

强化学习

第一讲：强化学习概述

教师：赵冬斌 朱圆恒 张启超

中国科学院大学
中国科学院自动化研究所



2021年3月

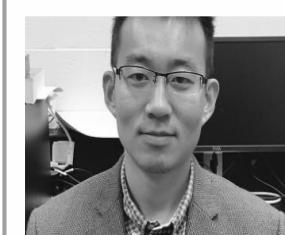
岁月静好，是因为有人负重前行！
春暖花开，不负韶华当全力以赴！

课程简介

- 课程名称: 强化学习(限120人)
- 时 间: 周五 5-7 节
- 地 点: 教1楼208
- 时 长: 40 课时 (上课 38 课时, 考试 2 课时)
- 评 分 标 准: 实验作业 1(20%)+ 实验作业 2(30%)+ 考试成绩(50%)



赵冬斌
研究员



朱圆恒
副研究员



张启超
副研究员



李论通
博士后



胡光政
博士研究生



李丁
博士研究生

课程简介-2020年评价



选课人数为：115人 参评人数为：108人 评估结果：97.01

首届“慧科杯”人工智能应用创新挑战赛获得最高一等奖（1/750+）

**2020首届“慧科杯”
人工智能应用创新挑战赛
获奖名单来啦！**

决赛获奖名单

名次	学校	作品名称	指导老师
一等奖	电子科技大学(成都)信息与电气工程学院	智能驾驶系统	黎晓峰 刘晓东 赵伟 周伟 钟伟
二等奖	北京邮电大学	AI语音识别系统	王立新
	中国科学院大学 - 中国科学院大学	基于深度学习的图像识别系统	郭雷 李小川 刘建伟 王立新
三等奖	清华大学	语音识别系统	李海林 刘洋 刘洋 刘洋 刘洋
	清华大学	无人驾驶系统	王浩宇 周洋 周洋 周洋
	北京理工大学	自动驾驶系统	王海峰 孙广亮 周洋 周洋
最佳创意奖	国防科技大学	人脸识别系统	李海林 刘洋 刘洋 刘洋
	北京邮电大学	语音识别系统	王立新
	中国科学院大学 - 中国科学院大学	基于深度学习的图像识别系统	郭雷 李小川 刘建伟 王立新
	北京理工大学	自动驾驶系统	王海峰 孙广亮 周洋 周洋

优秀奖获奖名单

名次	学校	作品名称	指导老师
一等奖	电子科技大学(成都)信息与电气工程学院	智能驾驶系统	黎晓峰 刘晓东 赵伟 周伟 钟伟
二等奖	北京邮电大学	AI语音识别系统	王立新
	中国科学院大学 - 中国科学院大学	基于深度学习的图像识别系统	郭雷 李小川 刘建伟 王立新
三等奖	清华大学	语音识别系统	李海林 刘洋 刘洋 刘洋
	清华大学	无人驾驶系统	王浩宇 周洋 周洋 周洋
	北京理工大学	自动驾驶系统	王海峰 孙广亮 周洋 周洋
最佳创意奖	国防科技大学	人脸识别系统	李海林 刘洋 刘洋 刘洋
	北京邮电大学	语音识别系统	王立新
	中国科学院大学 - 中国科学院大学	基于深度学习的图像识别系统	郭雷 李小川 刘建伟 王立新
	北京理工大学	自动驾驶系统	王海峰 孙广亮 周洋 周洋

课程简介-2020年评价

2020年IEEE CoG Fighting Game
AI Competition, FTGAIC格斗游戏，
第5名



	JPN	GUARDA	LUO
Fuzzy_ZYQAI	0	0	0
TeraThunder	32	0	10
SpringAI	100	0	0
ErheaAI	79	10	17
ERHEA_A	0	10	20
Caseline	0	0	0
MrTwo	21	8	0
CVR_AI	98	8	15
LuoyiAI	6	2	0
ButcherPudge	8	12	0
HTAI	0	8	0
Uterus_Zen	4	10	0
MonkeyLink_TriplePM	0	0	0
Noobbot	0	0	0
SampleMCTSAI	1	2	0

	JPN	GUARDA	LUO
Fuzzy_ZYQAI	0	0	0
TeraThunder	18	0	17
SpringAI	121	0	15
ErheaAI	0	8	4
ERHEA_A	20	10	12
Caseline	0	8	1
MrTwo	0	10	4
CVR_AI	0	4	15
LuoyiAI	10	0	0
ButcherPudge	18	10	0
HTAI	0	8	0
Uterus_Zen	6	10	0
MonkeyLink_TriplePM	0	0	0
Noobbot	0	0	0
SampleMCTSAI	0	10	0

	SUM	RANK
Fuzzy_ZYQAI	0	13.5
TeraThunder	30	0
SpringAI	98	8
ErheaAI	73	4
ERHEA_A	128	1
Caseline	11	11
MrTwo	29	9
CVR_AI	73	5
LuoyiAI	26	8.5
ButcherPudge	13	10
HTAI	0	13.5
Uterus_Zen	36	7
MonkeyLink_TriplePM	0	13.5
Noobbot	0	13.5
SampleMCTSAI	20	9.5

- Winner AI: ERHEA_AI by Zhentao Tang*, Rongqin Liang, and Mengchen Zhao (*2019 runner-up), University of Chinese Academy of Sciences and Huawei Noah's Ark Lab, China
 - Rolling Horizon Evolutionary Algorithm combined with an adaptive learning-based opponent model (Deeplearning4j) utilizing two simulation modules from ReiwaThunder (2019 Winner) (cf. the ArXiv paper in slide 5)
- Runner-up AI: Tera Thunder by Eita Aoki (winner for the last four consecutive years), Japan
 - 1. Prioritize certain actions in advance. 2. Predict the most possible three actions by the opponent. 3. Select the best AI action against the opponent's three actions using his original simulator.
- 3rd Place AI: ButcherPudge by Wen Bai (newcomer), Nanyang Technological University, Singapore
 - Reinforcement Learning Algorithm SAC (Soft-Actor-Critic) trained against 2019 top AIs with the OpenAI gym interface and Pytorch library.

1. 强化学习概述
2. 马尔可夫过程+ 第1次作业安排
3. 动态规划
4. 无模型预测方法
5. 无模型控制方法
6. 基于逼近器实现的强化学习算法
7. 策略梯度方法1
8. 策略梯度方法2+小组研讨课1
9. 强化学习应用+第2次作业安排
10. 逆强化学习
11. 深度强化学习算法
12. 深度强化学习与智能驾驶
13. 小组研讨课2

课程简介



部分讲义取自以下参考资料，仅用于教学和交流。

Sutton & Barto, 1998/2018, “Reinforcement Learning: An Introduction”

<http://incompleteideas.net/book/the-book-2nd.html>



Andrew G. Barto



Richard S. Sutton

David Silver, University College London Course on Reinforcement Learning

<http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Teaching.html>



David Silver

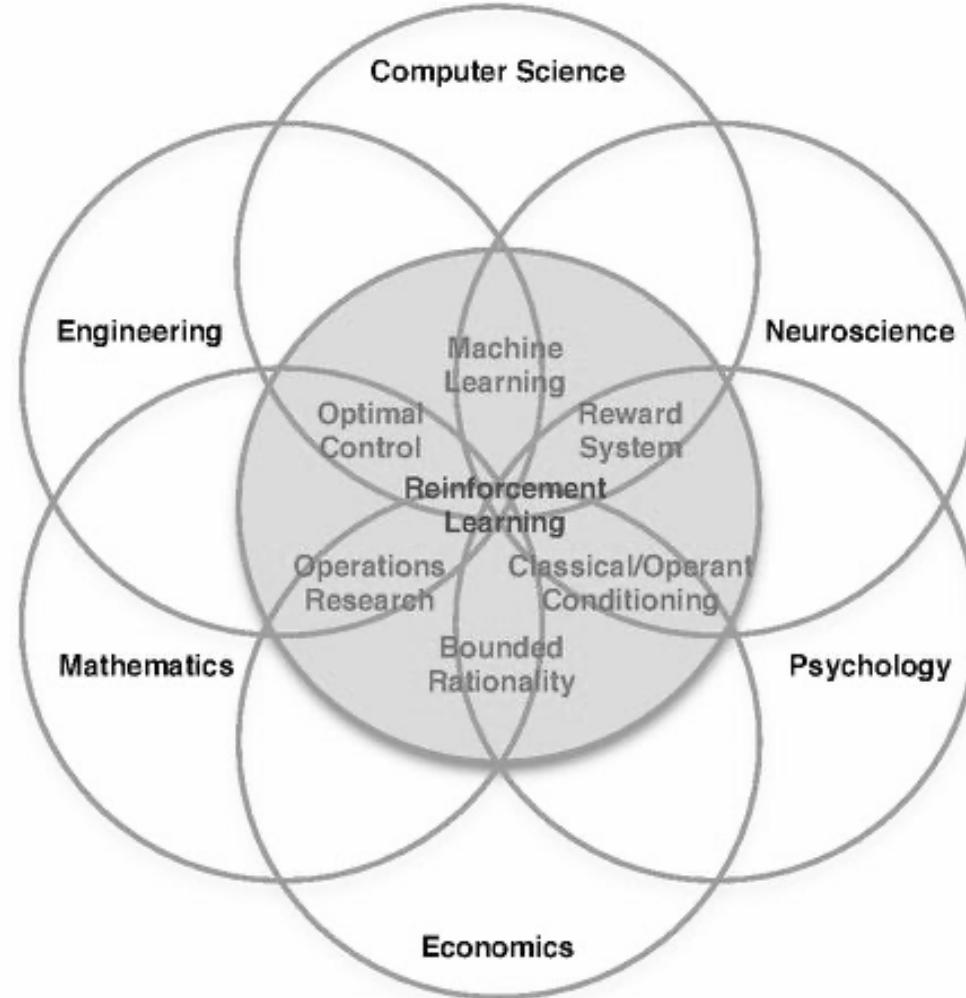
Emma Brunskill, Stanford CS234 Reinforcement Learning

Sergey Levine, UC Berkeley CS 294 Deep Reinforcement Learning

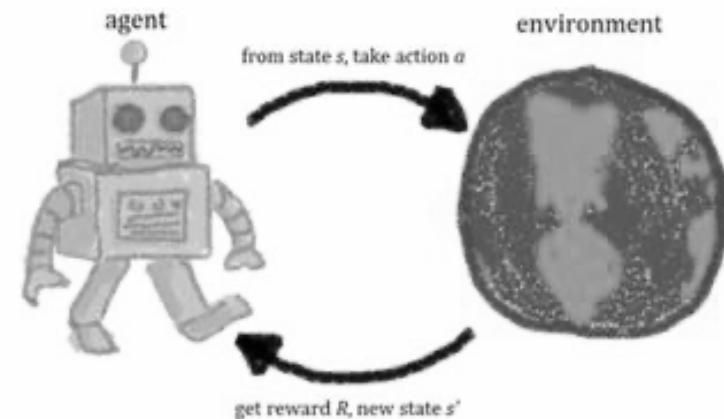
.....

李宏毅，机器学习

强化学习介绍



- 强化学习是一种优化智能体在环境中行为的一种方法。根据环境反馈的奖励，调整智能体的行为策略，提升智能体实现目标的能力



■ 巴普洛夫实验



■ 巴普洛夫实验



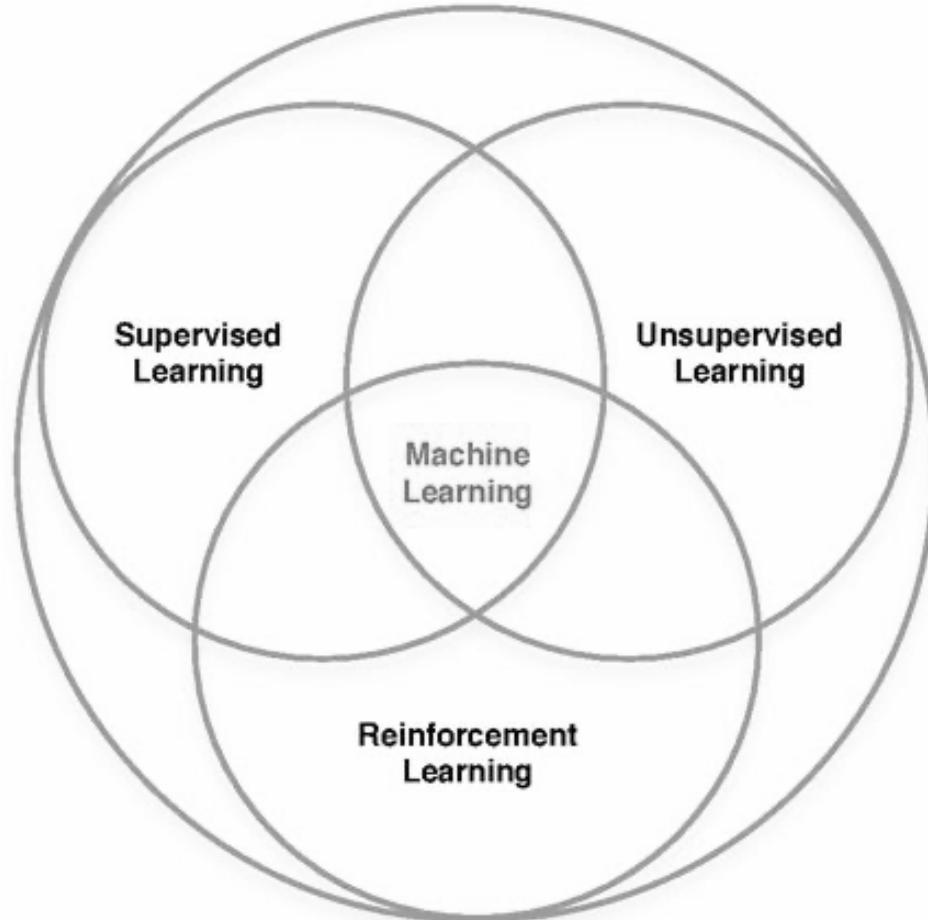
- 智能体：小狗
- 状态：有无听到铃声
- 动作：流口水
- 奖励：骨头

- 根据环境反馈的奖励，调整智能体的行为策略，提升智能体实现目标的能力。
 - 没有明确告诉采取哪些动作是可以实现目标
 - 通过间接的奖励信号反映完成目标的情况(稀疏)
 - 例如：下棋输赢 +1/-1, 汽车行驶碰撞 +1/-1, 机器人离目标点的距离
 - 好处：简单，便宜

- 强化学习也称为试错法 (trial-and-error) , 通过智能体和环境的交互得到反馈的信号
 - 有失败也有成功
- 强化正反馈的策略，避免负反馈的策
- 不太适合于无法进行大量实验的场景
 - 比如安全因素 (开车碰撞)、成本原因 (读博)
- 但是如果能够建立足够精确的仿真模型，在仿真环境使用强化学习方法得到的策略，在现实世界依然好用

强化学习与其它机器学习的不同

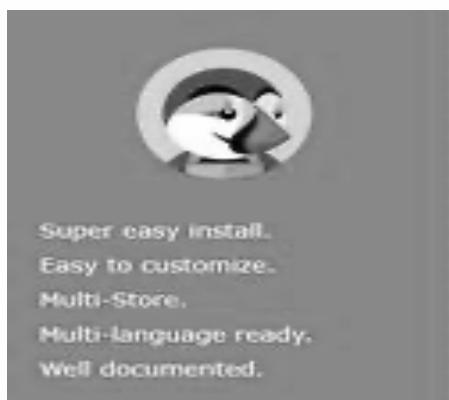
机器学习分支 Machine Learning (ML)



■ 监督学习/非监督学习应用领域



图像分类



自然语言处理

Frequently Bought Together

total price: \$63.09

Add both to Cart

Add both to List

B The Little Schemer - 4th Edition

B Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and ... by Harold Abelson Paperback \$32.00

B The Pragmatic Programmer: From Journeyman to Master by Andrew Hunt Paperback \$32.00

Customers Who Bought This Item Also Bought

The Little Schemer - 4th Edition

Introduction to Algorithms, 3rd Edition (MIT Press)

The Pragmatic Programmer: From Journeyman to Master

Introduction to Functional Programming Using Lambda Calculus

An Introduction to Functional Programming Through Lambda Calculus

Purely Functional Data Structures

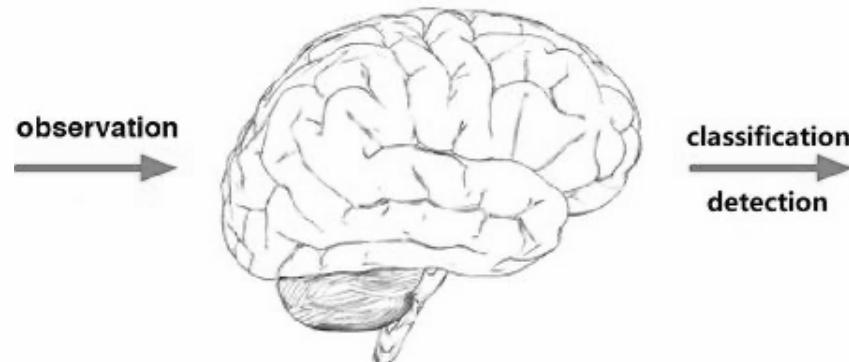
Code: The Hidden Language of Computer Hardware and Software

The Little Prover (MIT Press)

推荐系统

- 智能体处在特定的环境中产生一系列的动作，而这些动作改变智能体的状态。
- 举例
 - 1 遥控直升飞机的特技表演
 - 2 打败围棋世界冠军
 - 3 管理股票证券
 - 4 发电厂调控
 - 5 控制人型机器人双足行走
 - 6 视频游戏上超越人类
- 强化学习考虑的是 序贯决策过程

■ 感知：识别或估计观测的内容

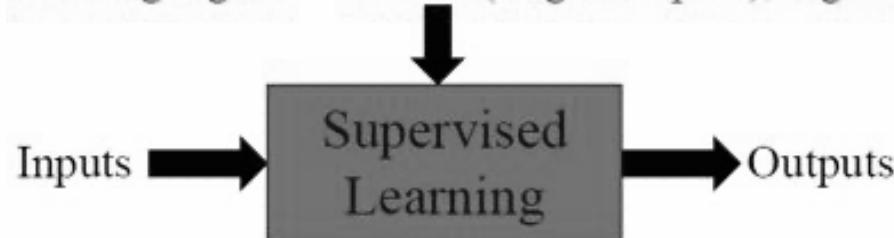


■ 决策：根据观测做出行为

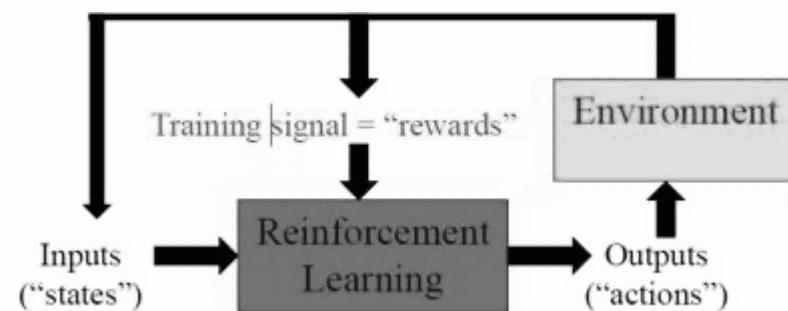


■ 感知：识别或估计观测的内容

Training signal = desired (target outputs), e.g. class



■ 决策：根据观测做出行为



- 强化学习：
- 监督学习 (Supervised Learning, SL)/非监督学习 (Unsupervised Learning, USL):

- 强化学习：

- ① 产生的结果（动作）能够改变数据的分布（状态）

- 监督学习 (Supervised Learning, SL)/非监督学习 (Unsupervised Learning, USL):

- ① 产生的结果（输出）不会改变数据的分布

■ 强化学习：

- 1 产生的结果（动作）能够改变数据的分布（状态）
- 2 最终的目标可能要很长时间才能观察到/奖励稀疏 (e.g. 下棋)

■ 监督学习 (Supervised Learning, SL)/非监督学习 (Unsupervised Learning, USL):

- 1 产生的结果（输出）不会改变数据的分布
- 2 结果是瞬时的/输出误差

■ 强化学习：

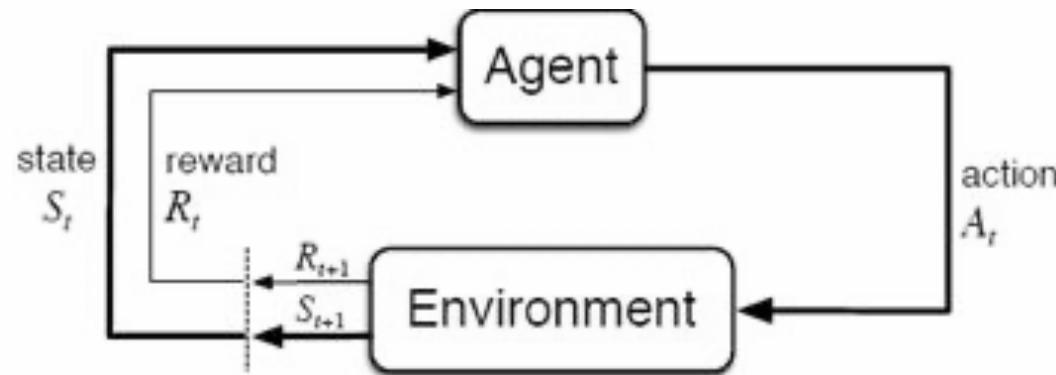
- 1 产生的结果（动作）能够改变数据的分布（状态）
- 2 最终的目标可能要很长时间才能观察到/奖励稀疏 (e.g. 下棋)
- 3 没有明确的标签 (label) 数据
- 4 根据当前的奖励，最终实现长远的目标

■ 监督学习 (Supervised Learning, SL)/非监督学习 (Unsupervised Learning, USL):

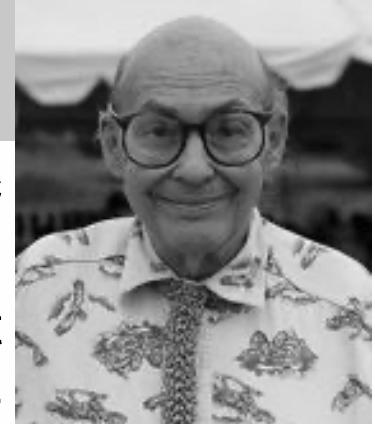
- 1 产生的结果（输出）不会改变数据的分布
- 2 结果是瞬时的/输出误差
- 3 要么有明确的标签数据 (SL)
- 4 要么完全没有任何标签数据 (USL)

强化学习发展历史

强化学习和马尔可夫决策过程(第2讲)



Stochastic
Neural-Analog
Reinforcement
Calculator, 1954



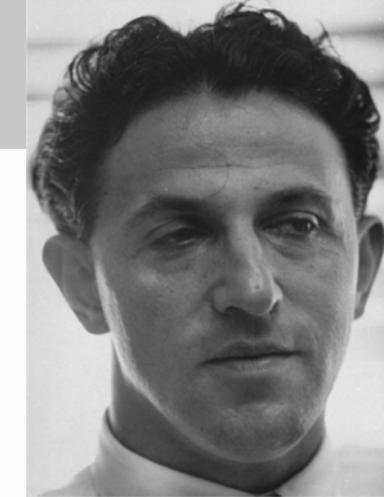
Marvin Minsky
1927-2016

- **马尔可夫决策过程**：个体未来的状态只与当前时刻的状态 S_t 有关，而与过去的状态 $\{S_1, \dots, S_{t-1}\}$ 无关
- 状态 S (观测 O)，动作 A ，奖赏 R ，策略 π
- 智能体通过直接与环境交互，学习出能够最大化长期的累积期望奖赏的策略。
- 目标：使值函数最大

$$v_\pi(s) = \mathbb{E}_\pi [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s]$$

A 10x10 grid representing a reinforcement learning environment. The grid contains various symbols: arrows pointing up, down, left, and right, representing possible actions; numbers like 0.00, 0.01, 0.02, etc., representing rewards; and some empty squares. The grid is divided into four quadrants by thick lines, suggesting different regions or states.

