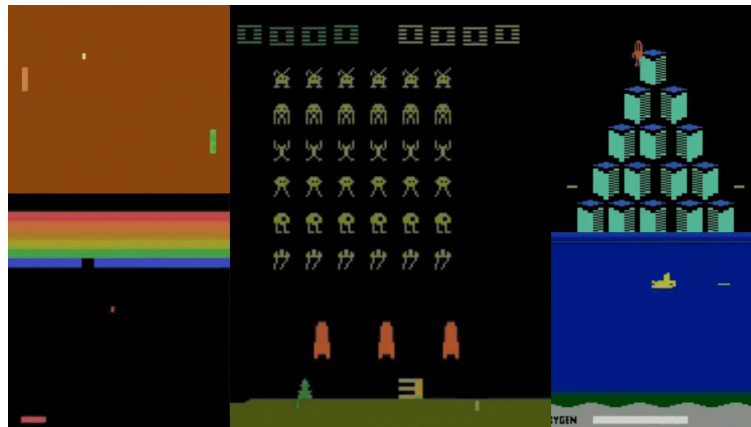
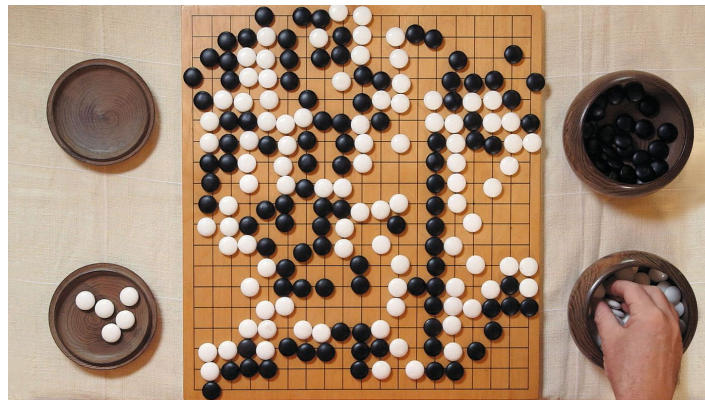
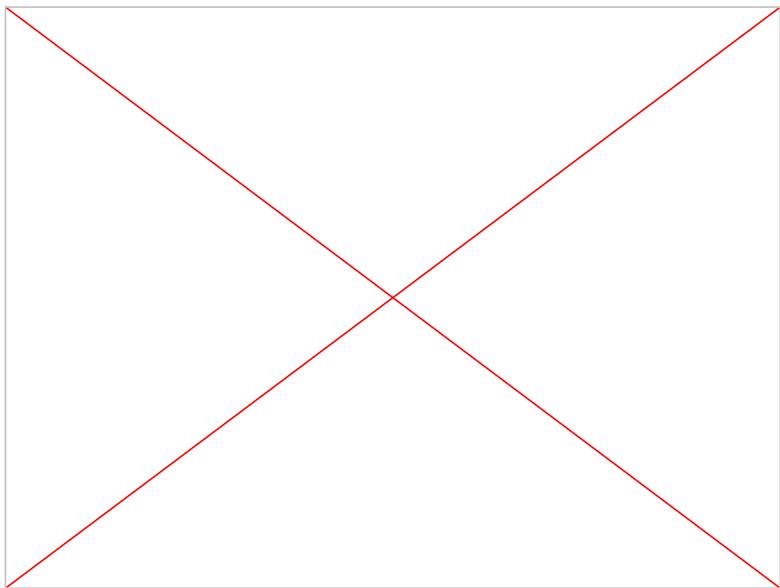


Dynamic pricing (multiple price point)



Игры и когнитивное поведение

RL прошел игры Atari, обыграл человека в Go, добыл алмаз в майнкрафте



LLM (Alignment)

RL используется для улучшения качества ответов генеративных моделей.

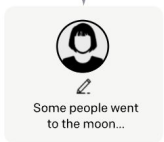
Step 1

Collect demonstration data, and train a supervised policy.

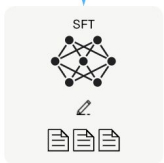
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



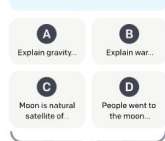
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

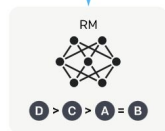
A prompt and several model outputs are sampled.



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 reinforcement learning.

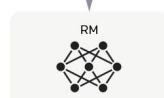
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



LLM (Reasoning)

RL используется для улучшения когнитивных способностей LLM (по крайней мере так кажется)

