

# 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



2023年2月

# 课程简介

# 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

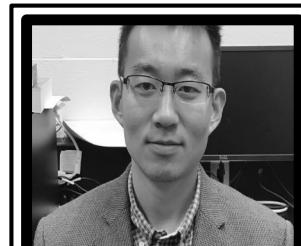
时 长: 40 课时 (上课 38 课时, 考试 2 课时)

评分标准: 实验作业 1(20%)+ 实验作业 2(30%)+ 考试成绩(50%)

华为支持: 课后学生用华为云和MindSpore完成作业



赵冬斌  
研究员



朱圆恒  
副研究员



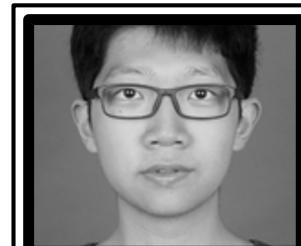
张启超  
副研究员



陈亚冉  
副研究员



李丁  
博士研究生



陆润宇  
博士研究生

# 课程简介-学生评价+学校评价

2022年：选课118人，评估结果：98.03

2021年：选课105人，评估结果：96.51

2020年：选课115人，评估结果：97.01

2022年，“智能基座”产教融合协同育人基地项目优秀教师奖

2021年，“智能基座”产教融合协同育人基地项目优秀教师奖

2022年，中国科学院大学研究生优秀课程奖；



# 课程简介-2020年学生成果

2020年，首届“慧科杯”人工智能应用创新挑战赛最高一等奖（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**

奖项	作品名称	参赛高校	学生团队
一等奖	《基于AutoML的遥感场景分类系统》	中国科学院大学	张佳锋、张浩天、夏彤、刘熠、韩竹
二等奖	《AiPhrase英文仿写系统》	江汉大学	廖晗、冯仔、马丽萍、黄程
	《照骗卸“装”术——AI修图检测小程序》	上海立信会计金融学院、复旦大学、上海财经大学	葛星、刘念一、刘唯洁、任慧颖、沈和威
三等奖	《高空抛物AI监测预警系统》	贵州民族大学人文科技学院	余震武、罗煜星、令狐略
	《基于人工智能的机械臂控制》	上海电机学院	王阿庆、张超、黄炜、蒋安尧、徐果
	《光伏板污迹检测》	上海电力大学	李双圻、闻卫、李家乐、周菲
特别优秀奖	《实时人脸识别》	西北工业大学	牛嘉兴、张露、谢心怡、王志平、李瑞强
	《基于YOLOv4的口罩识别系统》	北京城市学院	夏翔宇、郝梓萍、姚敬凡、周绍翔
	《基于对抗网络进行数据扩充的残差医疗图像分析》	上海建桥学院	薛铁勐、杨振坚、潘博伦
	《AI口罩督查官》	贵州理工学院	苏建川、余永胜、马林、任聪、滕兴

**优秀奖获奖名单**

作品名称	参赛高校	学生团队
智能金融	福州职业技术学院	林汉文、黄峰、陈杨、罗奕鸿
疫情期间口罩智能识别系统	北京城市学院	杨昊琨、魏英瑄、谢高志、陈楷文
食材分析	福州职业技术学院	王斌斌、严世杰、吴琼、李凯辉
垃圾分类系统	贵州大学	冉顺伟、黄正贤、张桂森、龙豪、徐文镇
ELX—智能随身信用仪	上海立信会计金融学院	张畅通、何奕、贺昱帆、许瀛、谷悦嘉
基于强化学习方法的自动化神经网络结构微调	中国科学院大学	才玉、陈志扬、李建军、伍虹燕、张普
超市自动识别果蔬称重	贵州民族大学人文科技学院	林战、陈俊、谭文江、彭发港
基于DQN训练的吃豆人AI	上海电力大学	闫南、宋鹏飞、皇甫百香、戴宇轩
AI识相识病	四川华新现代职业学院	肖铭、何挂华、曹亚萍、张永洪、粟泽亮
服装推荐	贵州大学、温州大学	吴义琦、冉芳科、杨耀凯、汪召鹏、张旭
知心考勤	上海电机学院	王峯杰、袁露、郭智伟
基于带目标网络DQN的防抖和学步系统	中国科学院大学	徐文博、王常维、韩立元、范嗣祺、牛植方
肺炎诊断	福州职业技术学院	许富杰、黄锦祥、叶嘉晖、魏舒帆
好利来烘焙店销售数据分析系统	江西软件职业技术大学	朱博文、俞国龙、周学凡、俞持鑫、刘奔
LWT.人脸识别识别	贵州民族大学人文科技学院	卢垚、吴梦、刘庆辉、陶发丽、刘红梅
深度学习聊天机器人	渤海职业技术学院	胡清泉、常亚南、徐懿、王艳冬
零一人脸识别自动报警技术	贵州民族大学人文科技学院	钱兰、邓培超、孙娜、罗志凤
五子棋游戏	西安文理学院	雷颖、李冰洁、李笑、屈佳怡、贺雨丹
基于计算机视角的交通场景车牌识别系统	贵州民族大学人文科技学院	陆安隆、王荣情、张航
路面智能识别(山海)	渤海职业技术学院	邱成宏、邱海远、白苗苗、张昊、聂恒丹
证件识别	广西大学行健文理学院	李源、时京欣、王倩
用户反馈语言分析系统	西安铁路职业技术学院	时金刚、陈亮亮、黄佳、刘忻乐、张艺娜
机动车识别系统	北京城市学院	周键、于洋、韩国伟、刘佳豪
Artemis-基于Nodejs的简单人脸识别打卡系统	贵州大学	王海波、陈嘉琪、彭浩然
房价评估模型	湖南科技学院	何志豪、王海雨、孟基、宋立仁、冯俊淇
基于深度学习的垃圾分类识别	贵州大学、贵州民族大学人文科技学院、贵州理工学院	李梅艳、吴亚波、吴运嘉、吴胜宇