动态规划(Dynamic Programming, DP 1957)



- 最优策略：一个最优化策略具有这样的性质，不论过去状态和决策如何，对前面的决策所形成的状态而言，余下的诸决策必须构成最优策略。

$$v_{\pi}(s) = \mathbb{E}_{\pi} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

Richard Ernest Bellman
1920-1984

$$v_{\pi}(s) = \mathbb{E}_{\pi} [R_{t+1} + \gamma v_{\pi}(S_{t+1}) | S_t = s]$$

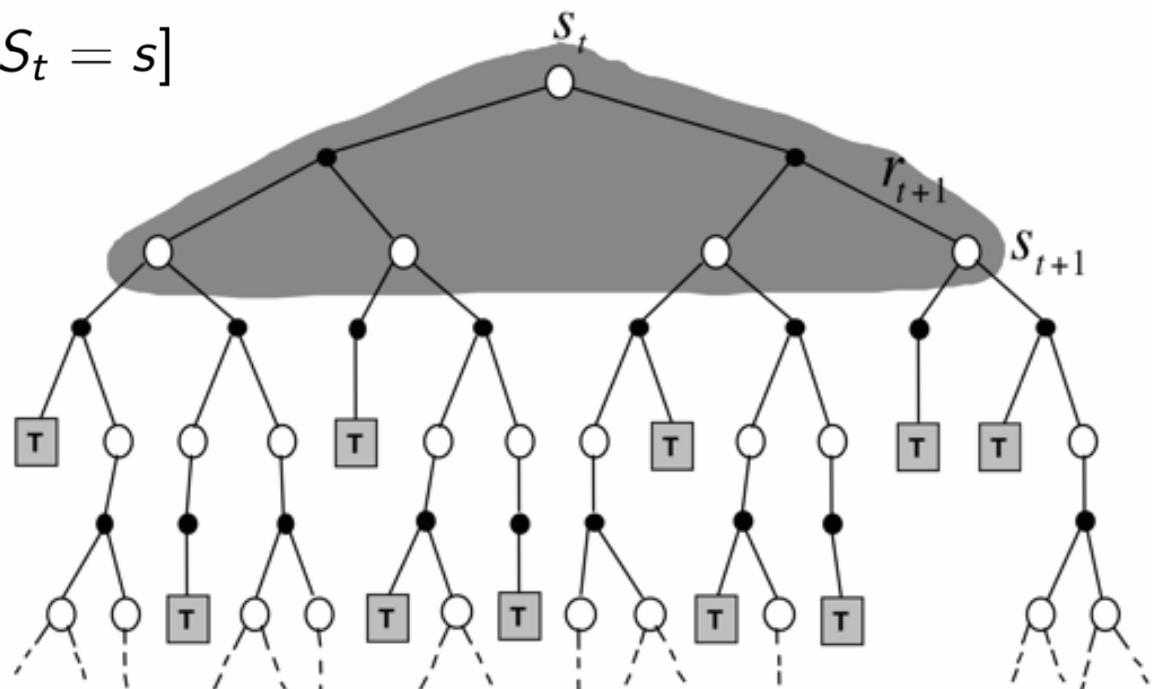
$$v_{\pi} = \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} v_{\pi}$$

$$v_{\pi} = (I - \gamma \mathcal{P}^{\pi})^{-1} \mathcal{R}^{\pi}$$

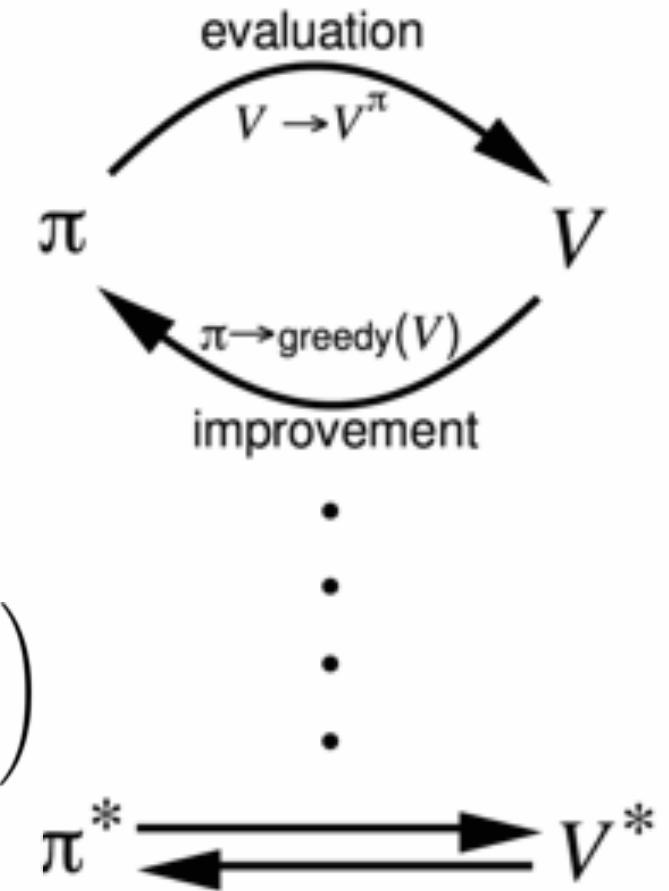
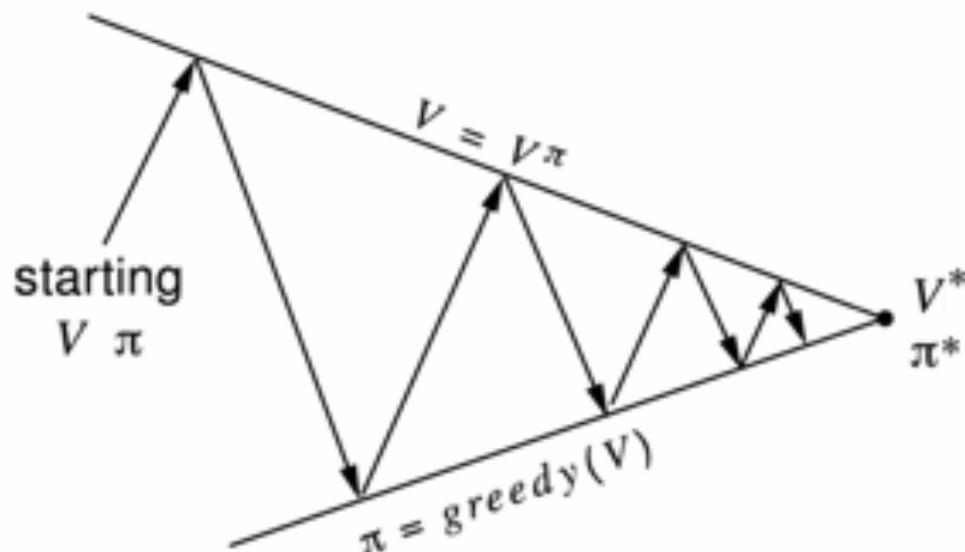
维数灾：离散状态、动作空间大

$$v_*(s) = \max_a \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

策略迭代、值迭代



策略迭代/值迭代（第3讲）



策略迭代

$$v_{k+1}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

$$\mathbf{v}^{k+1} = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathbf{v}^k$$

$$\pi' = \text{greedy}(v_\pi)$$

值迭代

$$v_{k+1}(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

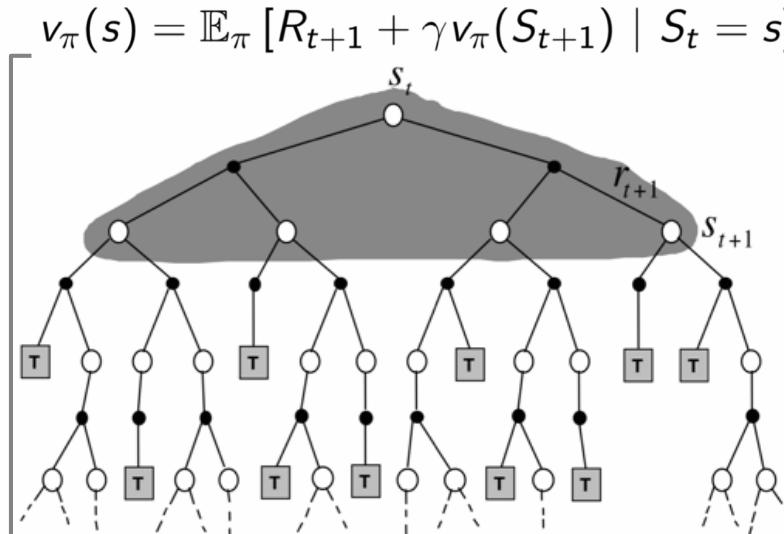
$$\mathbf{v}_{k+1} = \max_{a \in \mathcal{A}} \mathcal{R}^a + \gamma \mathcal{P}^a \mathbf{v}_k$$

蒙特卡洛算法和时间差分学习算法（第4讲）



Richard S. Sutton
Temporal-Difference
(TD) 1988

动态规划

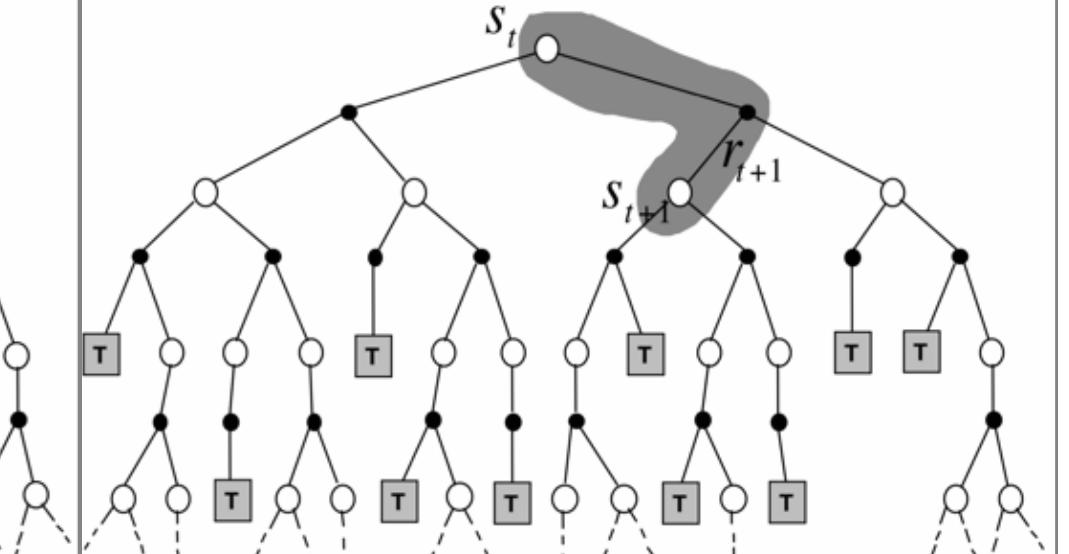
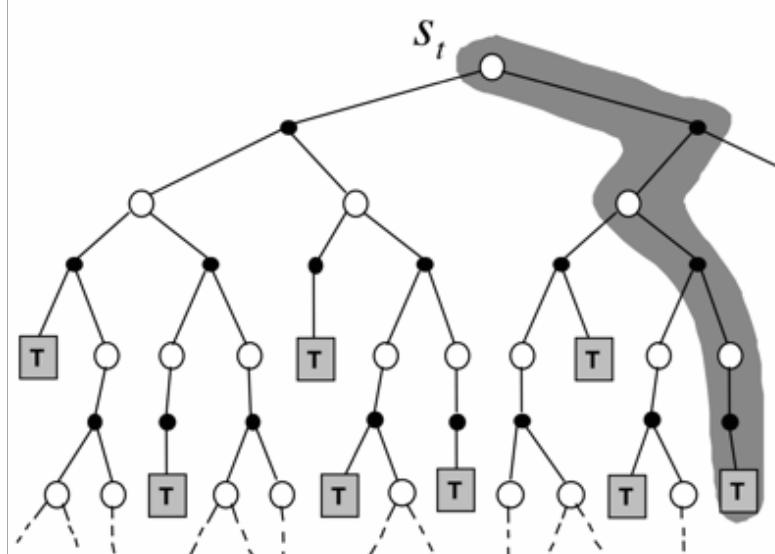


蒙特卡洛算法

时间差分学习算法

$$V(S_t) \leftarrow V(S_t) + \alpha (G_t - V(S_t))$$

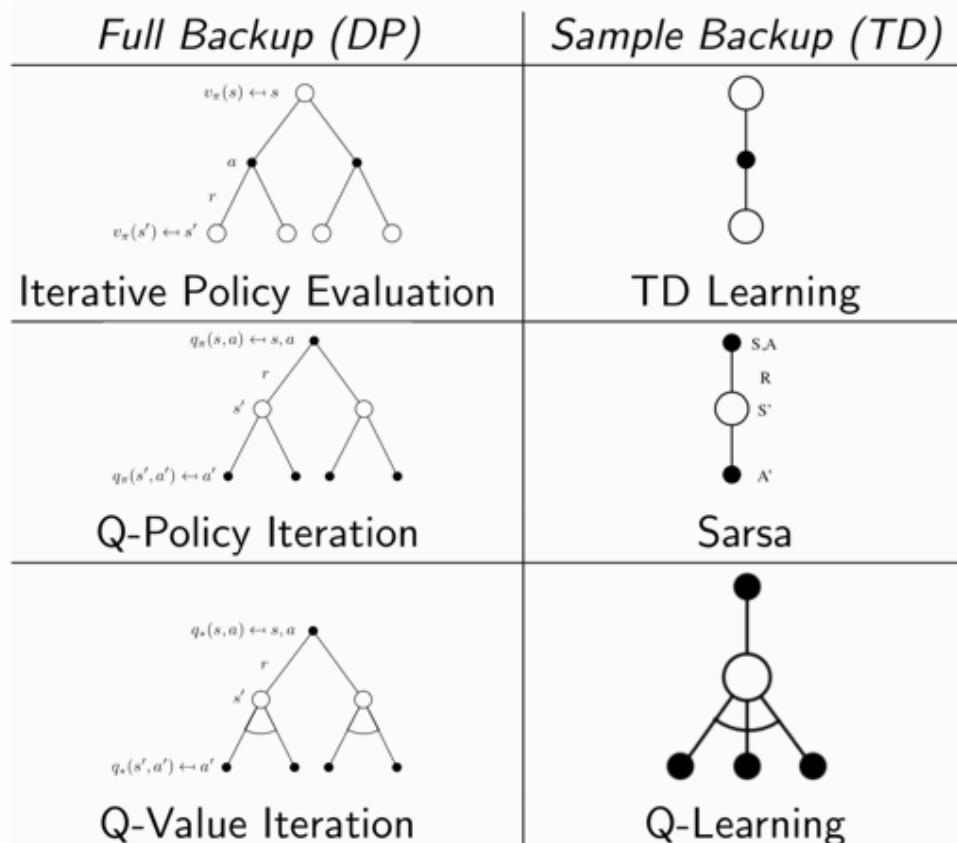
$$V(S_t) \leftarrow V(S_t) + \alpha (R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$



无模型学习控制：Sarsa和Q学习（第5讲）



$$q_{\pi}(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a \sum_{a' \in A} \pi(a'|s') q_{\pi}(s', a')$$



Sample Backup (TD)

TD Learning

$$V(S) \xleftarrow{\alpha} R + \gamma V(S')$$

Sarsa

$$Q(S, A) \xleftarrow{\alpha} R + \gamma Q(S', A')$$

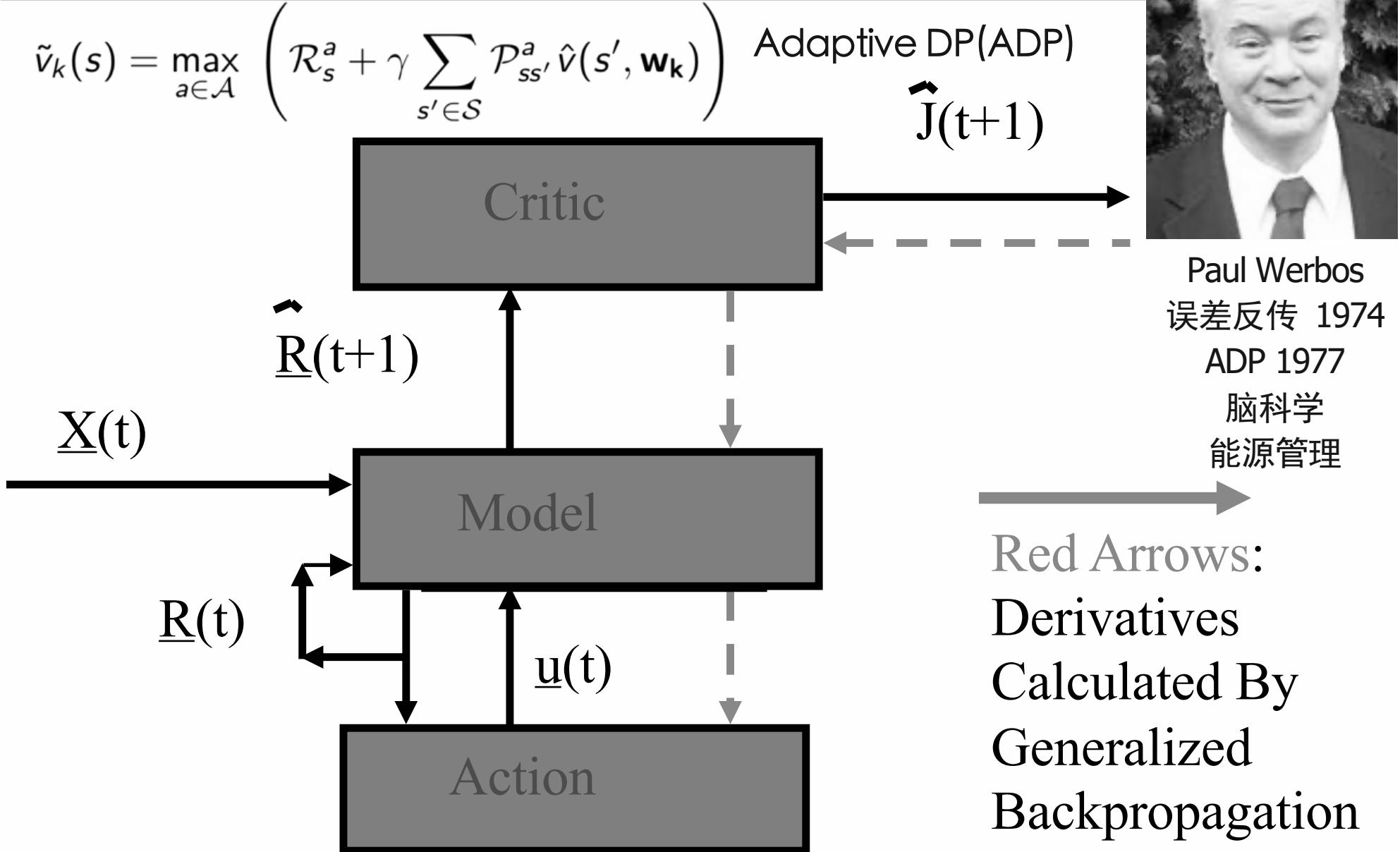
Q-Learning

$$Q(S, A) \xleftarrow{\alpha} R + \gamma \max_{a' \in A} Q(S', a')$$

Chris Watkins
Q学习, 1989

*Watkins, C. J. C. H. (1989). Learning from Delayed Rewards. Ph.D. thesis, University of Cambridge.

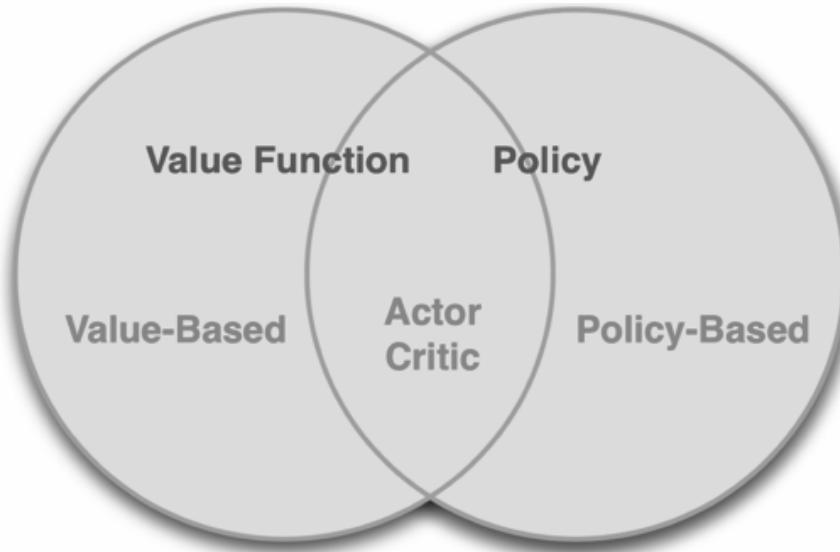
函数逼近(第6讲)



Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. [http://www.werkbund.org/](#) ↗ ↘ ↙
General Systems Yearbook, 22(12):25–38.

策略梯度 (第7-8讲)

Ronald J. Williams
REINFORCE 1992



$$J(\theta) = \mathbb{E}_{\pi_\theta} [r] \\ = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \pi_\theta(s, a) \mathcal{R}_{s,a}$$

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \pi_\theta(s, a) \nabla_\theta \log \pi_\theta(s, a) \mathcal{R}_{s,a} \\ = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) r]$$

$$\begin{aligned} \nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) v_t] \\ &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) Q^w(s, a)] \\ &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) A^w(s, a)] \\ &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) \delta] \\ &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) \delta e] \end{aligned}$$

$$G_\theta^{-1} \nabla_\theta J(\theta) = w$$

REINFORCE

Q Actor-Critic

Advantage Actor-Critic

TD Actor-Critic

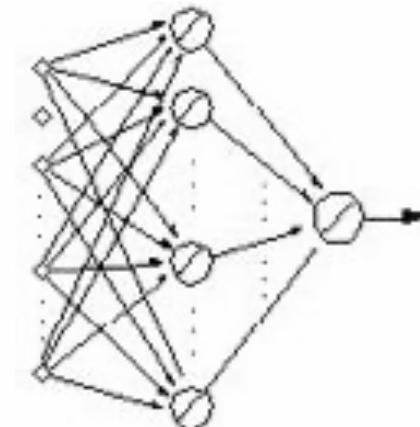
TD(λ) Actor-Critic

Natural Actor-Critic

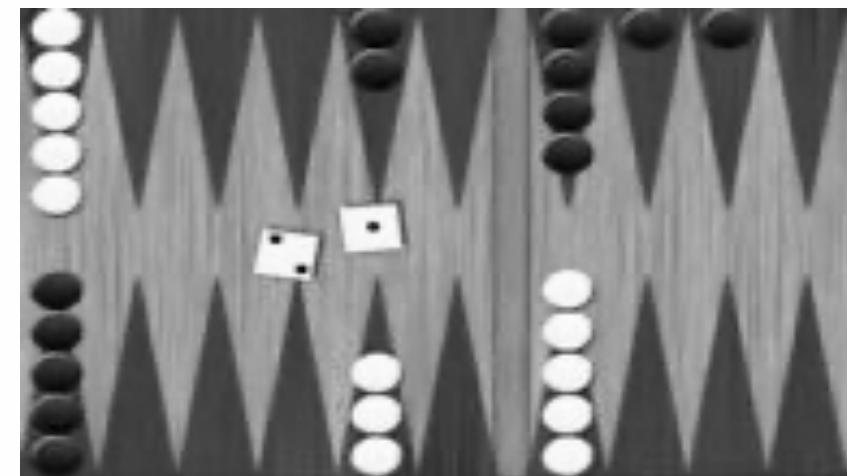
TD-Gammon

- 1992年，Tesauro等成功使用强化学习使西洋双陆棋达到了大师级水准。

- 完全信息零和博弈问题
- 特征：双方各执15枚棋子；投掷色子引入随机性
- 奖励回报设置：赢1输0
- 网络结构：隐层节点数为10-40的前馈网络
- 训练数据：20万盘数据
- 训练时间：2周

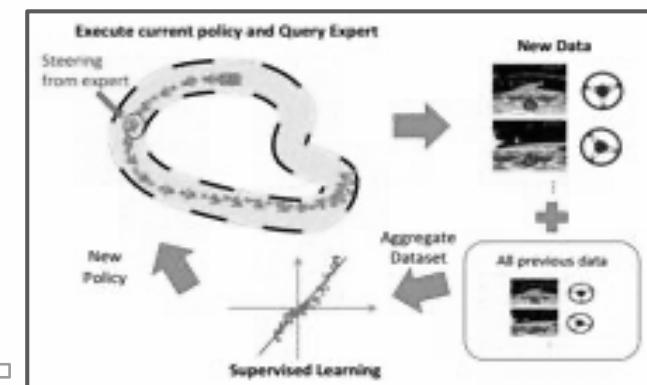


Gerald Tesauro
Backgammon 1992
IBM Watson
Deep Blue
.....



奖赏信号：难以人为事先给定

- 2000年，吴恩达，线性逆强化学习，执行最优策略所带来的样本（例如：人示范开车的行为）。
- 2004，Abbeel和吴恩达，根据从示范样本中学习奖赏函数形成了学徒学习算法(Apprentice Learning)。
- 2008年，Ziebart等人提出了最大熵逆强化学习，把原有的线性规划问题转化成了优化最大熵函数的问题，此时求得的奖赏函数是唯一的。
- 2011年，Ross等人提出了DAGGER算法：解决示范样本和学习过程中产生的样本可能不来自同一个分布的问题。

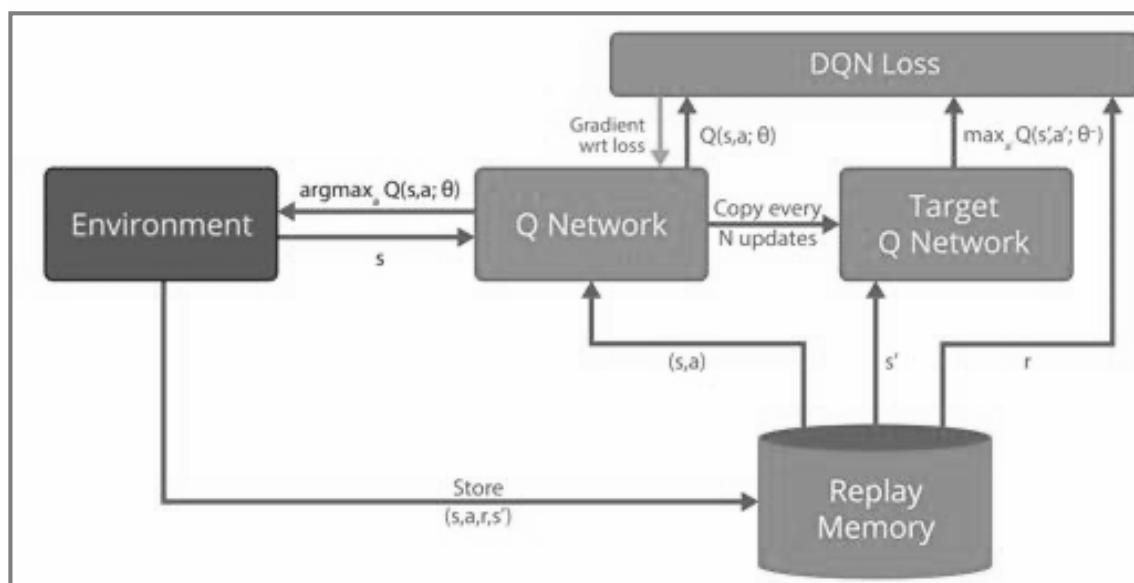
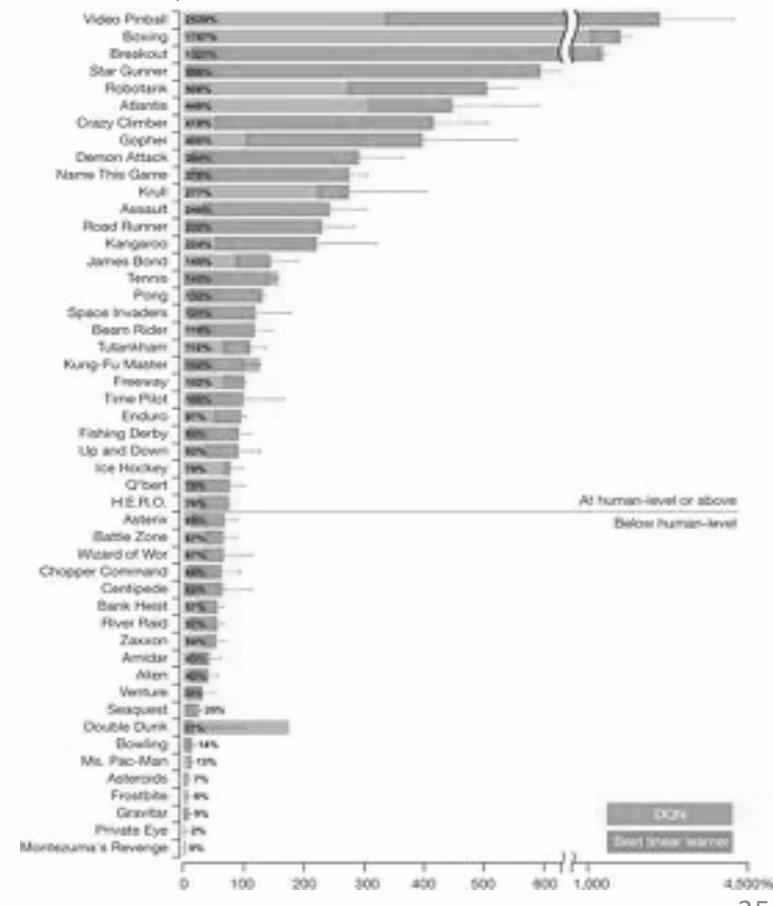


DQN (第11讲)

Volodymyr Mnih
多伦多大学
DQN, A3C

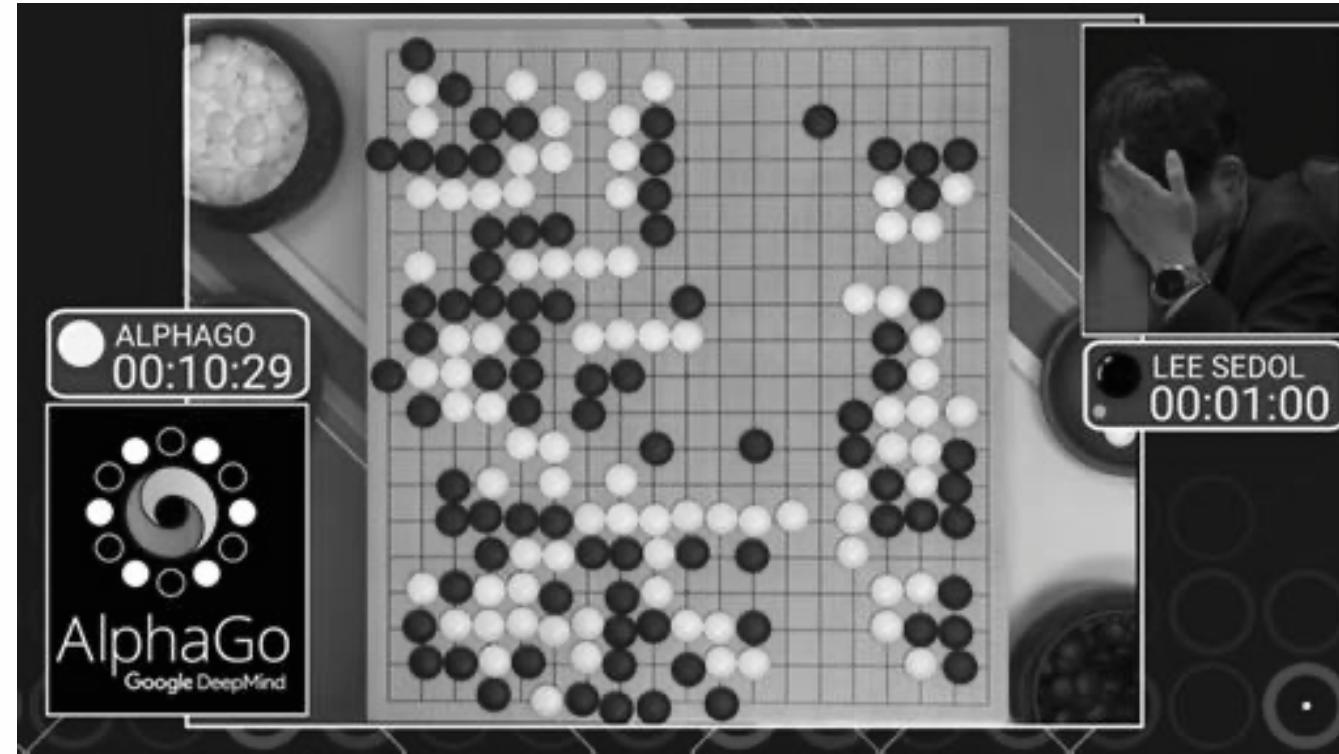


DeepMind2015年2月发表在Nature上的论文,提出了DQN算法,将卷积神经网络和Q学习结合,并集成了经验回放技术,在57款Atari游戏上超过了人类水平。



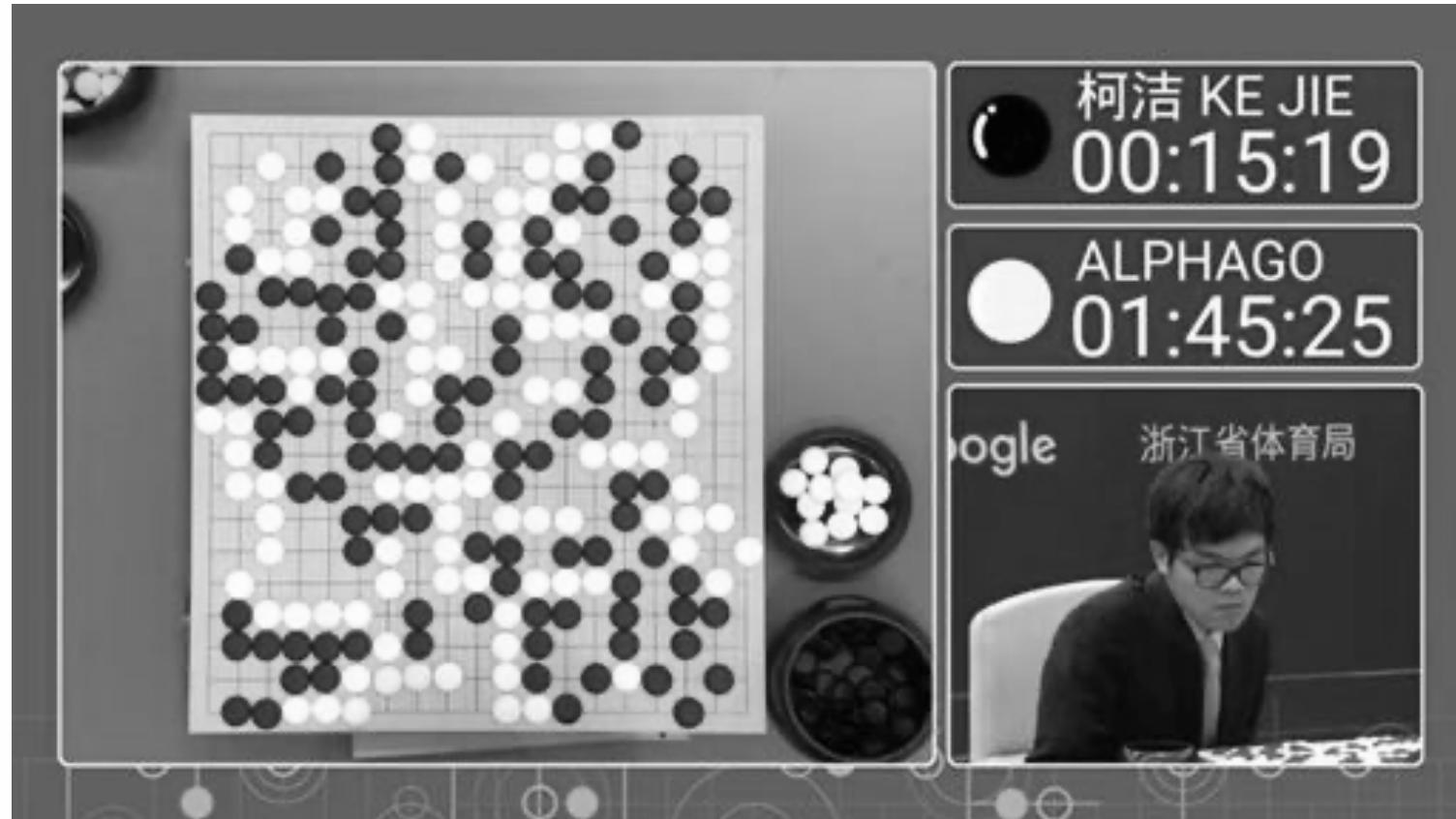
AlphaGo（第11讲）

David Silver
UCL教授
Alpha系列



- 2016 年 3 月, DeepMind 开发的 AlphaGo 围棋程序 4-1 战胜 Lee Sedol(前世界排名第一)
- DeepMind, *Nature*, 2016

赵, 邵, 朱, 李, 陈, 王, 刘, 周, 王。深度强化学习综述: 兼论计算机围棋的发展, 控制理论与应用, 2016. (入选科技部F5000, 本学科前1%高被引论文, 《控制理论与应用》年度优秀论文)

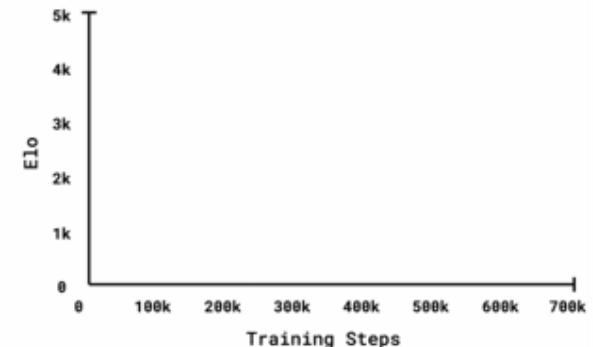
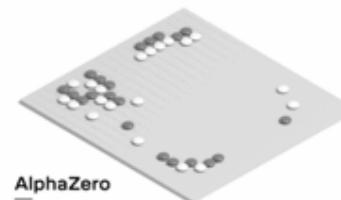
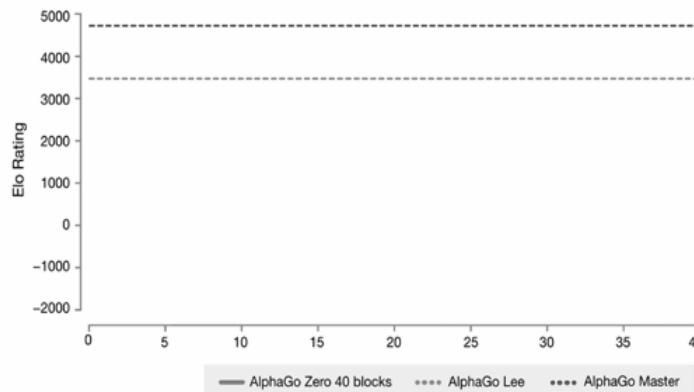


- 2017 年 5 月, AlphaGo 的升级级 AlphaGo Master 以 3-0 战胜 Ke Jie(当前世界排名第一)

AlphaGo Zero/AlphaZero

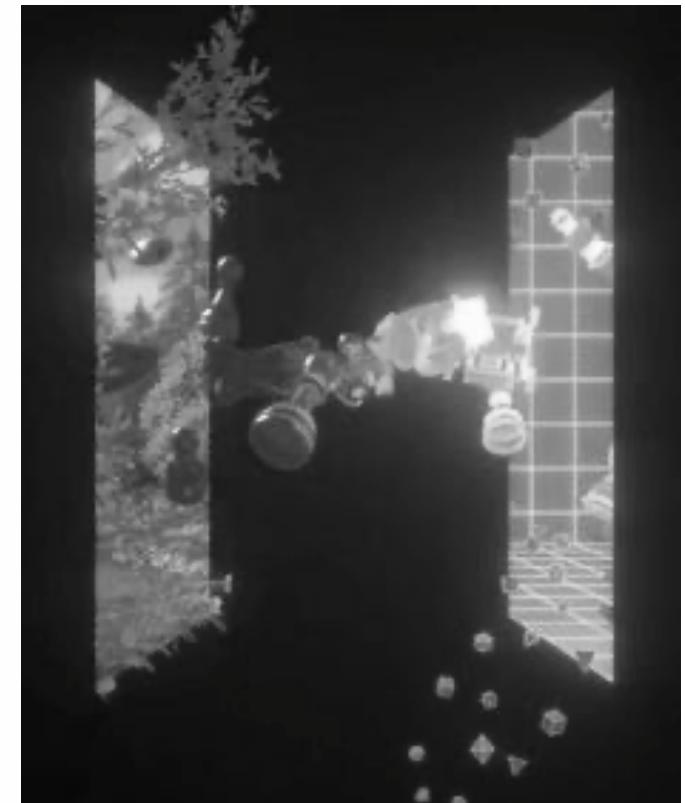
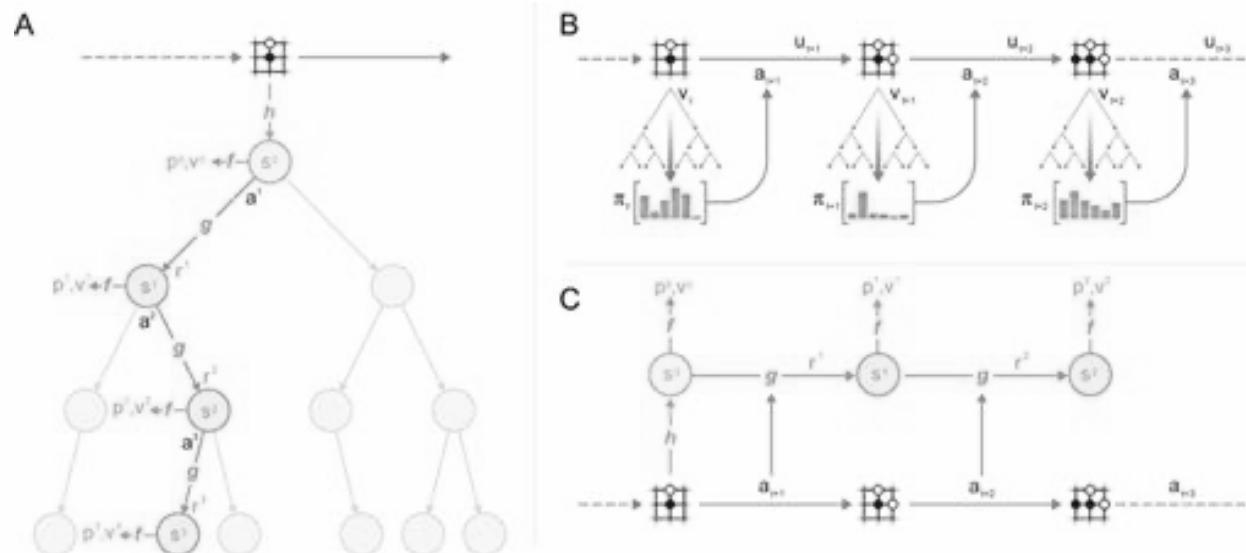


- Deepmind AlphaGo Zero masters Go without human knowledge (*Nature*, 2017)
- AlphaZero masters chess, shogi, and Go through self-play (*Science*, 2018)