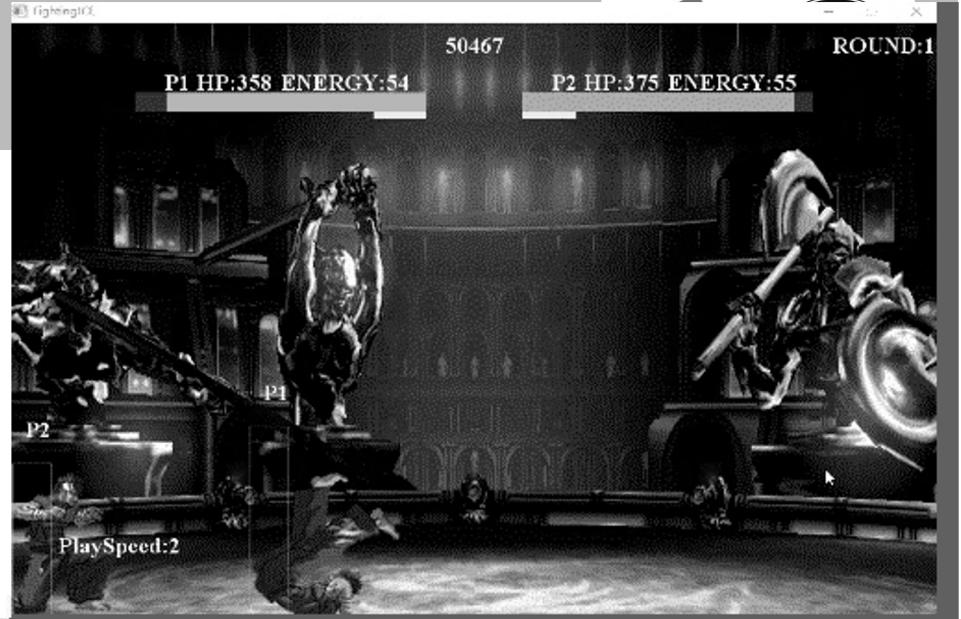
# 课程简介-2020年学生成果

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

	ZEN	GARNET	LUD
Fuzzy_ZYQAI	0	0	0
TeraThunder	12	18	10
SpringAI	10	4	18
EmcmAI	25	15	12
ERHEA	18	25	25
Caselene	0	0	2
MrTwo	2	6	0
CYR_AI	15	8	15
JayBot	6	2	1
ButcherPudge	8	12	8
YIYAI	0	0	0
Jitwisut_Zen	4	10	6
MonkeyLink_TriplePM	0	0	0
Noobot	0	0	0
SampleMctsAI	1	2	4