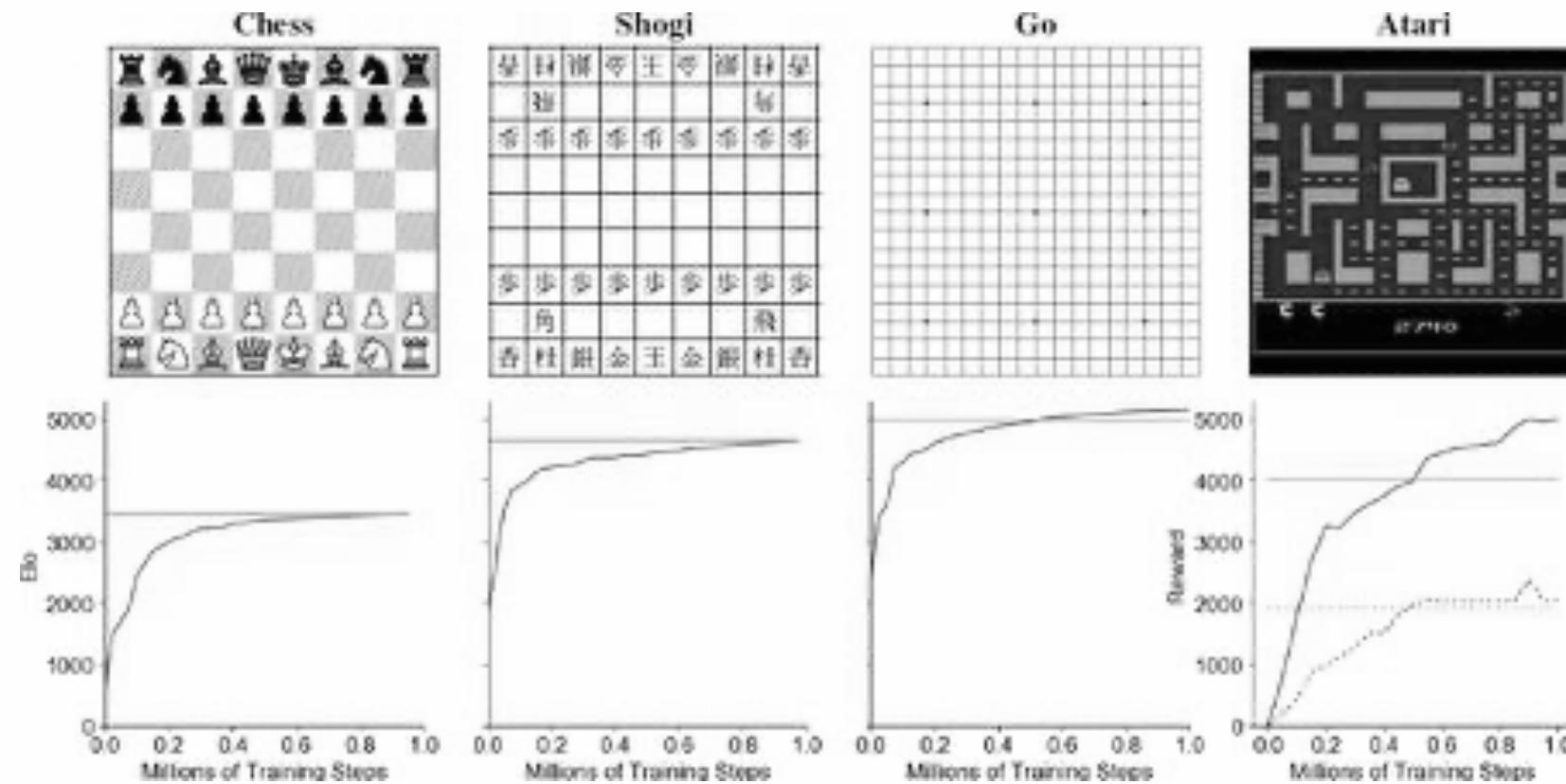


唐, 邵, 赵, 朱, 深度强化学习进展—从AlphaGo到AlphaGo Zero, 控制理论与应用, 2017(下载5000余次, 年度下载第一)

- A new approach to model-based RL that achieves state-of-the-art performance in Atari 2600, a visually complex set of domains, while maintaining superhuman performance in precision planning tasks such as chess, shogi and Go, 2020 *Nature*

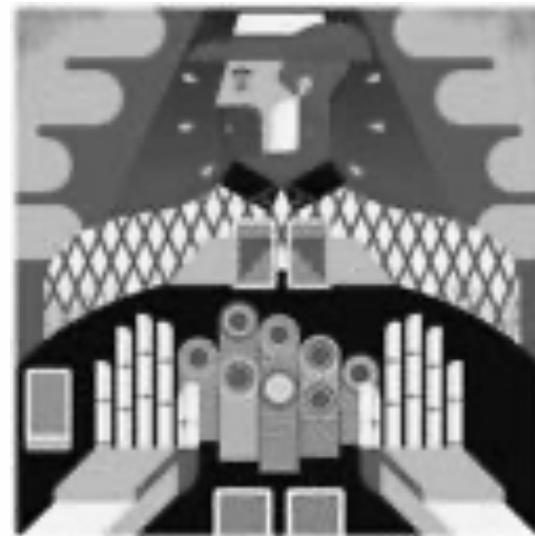
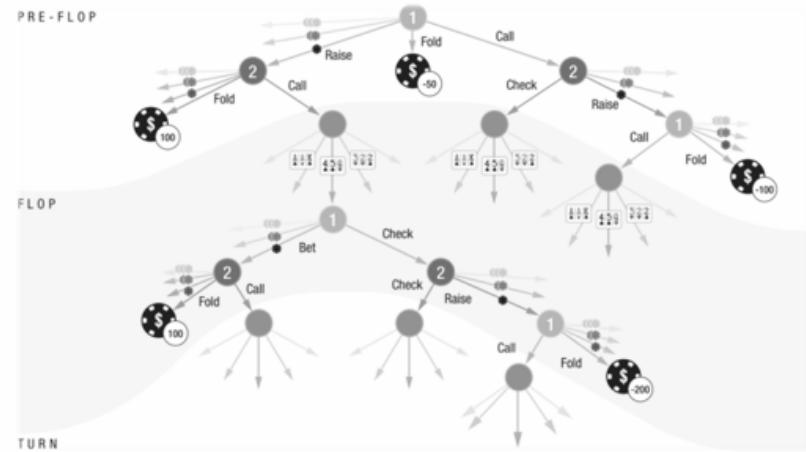


- A new approach to model-based RL that achieves state-of-the-art performance in Atari 2600, a visually complex set of domains, while maintaining superhuman performance in precision planning tasks such as chess, shogi and Go, 2020 *Nature*



- University of Alberta, *Science*, 2017, DeepStack: Expert-level artificial intelligence in heads-up no-limit poker
- 第一个在一一对一代注德州扑克中击败职业扑克玩家的计算机程序

不完全信息博弈



德州扑克 1v5



德州扑克1v5
Pluribus, CMU,
Science, 2019.07



技术突破：

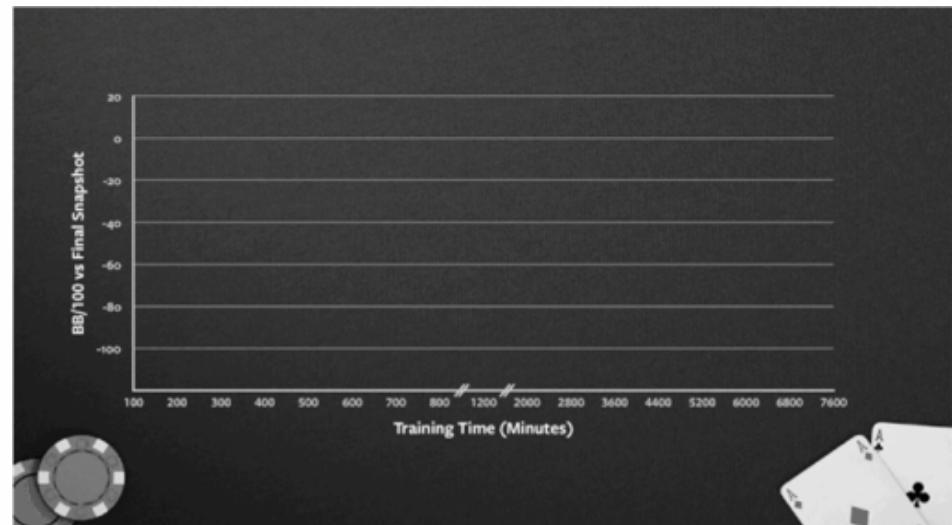
1. 率先在多人制德州扑克战胜职业选手
2. 较低成本的计算资源（64核CPU服务器）
3. 解决不完全信息下的多人零和博弈问题

技术路线：

1. 从“零”开始的多人自我博弈训练
2. 使用MCCFR算法，构造基于自我博弈下的“蓝图策略”集
3. 满足实时性需求，设计有限深度搜索
4. 使用线性CFR算法，将“蓝图策略”集应用到实时搜索过程求解最优应对策略



Pluribus 比赛环境



Pluribus 训练过程

麻将AI

Suphx, 10段, 微软, 2019.08

技术路线:

- 1. 初始化:** 用专家数据（天凤平台）做监督学习，得到初始模型；
- 2. 强化学习:** 用自我博弈的方式进行；
- 3. 先知教练:** 利用不可见的一些**隐藏信息**来引导AI模型的训练方向，倒逼AI模型更加深入地理解可见信息，从中找到有效的决策依据。
- 4. 全盘预测:** 将终盘的**奖励信号**分配回每一轮中，掌握大局观的高级技巧。
- 5. 探索:** 全新的机制对过程的多样性进行动态调控；根据本轮的**底牌**来动态调整；
- 6. 在线比赛:** 通过不断与人类玩家的对局中，得到自我更新和提高。



爆打
(东京大学/HEROZ)

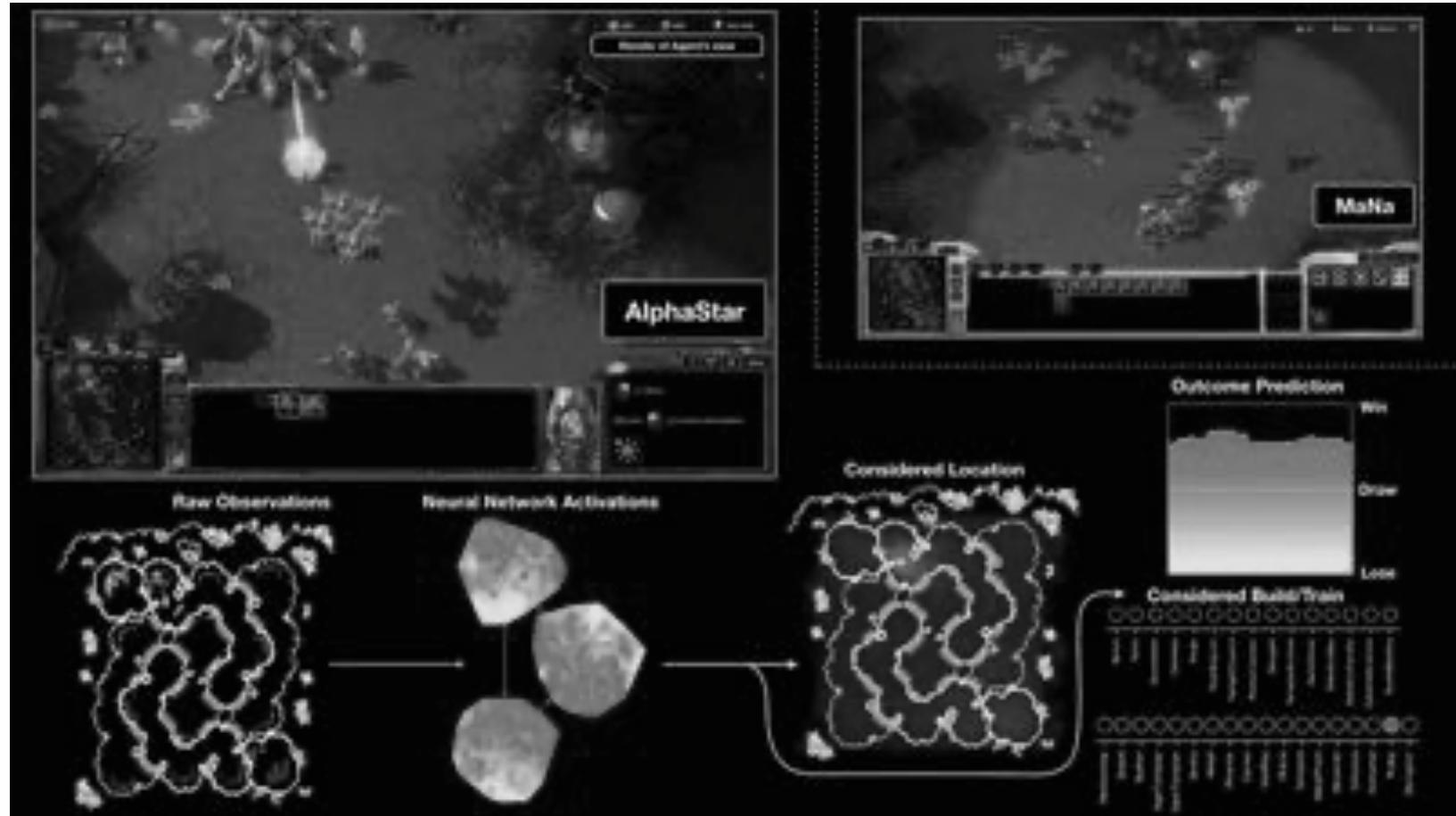
NAGA25
(Dwango公司AI)

顶级人类选手
(10段以上)

Suphx
(微软AI)

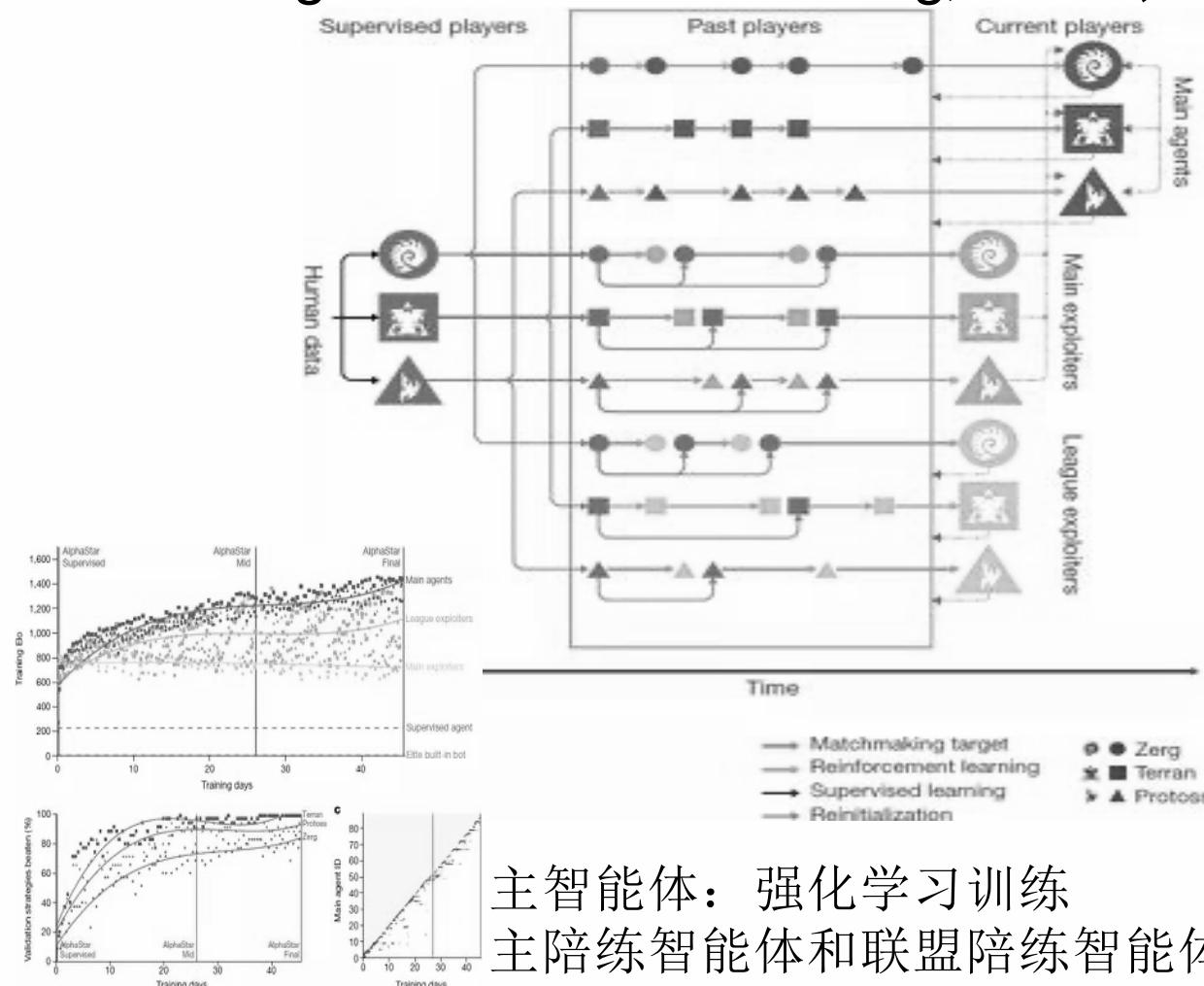
6 6.5 7 7.5 8 8.5 9

- 2019 年 1 月, DeepMind 公布了开发的 AlphaStar 与人类职业选手录像与比赛, 最终 10:1 获胜*



*<https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii>

- DeepMind, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, 2019



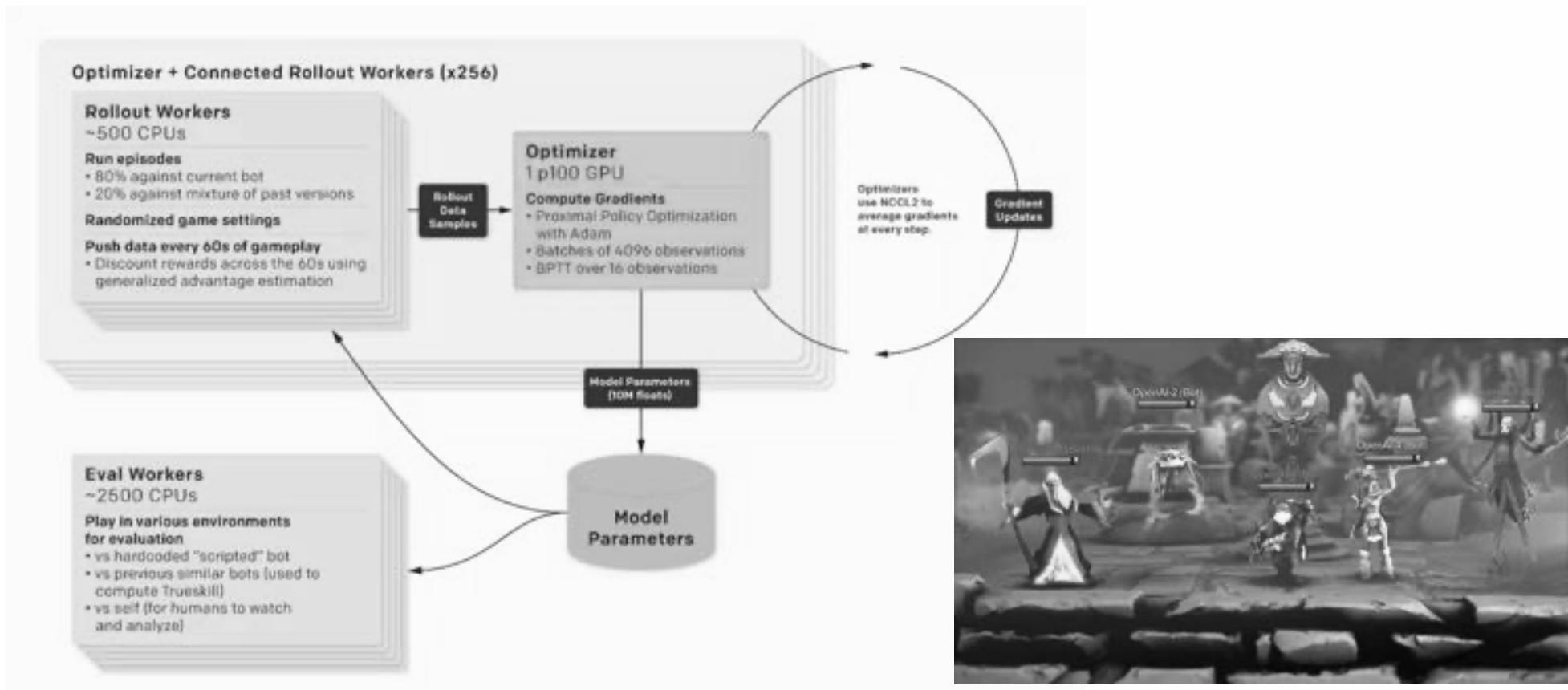
问题

- 没有单一最佳策略
- 非完全信息
- 有蝴蝶效应
- 实时决策
- 巨大动作空间
- 三种不同种族

改进

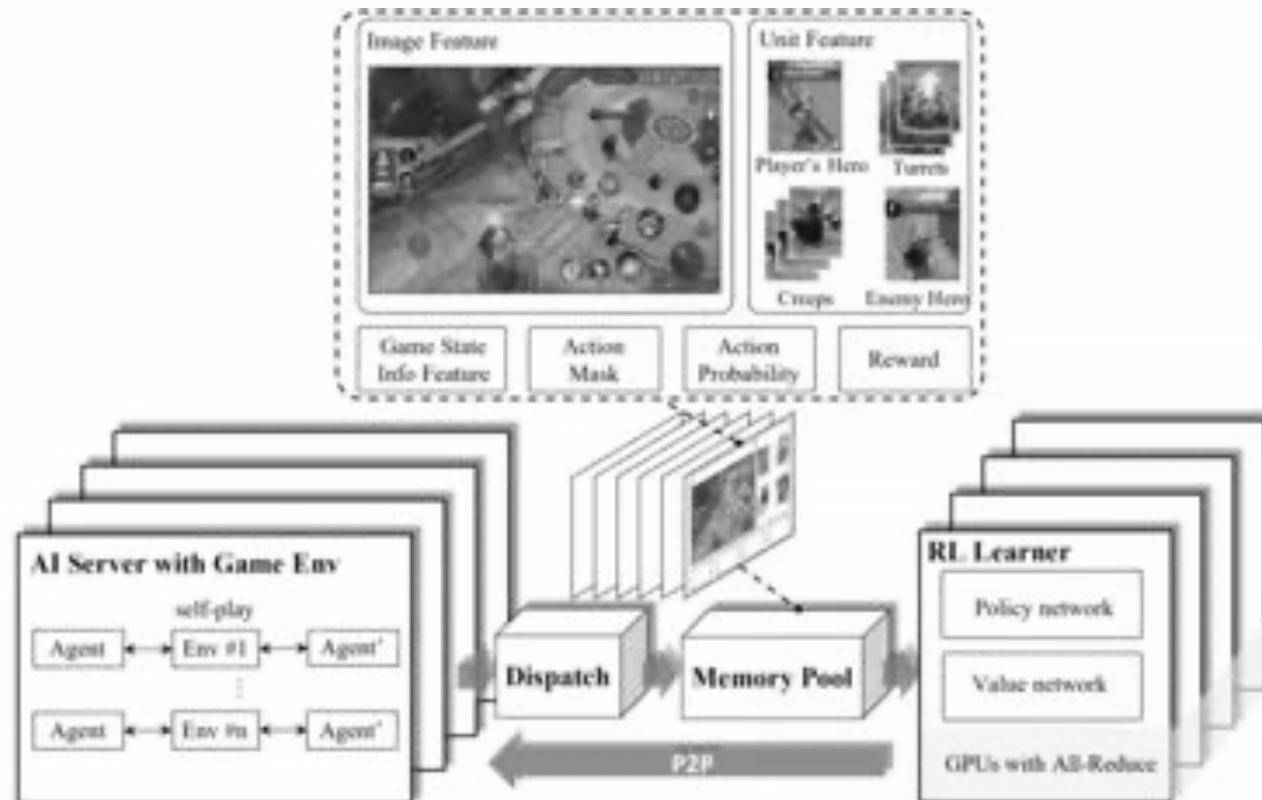
- ✓ 微观操作卓越
- ✓ 地形感知能力强
- ✓ 操作和人类相同
- ✓ 适应三大种族
- ✓ 训练过程全自动化
- ✓ 天梯比赛胜99.8%

- 2019 年 4 月, OpenAI Five 人工智能系统迎战去年 Ti8(第八届 Dota2 国际邀请赛) 冠军 OG 战队, 最终 2:0 获胜
- 近端策略优化(PPO), 公开赛中获得了 99.4% 胜率



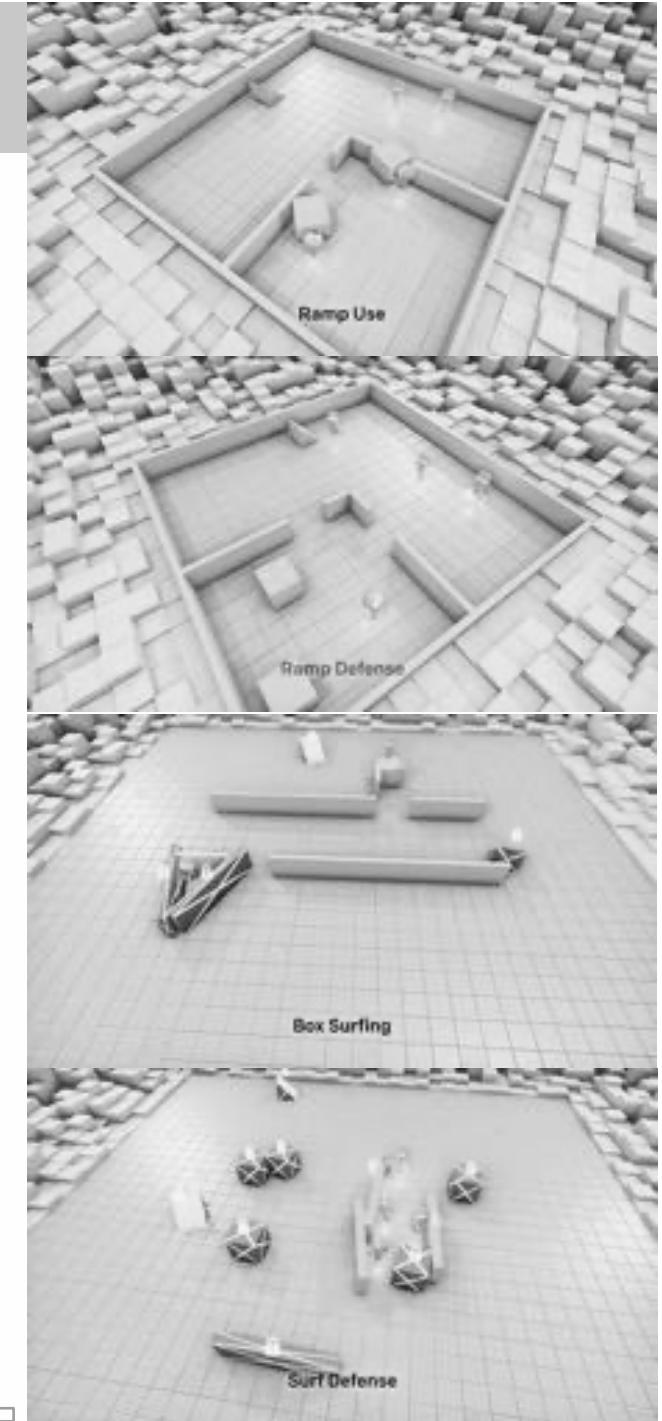
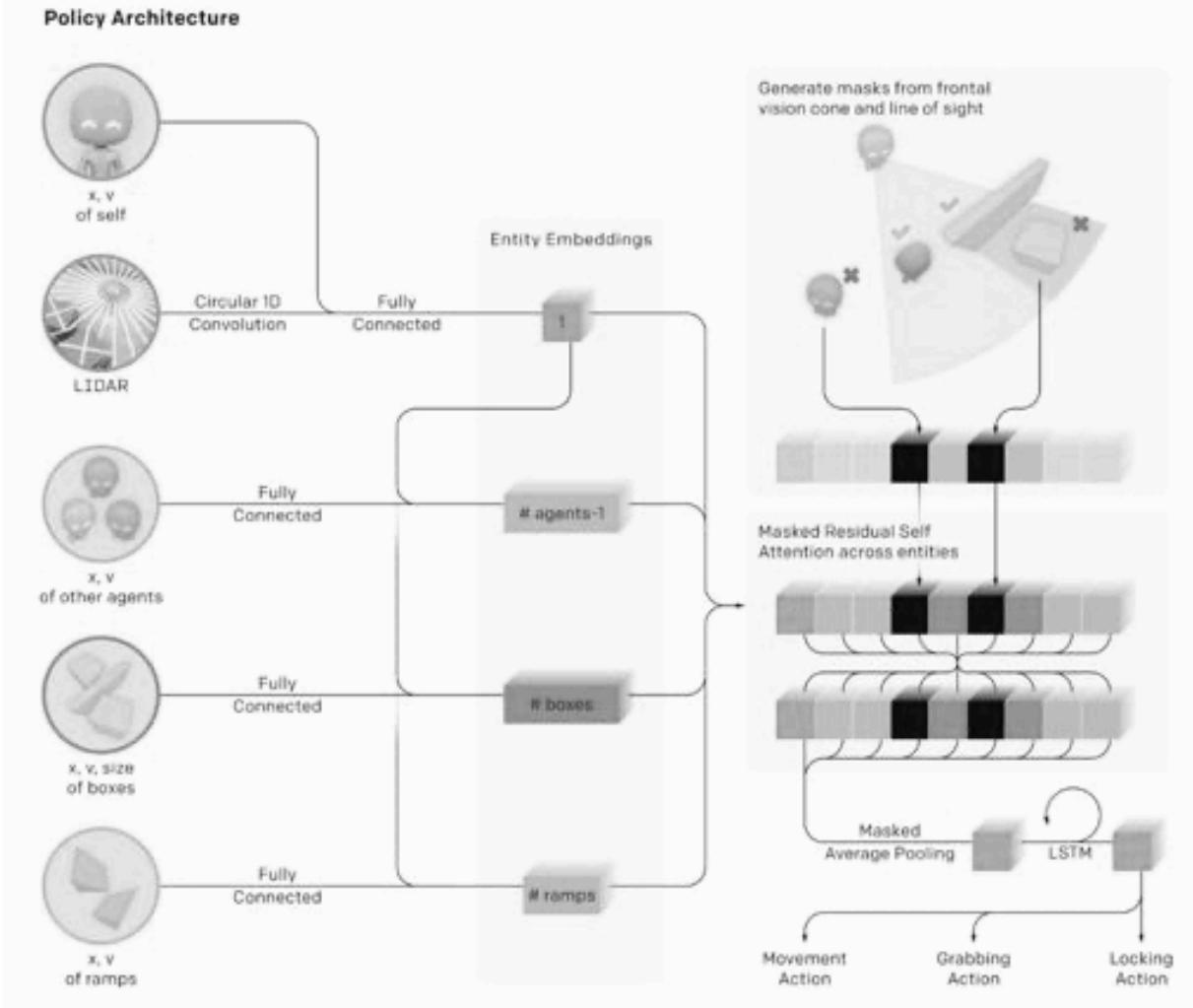
■ 2019年腾讯策略协作型AI「绝悟」升级至王者荣耀电竞职业水平

1. 目标注意力机制: 用于帮助AI在MOBA战斗中选择目标。
2. LSTM: 学习英雄的技能释放组合，在决策中快速输出大量伤害。
3. 动作依赖关系的解耦: 用于构建多标签近端策略优化（PPO）目标。
4. 动作掩码: 基于游戏知识的剪枝方法，引导强化学习的探索。
5. dual-clip PPO: 确保使用大量有偏差的数据批进行训练时的收敛性。



OpenAI--捉迷藏

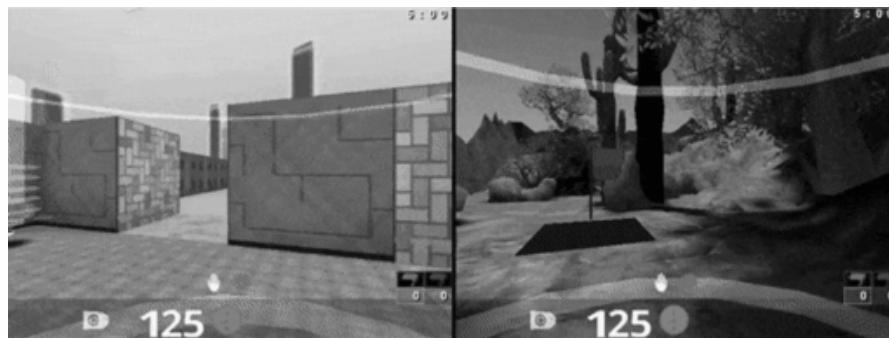
2019, 通过自动课程学习复杂的策略和反策略



三维策略游戏--雷神之锤竞技场



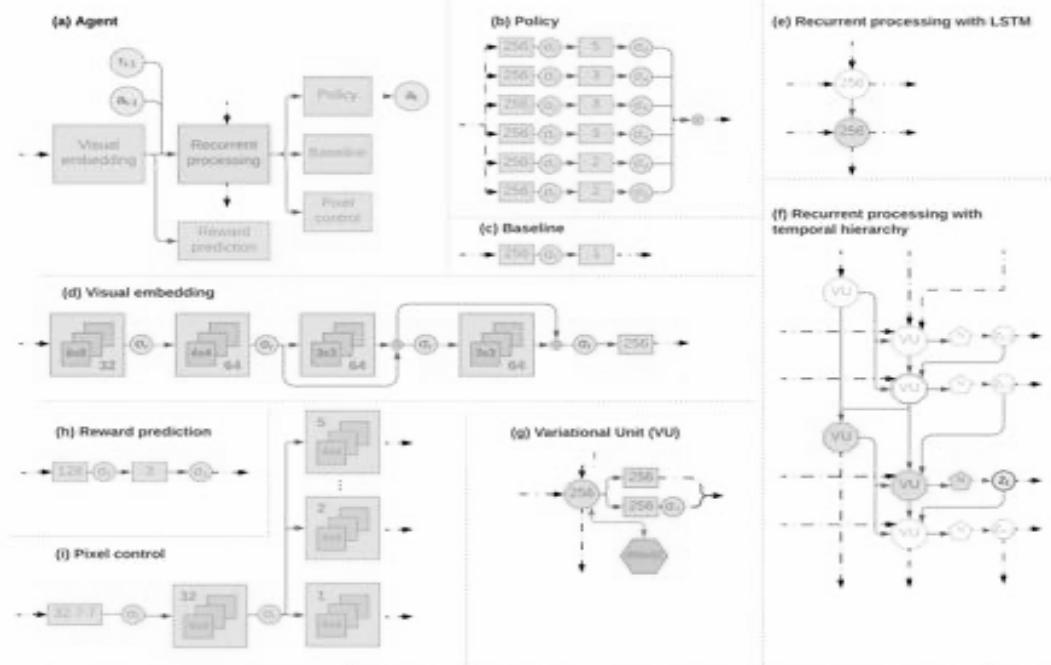
Human-level performance in 3D multiplayer games with population-based reinforcement learning, (2019.05, Science), competed against humans in Capture the Flag of the video game Quake III Arena by Deepmind.



In-door

Out-door

2V2: Compete and cooperate;
3D first person view;
Imperfect information;
Complex environment;
Detailed features.

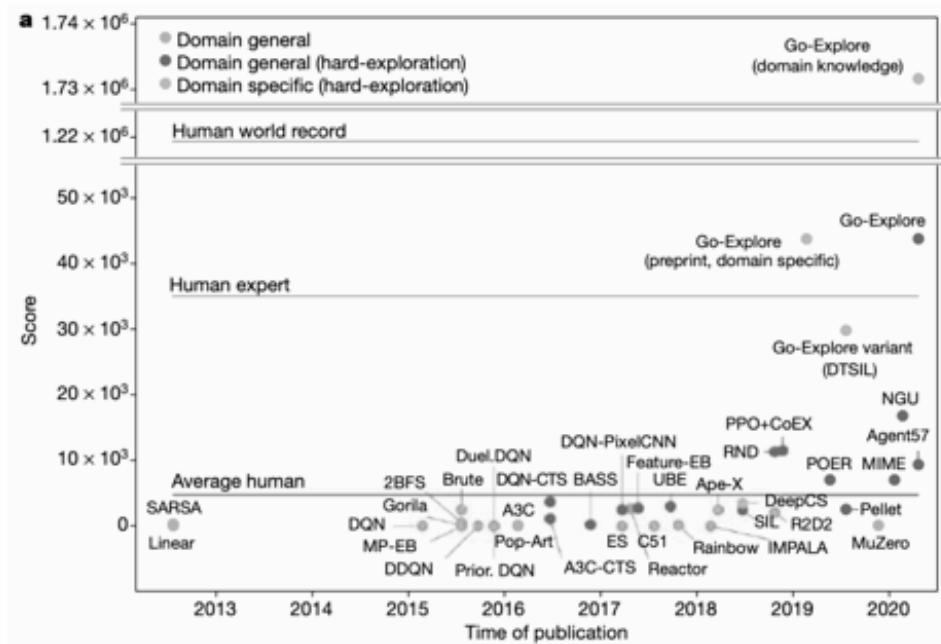
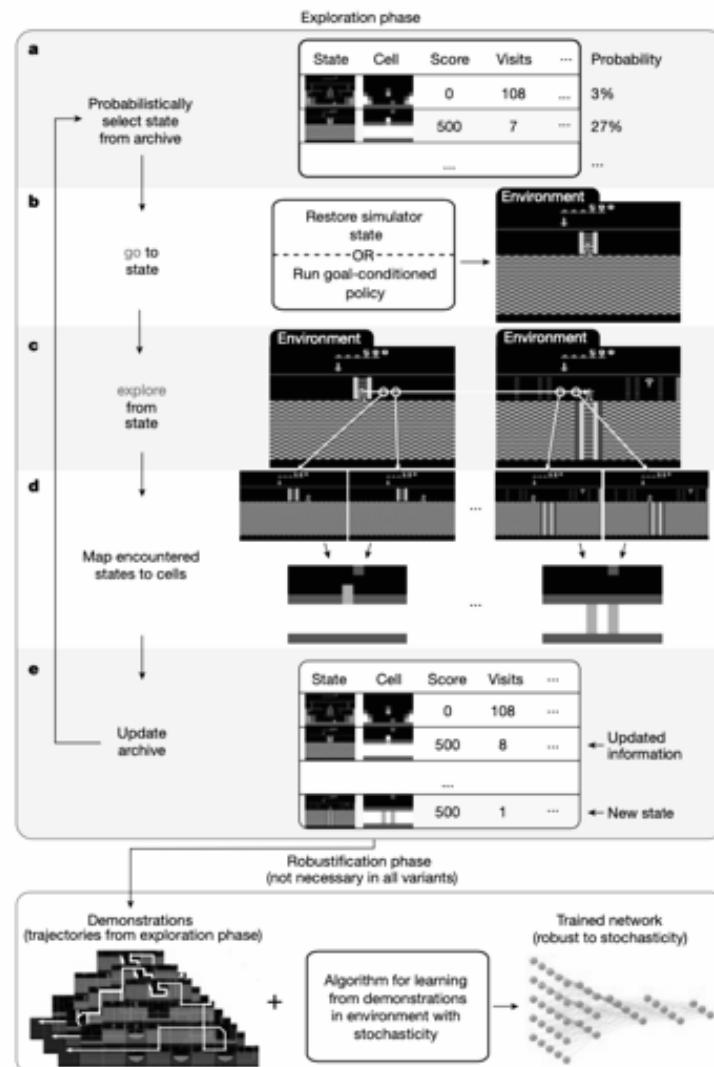
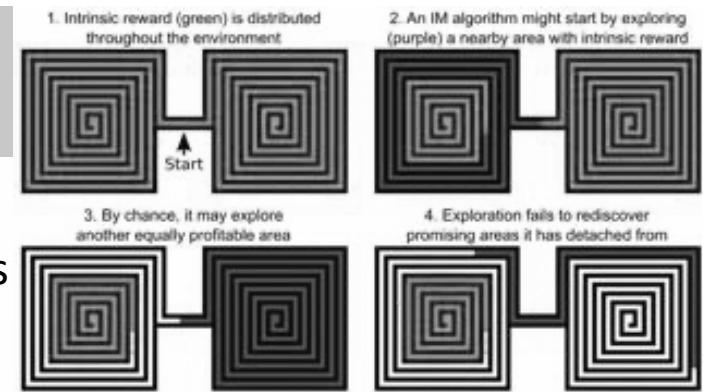


OpenAI&Uber– Go Explore

2021, Nature, the main impediment to effective exploration

Detachment: forgetting how to reach previously visited states

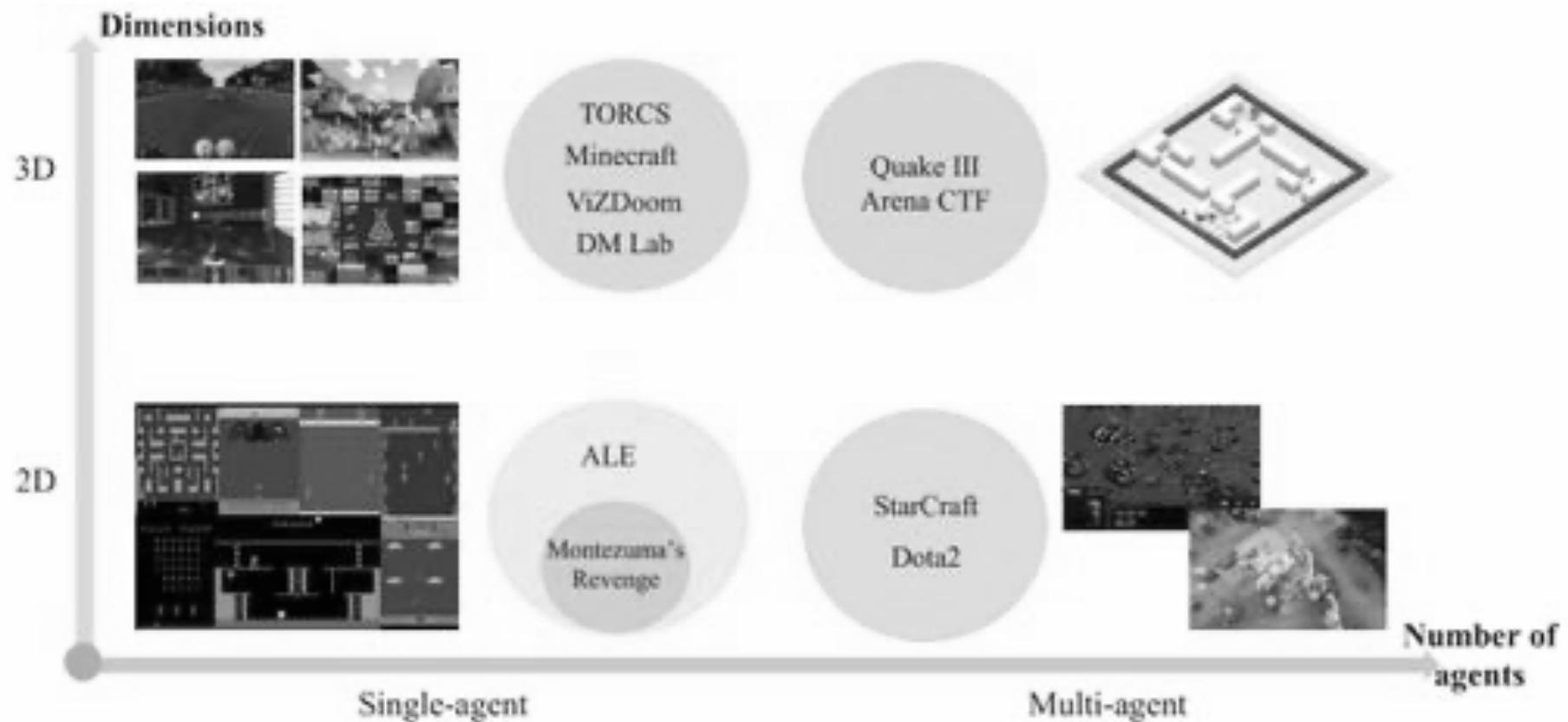
Derailment: failing to first return to a state before exploring from it.



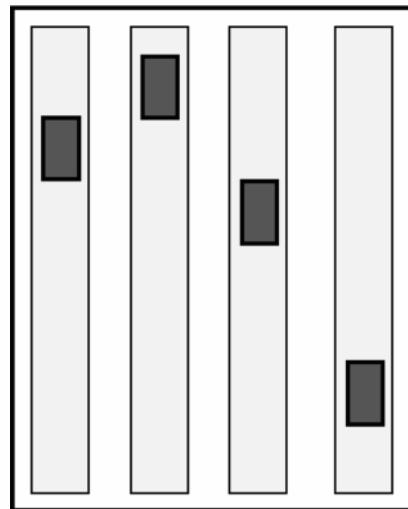
Game	Exploration phase	Robustification phase	State-of-the-art performance	Average human
Berzerk	131,216	197,376	1,383	2,630
Bowling	247	260	69	160
Centipede	613,815	1,422,628	10,166	12,017
Freeway	34	34	34	30
Gravitar	13,385	7,588	3,906	3,351
Montezuma's Revenge	24,758	43,791	11,618	4,753
Pitfall	6,945	6,954	0	6,463
Private Eye	60,529	95,756	26,364	69,571
Skiing	4,242	3,660	10,386	4,336
Solaris	20,306	19,671	3,282	12,326
Venture	3,074	2,281	1,916	1,187

强化学习典型应用

从二维完全信息到三维不完全信息，从单个体到多个体，从仿真到实体



- Crites and Barto*, 1996
 - 10 floors, 4 elevator cars



- STATES: button states, positions, directions, and motion states of cars; passengers in cars & in halls
- ACTIONS: stop at, go by, nextfloor
- REWARDS: roughly, -1 per time step for each person waiting

- conservatively about 10^{22} states
- Q-learning (Watkins, 1989)

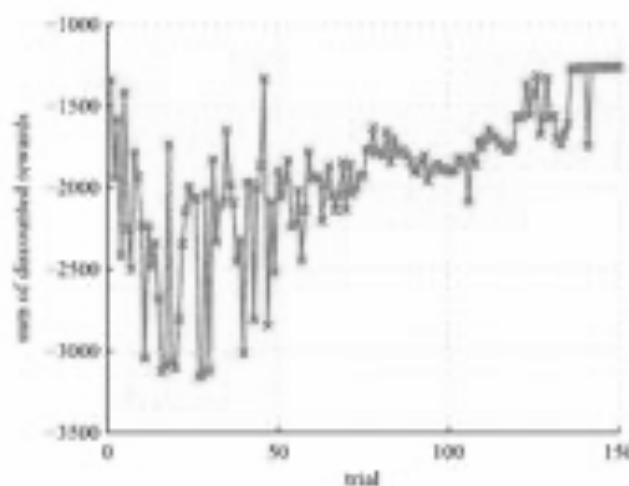
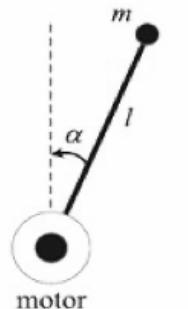
*Crites, R. H., & Barto, A. G. (1996). Improving elevator performance using reinforcement learning. NIPS.

常用的机器人benchmark（第8讲）



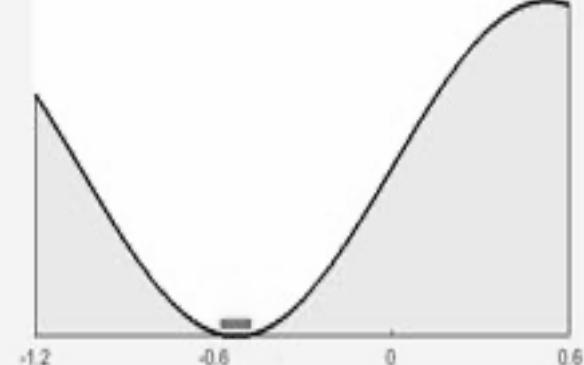
➤倒立摆的优化控制问题及一些常规的RL benchmark算法测试

单倒立摆摆起+稳定

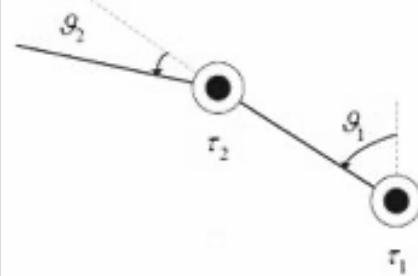


The 1-th Episode

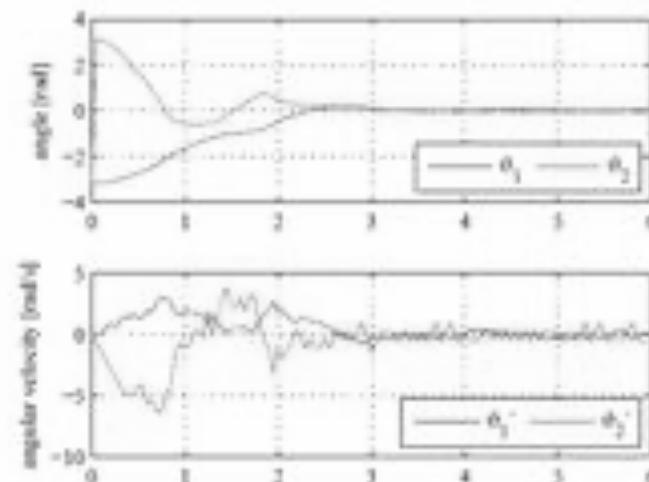
Moutain-car



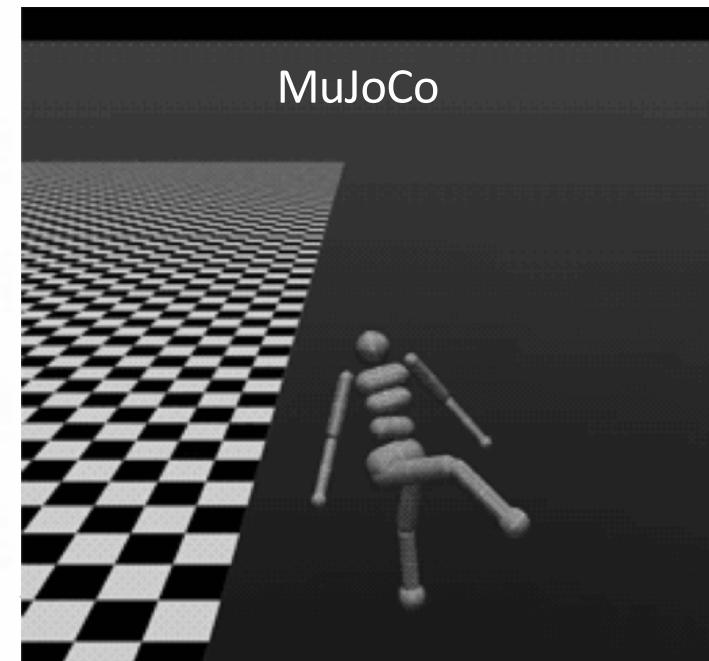
双连杆机械臂



性能变化曲线



MuJoCo



- 强化学习让小狗机器人学会前行 *
- 初始阶段，行走比较吃力，歪扭七八



*Kohl, N., & Stone, P. (2004). Policy gradient reinforcement learning for fast quadrupedal locomotion. ICRA.

■ 强化学习让小狗机器人学会前行 *

- 初始阶段，行走比较吃力，歪扭七八
- 学习中期，走路姿势有效，直线前行



*Kohl, N., & Stone, P. (2004). Policy gradient reinforcement learning for fast quadrupedal locomotion. ICRA.

■ 强化学习让小狗机器人学会前行 *

- 初始阶段，行走比较吃力，歪扭七八
- 学习中期，走路姿势有效，直线前行
- 最终结果，走路姿势更有效，前行更快



*Kohl, N., & Stone, P. (2004). Policy gradient reinforcement learning for fast quadrupedal locomotion. ICRA.

■ 双足机器人行走 (actor-critic + eligibility trace)^{*}



^{*}Tedrake, R., Zhang, T. W., & Seung, H. S. (2005). Learning to walk in 20 minutes.

- 遥控直升机倒立飞行 *

use the Pegasus reinforcement learning algorithm
(a policy-search method)



*Abbeel, P., Coates, A., Quigley, M., & Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. NIPS.

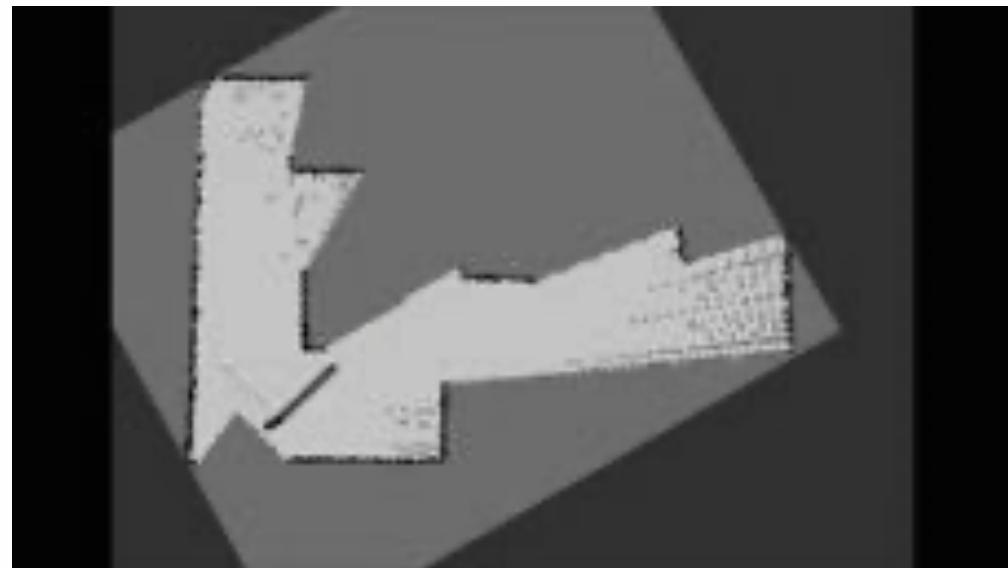
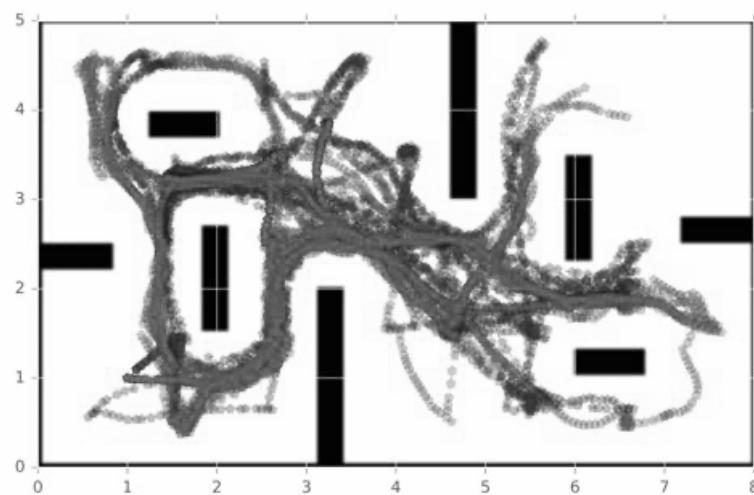
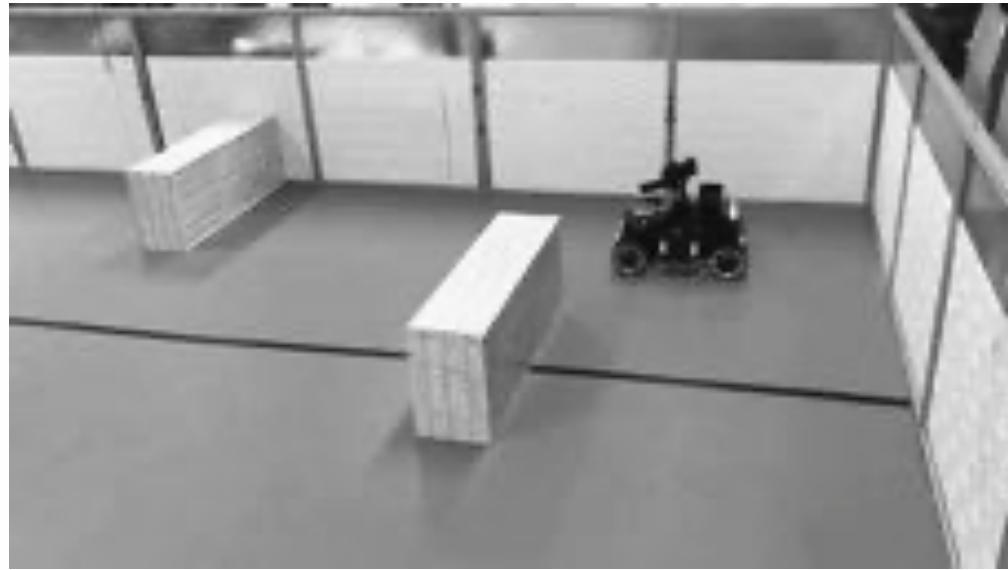
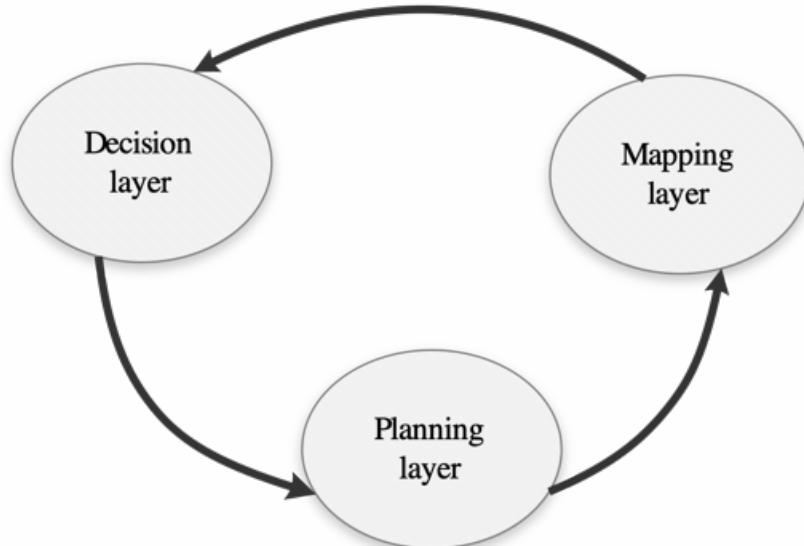
- a Barrett WAM arm uses the mixture of motor primitives (MoMP) algorithm to learn successful hitting movements in table tennis using imitation and reinforcement Learning.*



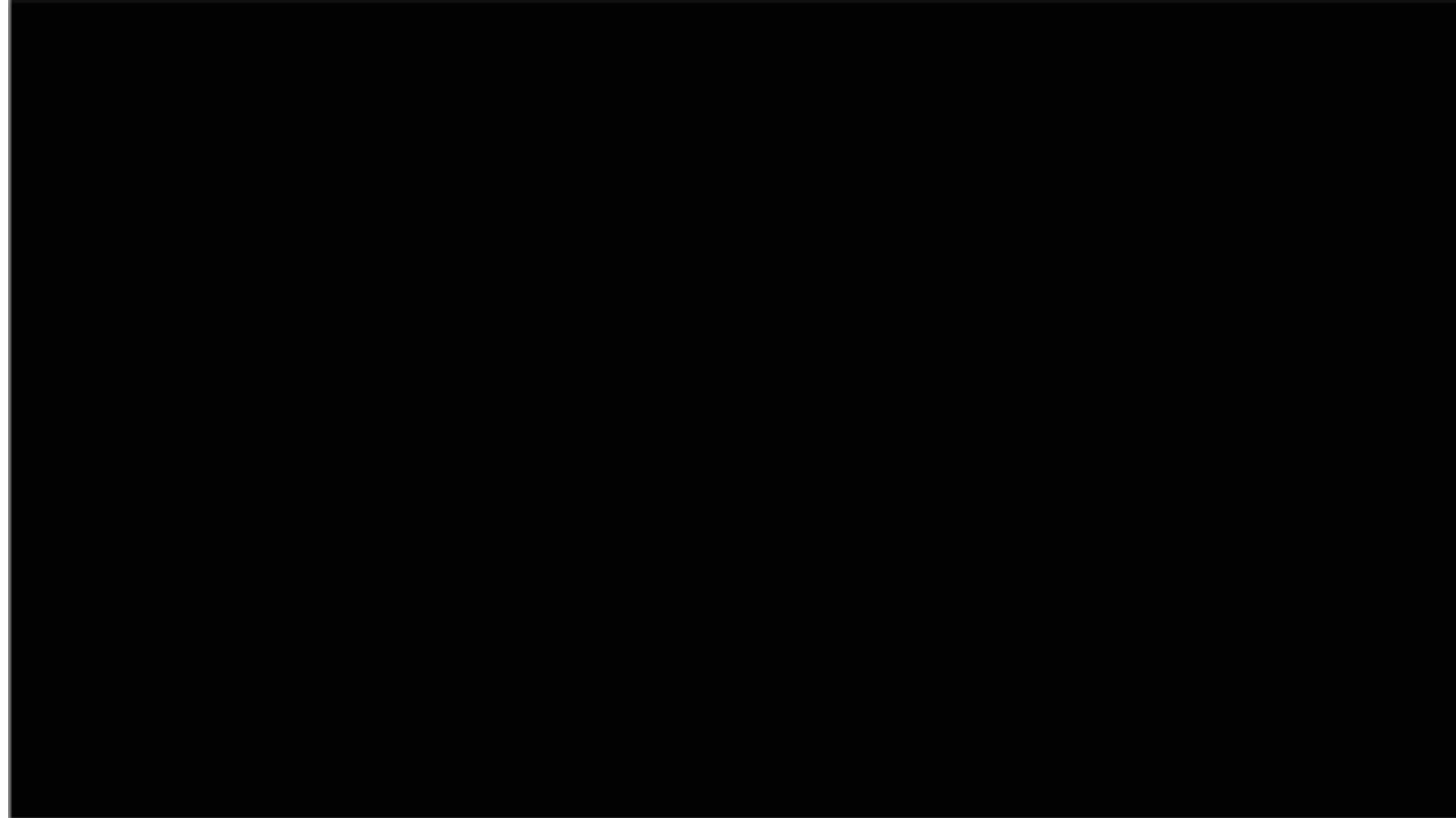
RoboMaster 机器人 — 环境探索



问题：机器人在全新的环境中，通过自主移动构建整个环境地图的过程。

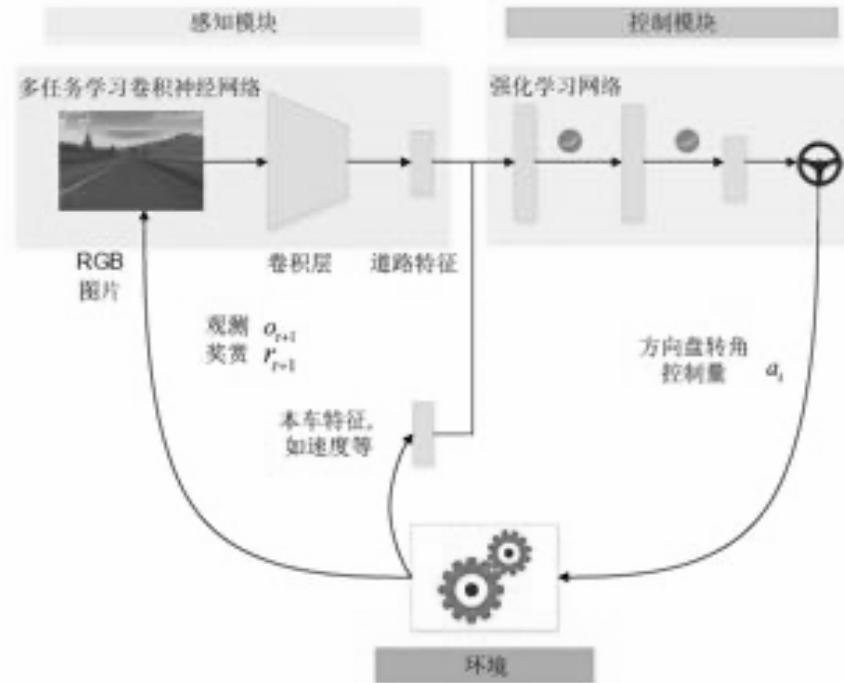


- 剑桥大学创业公司 wayve—The first example of reinforcement learning on-board an autonomous car^{*}

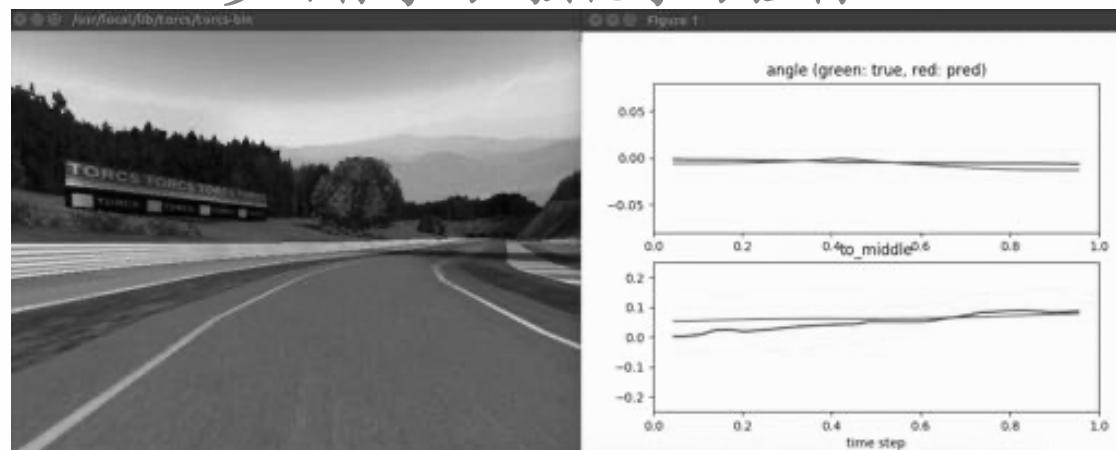


*<https://wayve.ai/blog/learning-to-drive-in-a-day-with-reinforcement-learning>

智能驾驶 - 横纵向控制（第12讲）



多目标学习+强化学习控制



单车道保持

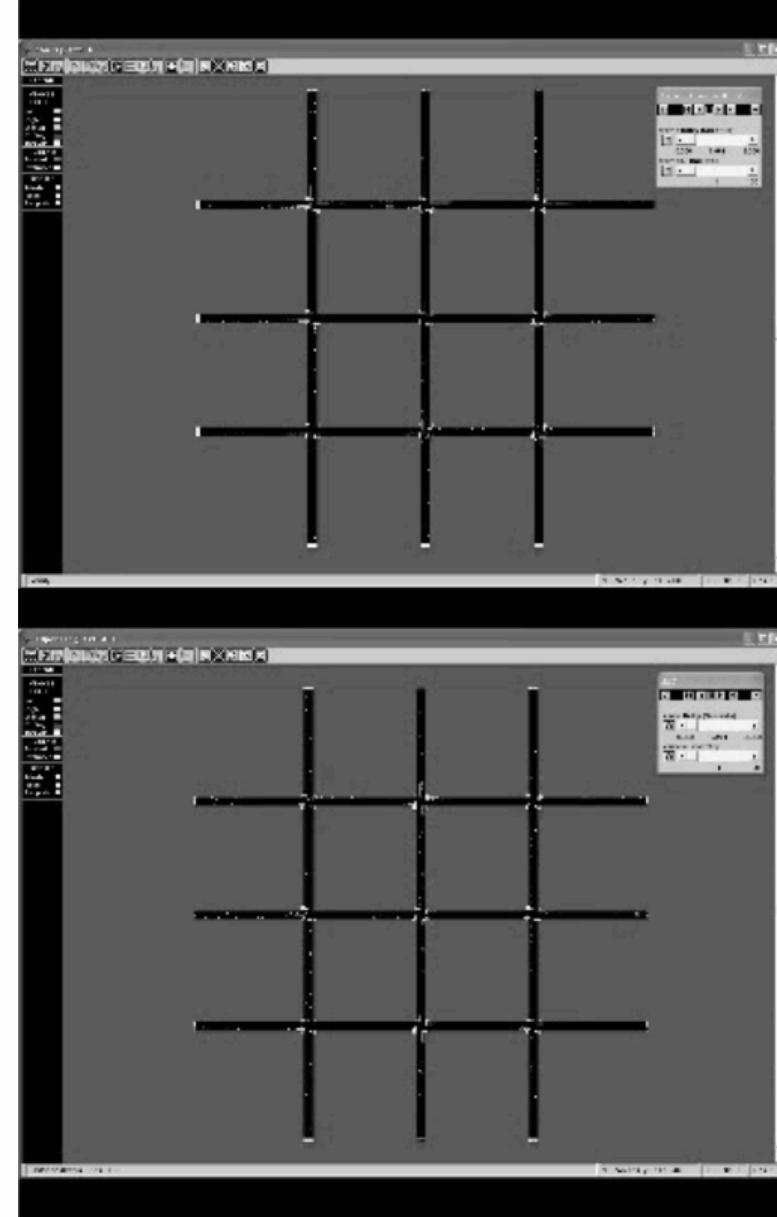
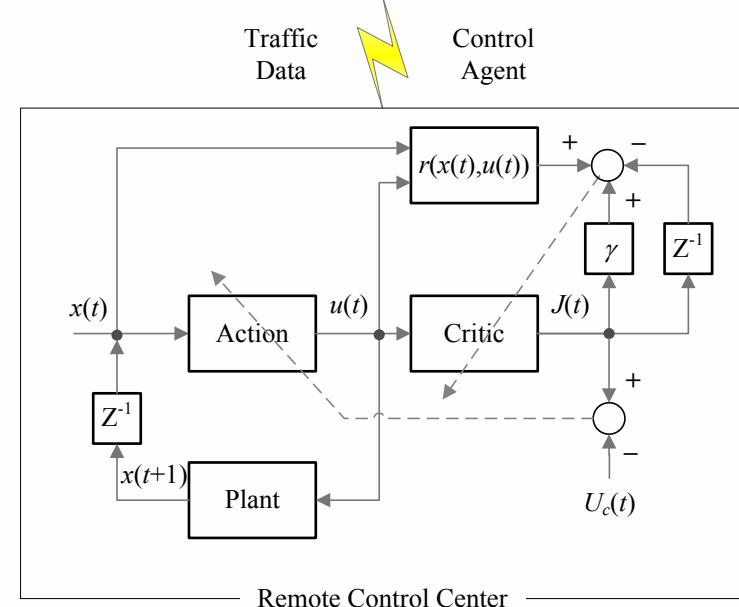
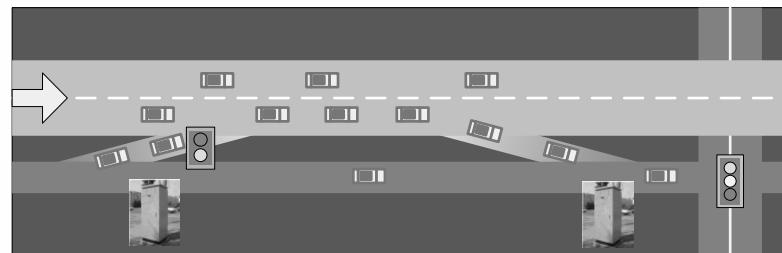


3车道保持



交通信号控制

➤ 实现街道路网、快速路入口匝道交通信号的协调优化控制，有助于减少城市交通拥堵；



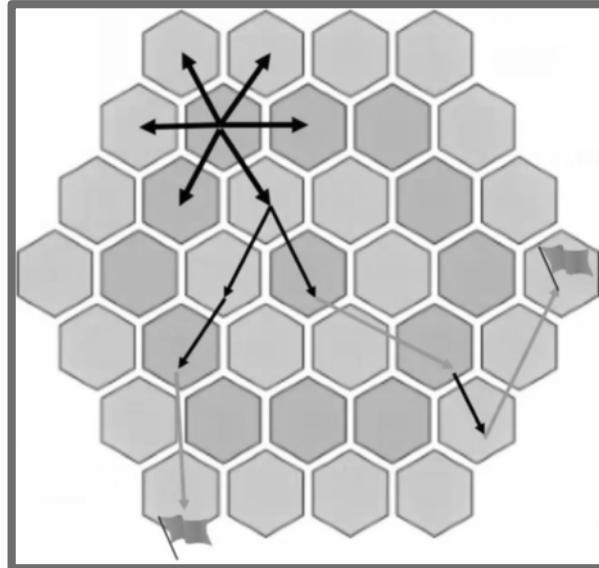
传统
控制
方法

A
D
P
方
法

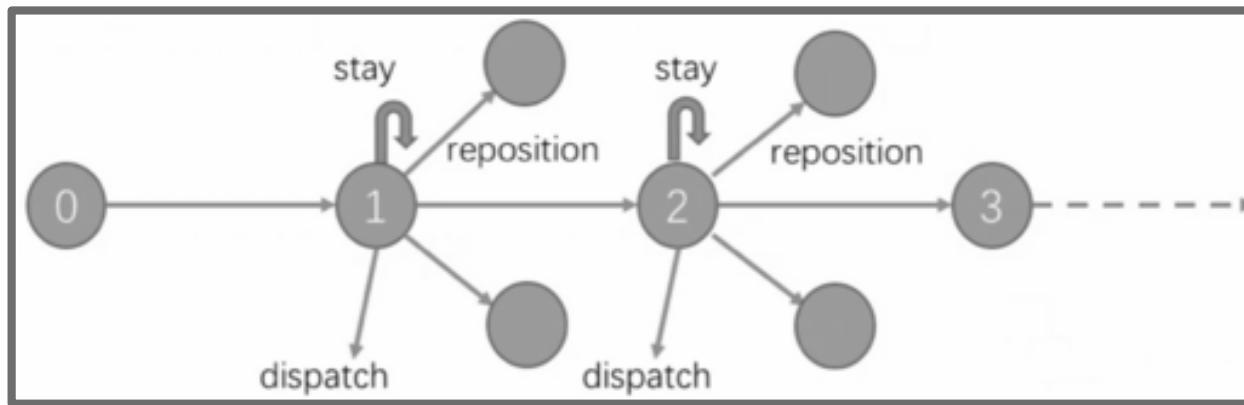
Li, Zhao, Yi. Adaptive dynamic programming for multi-intersections traffic signal intelligent control, ITSC 2008.

智慧城市

➤ 应用需求：派车、调度、物流、供应链、交通管理、智能电网等



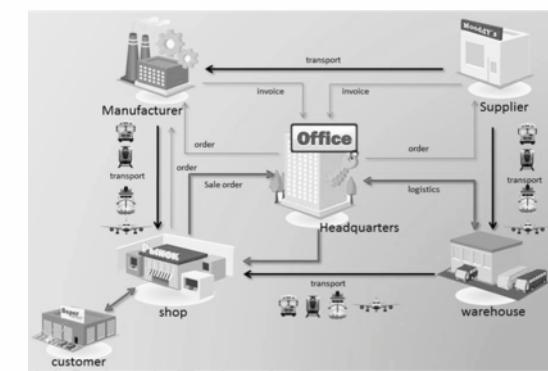
状态空间：六角形格
动作空间：当前格或相邻格的目的地，任意格的目的地（长距离任务），转换到派单任务
环境模型：如下图
特点：供需随机性变化，大规模智能体协调



滴滴城市网约车调度



场景	Well Define Solver	Data Driven	结果
仓内拣选	Batching	Learn to Define Optimization (Embedding)	拣选时间降低10%
Last Mile	DVRP	Can we learn the dispatching rule?	易于接受
包材推荐	Bin Packing	Multi-task selected learning	包材成本降低4.5%到6.6%
AE大脑	MIP	Offline Training + Online Prediction	荷兰109万件；法国、波兰等362万件
智能调度	Heuristic	Adaptive	Improved convergence speed and quality of solutions



• 智能诊断：深度学习的成果丰硕

- ✓ 斯坦福大学，Nature 2017，皮肤癌诊断与专业医生相当
- ✓ 谷歌，2016年，糖尿病的视网膜病变，略高于眼科医生。
- ✓ 北大，2017年，前列腺癌诊断准确率超过90%。
- ✓ 谷歌，Nature Medicine 2018，诊断31种相关的眼部疾病。

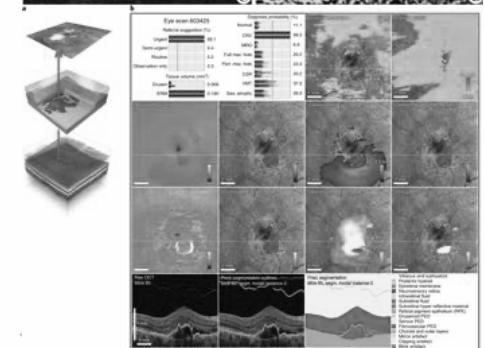
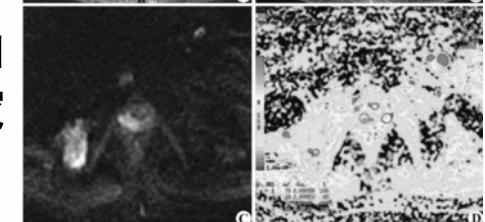
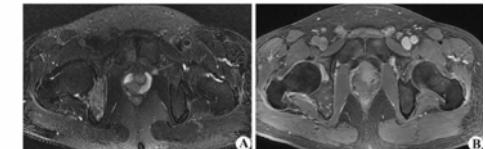
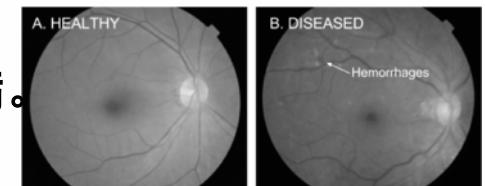
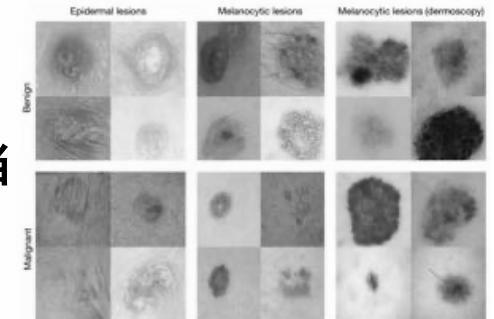
• 智能治疗：强化学习的优化决策

• 药物研发：传统药物开发周期长、耗资高

- ✓ 结合生物模型元素、人工智能、基因组学、蛋白质组学和代谢组学，从大量样本数据中创建病人“图谱”，从中挖掘出可用数据，使药物研发更便宜快捷。

• 机器人手术：精准治疗

-



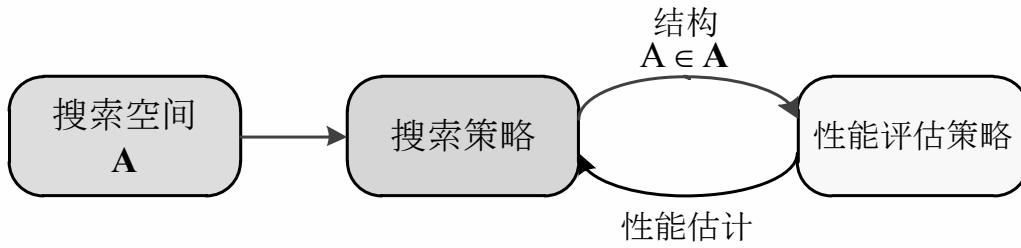
- 量化投资：采用统计、计算机、人工智能等技术，来实现复杂的金融市场的量化决策，选择合适的投资目标，提升投资成效。

- 纪律性；系统性；及时性。
- 美国的公募基金市场里排名第一第二的都是在做量化的基金。
- 2017年10月18日，推出了全球第一只应用人工智能、机器学习进行投资的ETF。

- 欺诈检测(Fraud Detection)
- 风险管理(Risk Management)
- 期权定价(Option Pricing)
-

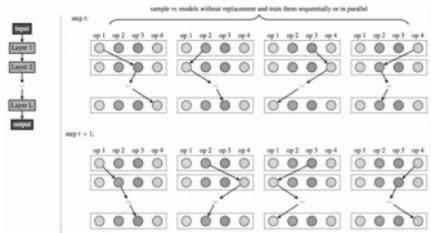


AutoML - 神经架构搜索 (第8/13讲)

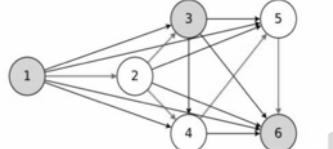


权重加权求和
梯度求解
1块卡1天

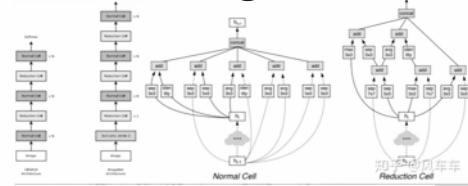
训练超网络



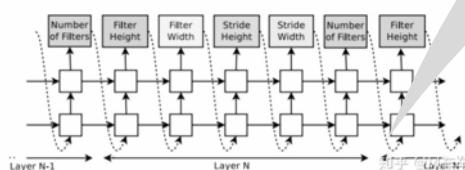
Weights sharing
1块卡0.5天



500块GPU跑了4天
Cifar10/ImageNet



800块GPU跑了快一个月



NAS
Google
ICRL'17

ENAS
Google
ICML'18

DARTS
CMU+Google
ICLR'19

One-shot

旷世、小米等

强化学习基本元素

- 状态：描述当前智能体位置、姿态等信息的变量
- 状态集 **State space**：智能体所有可能状态的集合 S
- 离散状态集：状态与状态之间是独立的
 - discrete, finite, but probably a large set
 - 电梯所在楼层，围棋棋盘
- 连续状态空间：状态与状态之间是连续变化
 - continuous, infinite
 - 车辆的位置，速度，加速度

- 动作：智能体能够执行，改变当前状态的变量
- 动作集 **Action space**：智能体所有可行的动作集合 A
- 离散动作集
 - 电梯的按钮，围棋下一步的落子位置
- 连续动作空间
 - 车辆油门/刹车踏板的深浅，方向盘转角

- 策略：状态空间到动作空间的映射

$$\pi: S \rightarrow A$$

- 代表了智能体是如何行为的
- 确定策略 deterministic:

$$a_t = \pi(s_t)$$

- 随机策略 stochastic:

$$a_t \sim \pi(s_t)$$
$$\pi(a_t|s_t) = P(a_t|s_t)$$

举例：

- 电脑游戏中 NPC (Non-Player Character) 的策略
 - 基于脚本/规则树，每次行为都一样，完全没有变化
- 石头 - 剪刀 - 布的策略
 - 石头： $1/3$ 概率，剪刀： $1/3$ 概率，布： $1/3$ 概率
 - 确定性的策略容易被对方利用 (exploitable)

- 也称为环境/模型
- 描述智能体在给定动作下状态的变化
- 离散时间 : $(s_t, a_t) \rightarrow s_{t+1}$
 - 确定型: $s_{t+1} = f(s_t, a_t)$
由 s_t 和 a_t 唯一决定
如 围棋每一步后棋盘的变化
 - 随机型: $s_{t+1} \sim P(s_t, a_t)$
满足一个和 s_t, a_t 相关的概率分布
如 减肥者饮食的控制对体重的变化
如 噪声干扰
- 连续时间: $\dot{x}(t) = f(x(t), u(t))$

- 奖励：环境（算法）对智能体当前的状态/动作好坏程度的反馈

- 奖励是一个标量的反馈信号
- 智能体的任务就是要最大化累加奖励

$$r_{t+1} = R(s_t, a_t)$$

$$r_{t+1} \sim R(s_t, a_t)$$

- 遥控直升飞机的特技表演
 - $+r$ 跟踪期望轨迹
 - $-r$ 坠机
- 打败围棋世界冠军
 - $+/-r$ 赢/输一场比赛
- 管理股票证券
 - $+r$ 帐户增加财富
- 发电厂调控
 - $+r$ 发电
 - $-r$ 超出安全运行条件
- 控制人型机器人双足行走
 - $+r$ 向前移动
 - $-r$ 摔倒
- 视频游戏上超越人类
 - $+r/-r$ 游戏得分增加/减少

举例：

1 下棋时双方依次落子，最后赢了对手

- 动作：每次的落子
- 奖励：中间阶段 $r = 0$ ，最后一步 $r = +1$

2 给机器人控制信号然后移动到工作区域

- 动作：每个时刻的控制信号
- 奖励： $r = -dist(robot, target)$

3 注射或服用药物让糖尿病人血糖长时间稳定

- 动作：各个疗程阶段的给药种类和给药量
- 奖励： $r = |level_{sugar} - level_{normal}|$

举例：

1 下棋时双方依次落子，最后赢了对手

- 动作：每次的落子
- 奖励：中间阶段 $r = 0$ ，最后一步 $r = +1$

2 给机器人控制信号然后移动到工作区域

- 动作：每个时刻的控制信号
- 奖励： $r = -\text{dist}(\text{robot}, \text{target})$

3 注射或服用药物让糖尿病人血糖长时间稳定

- 动作：各个疗程阶段的给药种类和给药量
- 奖励： $r = |\text{level}_{sugar} - \text{level}_{normal}|$

某一时刻的瞬时奖励不能完全反映最终目标完成的情况，需要考虑未来奖励的变化

- 回报：智能体从某一初始状态出发，在策略下产生的轨迹上的奖励累加和 (sum of rewards)

$$G_t = r_{t+1} + \gamma r_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- 折扣因子 $\gamma \in [0, 1]$ 代表未来的奖励对当前回报的贡献
- k 时刻后的奖励 r 对当前回报的贡献只有 $\gamma^k r$
- 这种定义形式更重视近期的奖励，忽视远期的奖励
 - γ 越接近 0，回报越是“目光短浅”
 - γ 越接近 1，回报越是“目光长远”

目标假设

所有的目标都可以通过 最大化期望累加奖励 实现

- 价值：智能体在当前状态下回报的期望 V

$$V(s_t) = \mathbb{E}[G_t] = \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \dots]$$

$$s_{k+1} \sim P(s_k, a_k), a_k \sim \pi(s_k)$$

- 最优价值 optimal value : 智能体在每个状态下能获得的最高价值

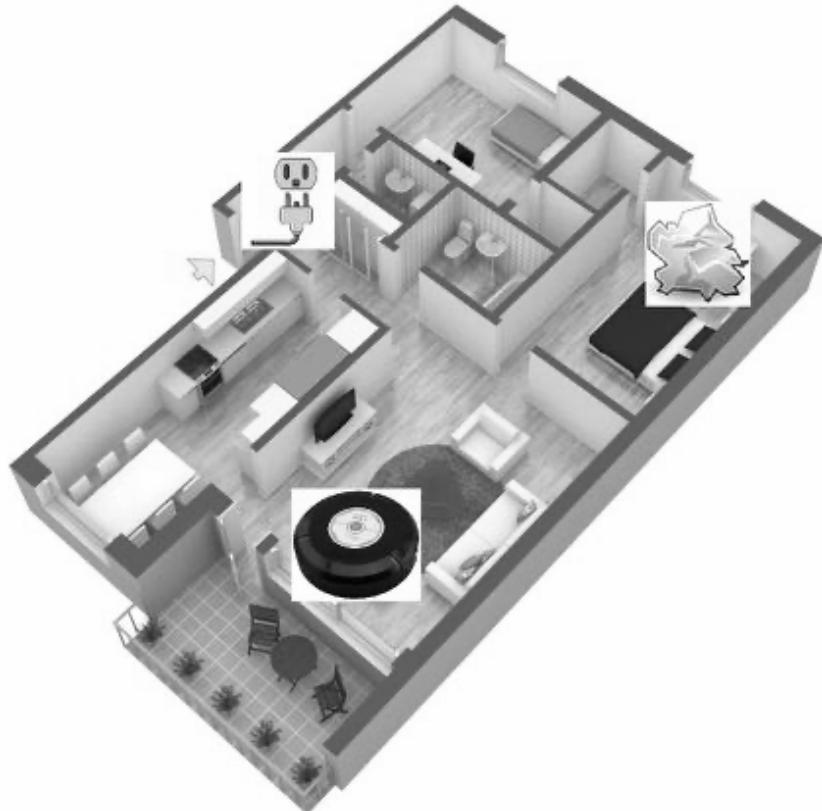
$$V^*(s) = \max V(s) = \max \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \dots]$$

- 最优策略 optimal policy : 能够使智能体获得最高价值的策略 π^*

$$\begin{aligned} V^*(s_t) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid a_k \sim \pi(s_k^*)\right] \\ &\geq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid a_k \sim \pi(s_k)\right], \forall \pi \end{aligned}$$

- 对每个马尔可夫决策问题, 有且只有一个最优价值
- 但最优策略不一定是唯一的

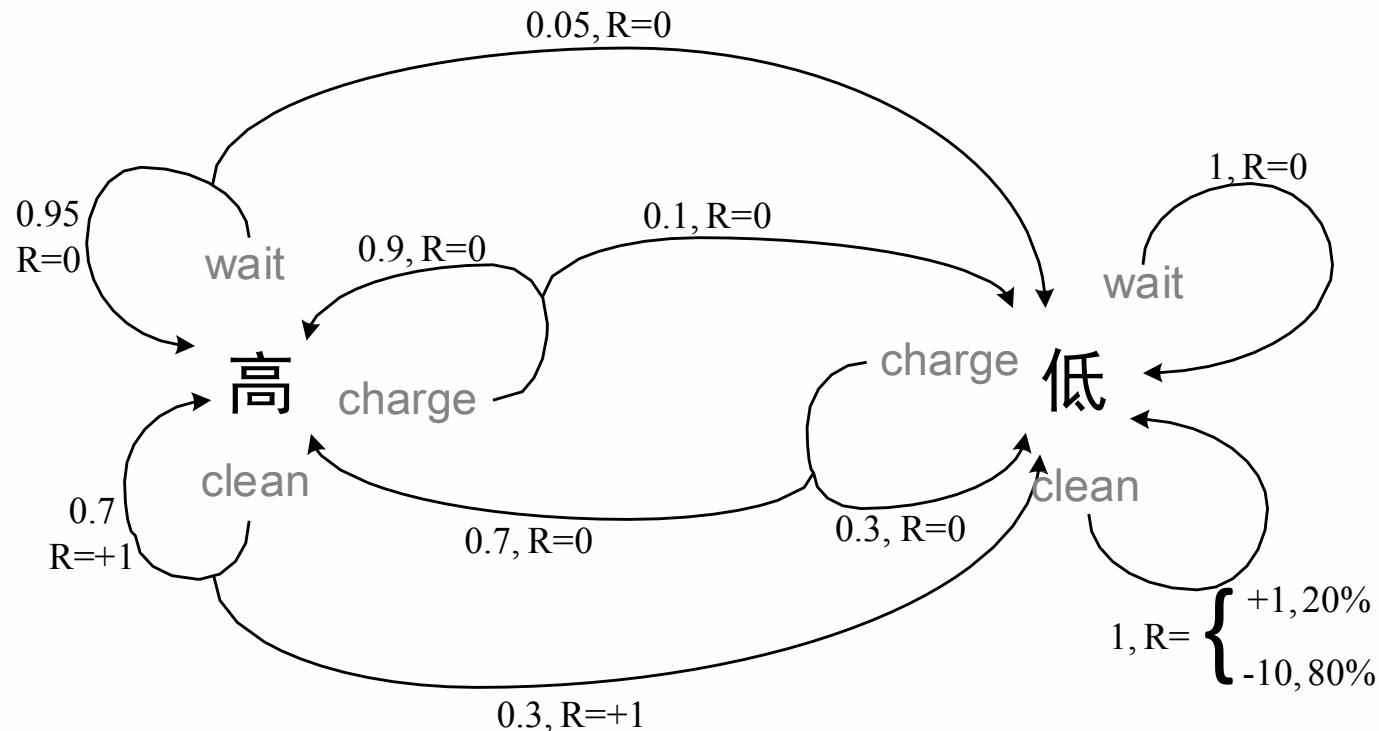
举例：扫地机器人建立强化学习问题



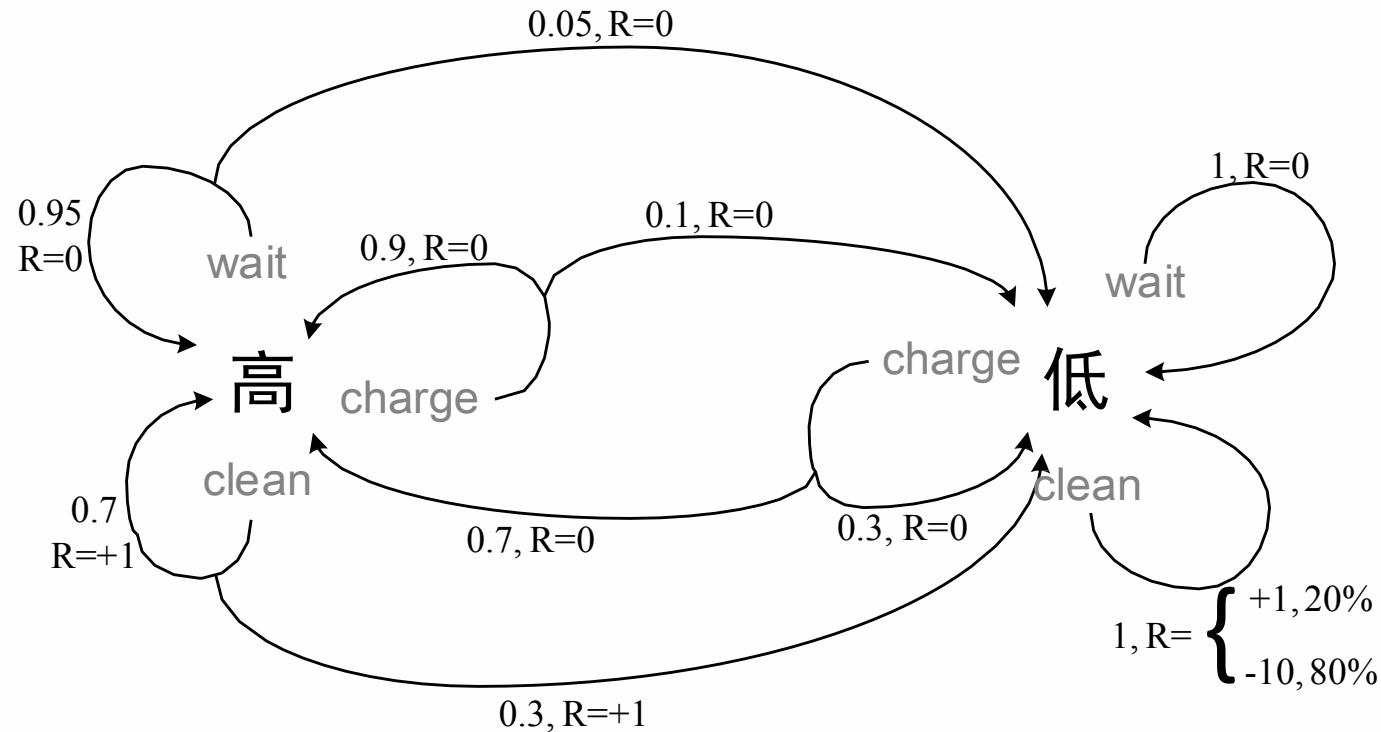
- 扫地机器人任务：保持房间清洁，同时避免耗尽电池关机
- 可能的三种决策：
 - 1 待在原地不动
 - 2 去房间打扫卫生
 - 3 找电源充电
- 决策时考虑的因素：当前电量
- 用户当然希望扫地机器人能经常打扫房间，但如果打扫过程中用尽了电池导致了关机，需要手动把机器人搬回充电，对用户是很差的体验
- 目标：最大化提升使用体验

- 状态: 机器人电量高, 低
- 动作: wait, clean, charge
- 奖励: 扫到垃圾 +1, 停机 -10

- 状态: 机器人电量高, 低
- 动作: wait, clean, charge
- 奖励: 扫到垃圾 +1, 停机 -10



- 状态: 机器人电量高, 低
- 动作: wait, clean, charge
- 奖励: 扫到垃圾 +1, 停机 -10



充电不带来奖励, 却是保证机器人长期运行的重要动作

■ 策略 1: 电量高时 wait, 电量低时 charge (lazy)

- 策略 1: 电量高时 wait, 电量低时 charge (lazy)
- 策略 2: 电量高时 clean, 电量低时 clean (poor experience)

- 策略 1: 电量高时 wait, 电量低时 charge (lazy)
- 策略 2: 电量高时 clean, 电量低时 clean (poor experience)
- 策略 3: 电量高时 clean, 电量低时 charge (work too hard)

- 策略 1: 电量高时 wait, 电量低时 charge (lazy)
- 策略 2: 电量高时 clean, 电量低时 clean (poor experience)
- 策略 3: 电量高时 clean, 电量低时 charge (work too hard)
- 策略 4: 电量高时 50%clean, 50%wait, 电量低时 charge (very good)

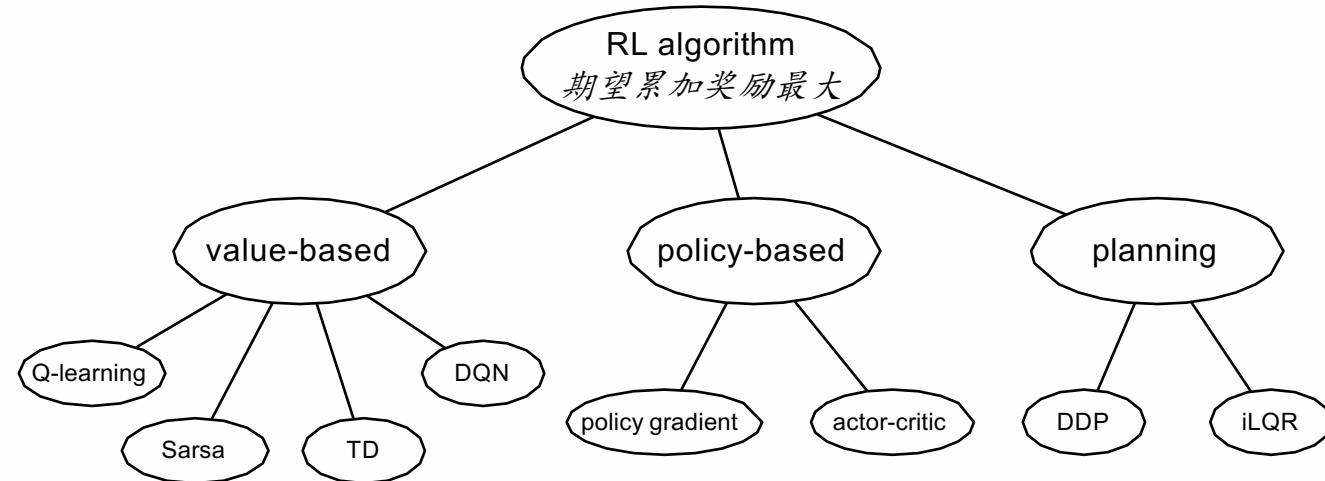
- 策略 1: 电量高时 wait, 电量低时 charge (lazy)
- 策略 2: 电量高时 clean, 电量低时 clean (poor experience)
- 策略 3: 电量高时 clean, 电量低时 charge (work too hard)
- 策略 4: 电量高时 50%clean, 50%wait, 电量低时 charge (very good)
- 回报: 机器人在未来一段时间获得的奖励

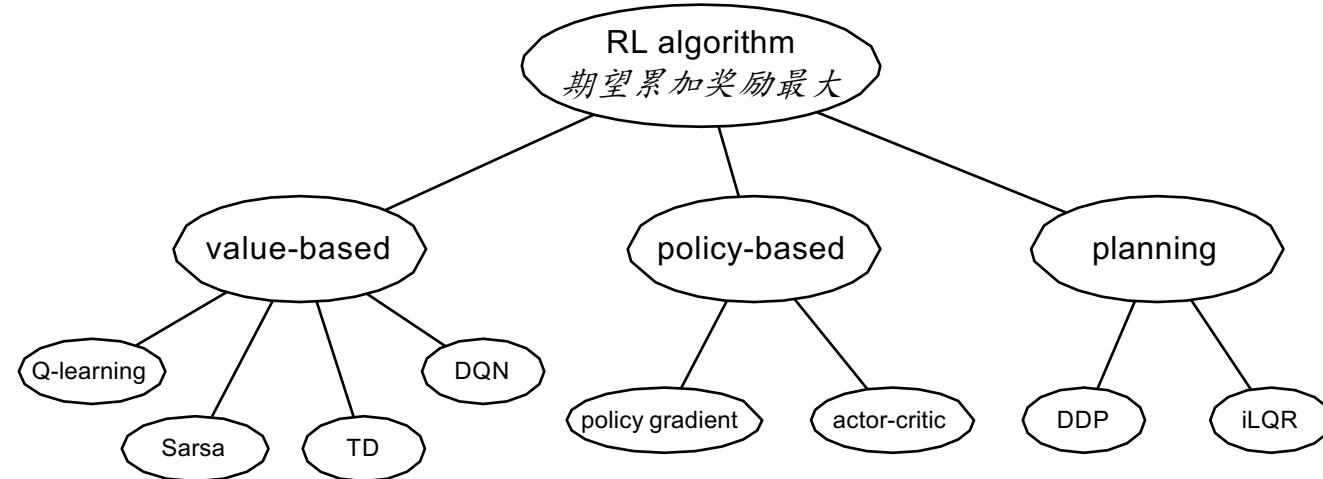
$$r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \dots$$

- 最优策略/最优价值: 能够让机器人在未来一段时间保持最高收益的策略及相应的回报

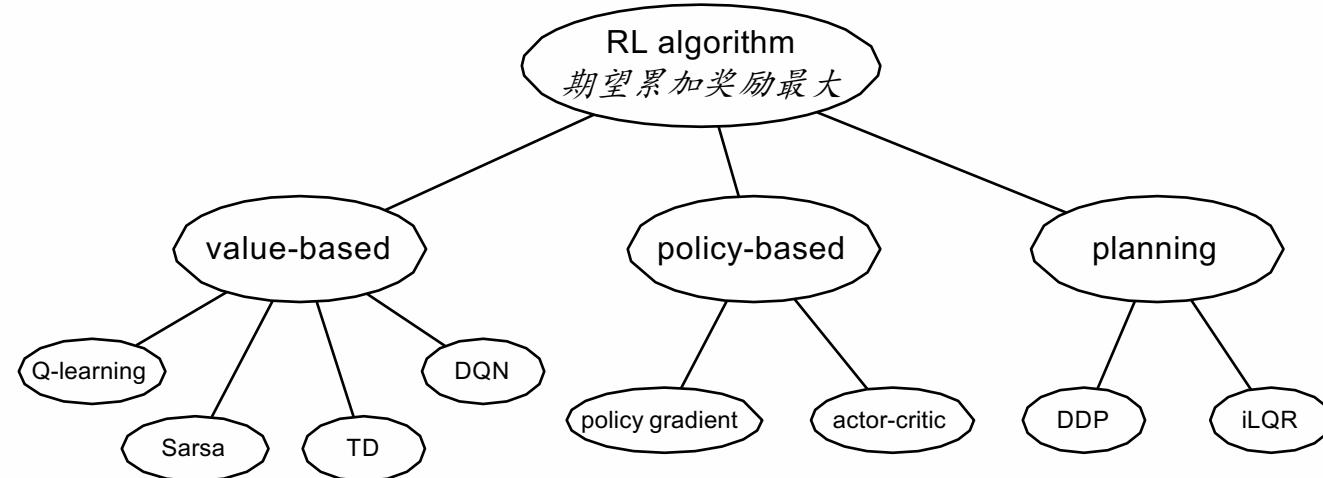
强化学习算法分类

算法分类

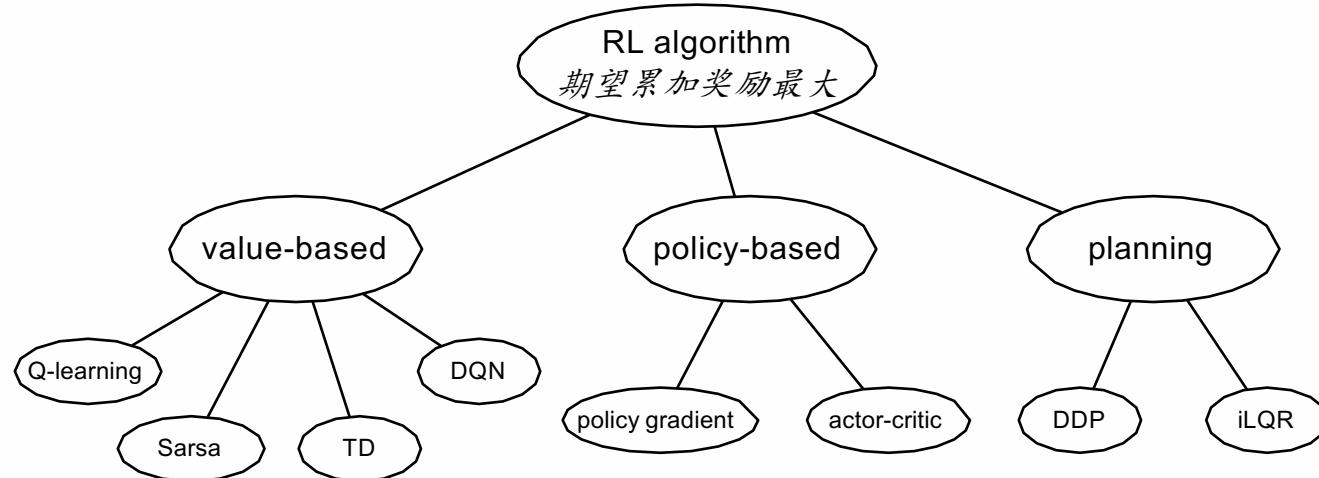




- 基于价值 **value-based**: 主要学习价值函数 $V(s_t)$
 - 策略可由价值函数提取出来



- 基于价值 **value-based**: 主要学习价值函数 $V(s_t)$
 - 策略可由价值函数提取出来
- 基于策略 **policy-based**: 明确定义并学习一个策略函数 $\pi(s_t)$
 - 可以使用价值函数辅助策略的训练，也可以不使用



- 基于价值 **value-based**: 主要学习价值函数 $V(s_t)$
 - 策略可由价值函数提取出来
- 基于策略 **policy-based**: 明确定义并学习一个策略函数 $\pi(s_t)$
 - 可以使用价值函数辅助策略的训练，也可以不使用
- 规划 **planning**: 直接优化动作序列 $\{a_t, a_{t+1}, a_{t+2}, \dots\}$
 - 不借助价值函数和策略函数，通常依赖模型
 - 如 树搜索算法 (Monte-Carlo tree search)

- 在线学习 **online**: 智能体与环境一边交互，一边学习
 - 利用在线的观测数据，不依赖模型，或本身模型就未知
 - 智能体时时使用的是最新的策略
- 离线学习 **offline**: 智能体在线下学习
 - 利用模型或是收集的观测数据进行训练
 - 训练结束后的策略再由智能体在环境中使用

- 基于模型 model-based:
 - 使用模型 P 或模型生成的数据 $\{s' \sim P(s, a)\}$ 训练
 - 利用观测数据构造一个辨识模型 \hat{P} , 基于辨识模型训练
- 不基于模型 model-free:
 - 直接利用观测数据训练价值或策略

课程简介

强化学习介绍

强化学习与其它机器学习的不同

强化学习发展历史

强化学习典型应用

强化学习基本元素

强化学习算法分类

深度强化学习综述

2016年发表，下载6000余次，年度第1
入选F5000提名论文，年度优秀论文

2016年6月

Control Theory & Applications

Jun. 2016

DOI: 10.7641/CTA.2016.60173

深度强化学习综述：兼论计算机围棋的发展

赵冬斌^{1†}, 邵 坤¹, 朱圆恒¹, 李 栋¹, 陈亚冉¹, 王海涛¹

(1. 中国科学院自动化研究所 复杂系统管理与控制国家重点实验室, 北京 100190)

刘德荣², 周 彤³, 王成红⁴

(2. 北京科技大学 自动化学院, 北京 100083; 3. 清华大学 自动化系, 北京 100084;

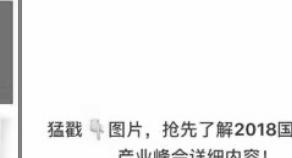
4. 国家自然科学基金委 信息科学部, 北京 100085)

摘要: 深度强化学习将深度学习的感知能力和强化学习的决策能力相结合, 可以直接根据输入的图像进行控制, 是一种更接近人类思维方式的人工智能方法。自提出以来, 深度强化学习在理论和应用方面均取得了显著的成果, 尤其是谷歌深智(DeepMind)团队基于深度强化学习方法研发的计算机围棋“初弈号—AlphaGo”, 在2016年3月以4:1的大比分战胜了世界围棋顶级选手李世石(Lee Sedol), 成为人工智能历史上一个新里程碑。为此, 本文综述深度强化学习的发展历程, 兼论计算机围棋的历史、分析算法特性, 探讨未来的发展趋势和应用前景, 期望能为控制理论与应用新方向的发展提供有价值的参考。

关键词: 深度强化学习; 初弈号; 深度学习; 强化学习; 人工智能

中图分类号: TP273 文献标识码: A

Review of deep reinforcement learning and discussions on the development of computer Go

中国移动 4G 下午5:23 89% 返回 控制理论与应用	中国移动 4G 下午5:22 90% 返回 中国科学院自动化...	中国移动 4G 下午5:22 90% 返回 中国自动化学会 ...	中国移动 4G 下午5:23 89% 返回 德先生 ...	中国移动 4G 下午5:22 90% 返回 人工智能学家 ...	中国移动 4G 下午12:12 88% 返回 新智元 ...
论报道 深度强化学习进展:从Alpha Go到Alpha Go Zero 2018-01-30 控制理论与应用 	【团队新作】深度强化学习进展: 从AlphaGo到AlphaGo Zero 原创 2018-01-31 赵冬斌等 中国科学院自动化研究所 	【前沿】深度强化学习进展: 从AlphaGo到AlphaGo Zero 2018-01-31 中国自动化学会 	中科院自动化所介绍深度强化学习进展: 从AlphaGo到AlphaGo Zero 2018-01-30 德先生 	中科院自动化所介绍深度强化学习进展: 从AlphaGo到AlphaGo Zero 2018-01-31 人工智能学家 	【深度】自动化所解读“深度强化学习”: 从AlphaGo到AlphaGoZero 2017-10-21 新智元 
深度强化学习进展: 2018年发表, 下载6000余次, 年度第1 入选F5000提名论文, 年度优秀论文	2017年底发表, 下载5000余次, 年度第1				

领域顶刊专刊

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

Editorial Special Issue on Deep Reinforcement Learning and Adaptive Dynamic Programming

IN THE first issue of *Nature* 2015, Google DeepMind published a paper "Human-level control through deep reinforcement learning." Furthermore, in the first issue of *Nature* 2016, it published a cover paper "Mastering the game of Go with deep neural networks and tree search" and proposed the computer Go program, AlphaGo. In March 2016, AlphaGo beat the world's top Go player Lee Sedol by 4:1. This becomes a new milestone in artificial intelligence history, the core of which is the algorithm of deep reinforcement learning (RL). Deep RL is able to output control signal directly based

on only action learning capability but also images learning efficiency and testing accuracy.

S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi propose a deep NN-based visual tracker to directly capture the target object in a video with a bounding box. Various training video sequences are used to pretrain the proposed deep NN-based visual tracker, which is further fine-tuned with online adaptation. Deep RL and supervised learning are used for

systems, they discuss Q-learning and integral RL algorithms as core algorithms, respectively. Furthermore, a new direction of off-policy RL for both CT and DT systems is pointed out. Finally, several applications are presented and discussed.

RL in environments with many action-state pairs is challenging. I. J. Sledge, M. S. Emigh, and J. C. Principe propose an uncertainty-based information-theoretic approach for performing guided stochastic searches. They present the value of information as a criterion for the optimal tradeoff between expected costs and the granularity of the search process.

DONGBIN ZHAO, *Guest Editor*

DERONG LIU, *Guest Editor*

F. L. LEWIS, *Guest Editor*

JOSE C. PRINCIPE, *Guest Editor*

STEFANO SQUARTINI, *Guest Editor*



IEEE ToG
创刊主编



IEEE ToG
现任主编

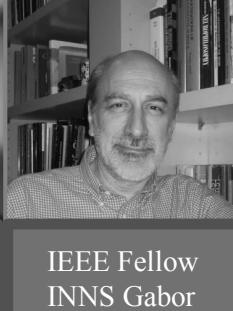
Deep Reinforcement Learning and Games



IEEE Fellow
INNS Fellow



IEEE Fellow
IFAC Fellow



IEEE Fellow
INNS Gabor

international conferences.

Recently, there has been tremendous progress in artificial intelligence (AI), computational intelligence (CI) and games. In 2015, Google DeepMind published a paper "Human-level control through deep reinforcement learning" in *Nature*, showing the power of AI & CI in learning to play Atari video games directly from the screen capture. Furthermore, in *Nature* 2016, it published a cover paper "Mastering the game of Go with deep neural networks and tree search" and proposed the computer Go program, AlphaGo. In March 2016, AlphaGo beat the world's top Go player Lee Sedol by 4:1. In early 2017, the Master, a variant of AlphaGo, won 60 matches against top Go players,

Tremendous progress of deep learning is paving the way to artificial intelligence, games and beyond

DRL is able to output control signals directly based on input images, and integrates the capacity for perception of deep learning (DL) and the decision making of reinforcement learning (RL). This mechanism has many similarities to human modes of thinking. However, there is much work left to do. The theoretical analysis of DRL, e.g., the convergence, stability, and optimality, is still in early days. Learning efficiency needs to be improved by proposing new algo-