## Results

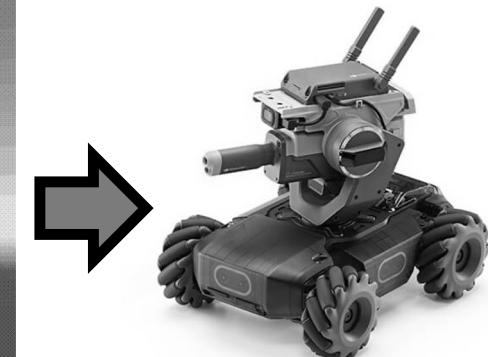
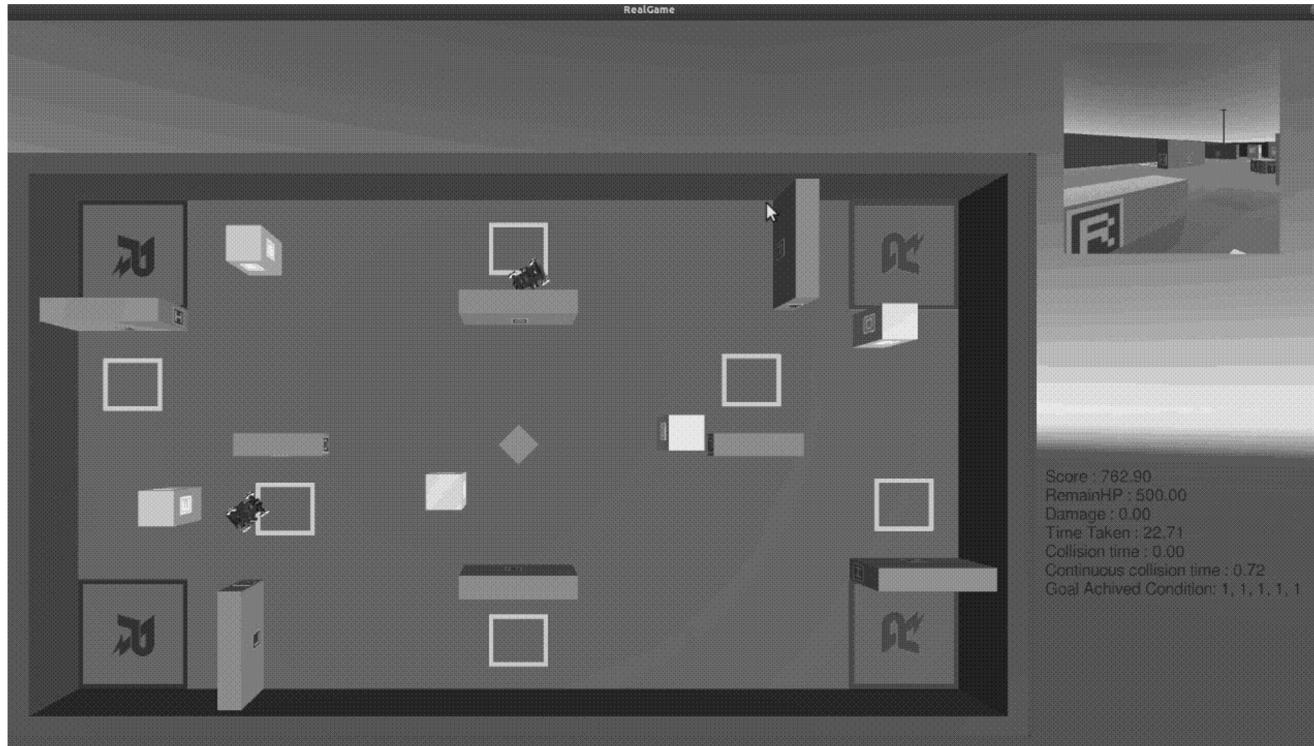
	SUM	RANK
Fuzzy_ZYQAI	0	13.5
TeraThunder	88	2
SpringAI	59	6
EmcmAI	72	4
ERHEA	128	1
Caselene	11	11
MrTwo	29	8
CYR_AI	71	5
JayBot	20	9.5
ButcherPudge	73	3
YIYAI	0	13.5
Jitwisut_Zen	36	7
MonkeyLink_TriplePM	0	13.5
Noobot	0	13.5
SampleMctsAI	20	9.5



	ZEN	GARNET	LUD
Fuzzy_ZYQAI	0	0	0
TeraThunder	18	18	12
SpringAI	12	0	15
EmcmAI	8	6	6
ERHEA	25	25	10
Caselene	0	8	1
MrTwo	2	15	4
CYR_AI	4	4	25
JayBot	10	1	0
ButcherPudge	15	12	18
YIYAI	0	0	0
Jitwisut_Zen	6	2	8
MonkeyLink_TriplePM	0	0	0
Noobot	0	0	0
SampleMctsAI	1	10	2

- **Winner AI: ERHEA\_PI 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.

# 课程简介-2022年学生成果



**2022 IEEE Conference on Games**组织了**RoboMaster Sim2Real Competition**

**内容：**提供机器人在地图中的位置，机器人根据车载相机拍摄的图像、目标点的位置等信息，AI算法输出对机器人的导航和对抗控制；

**比赛奖励：**冠军500美元+大疆产品实物+证书；

**报名阶段：**4月1日-5月10日，8月16日结束

<https://eval.ai/web/challenges/challenge-page/1513/overview>

**学生结果：**桑明，刘晓东，第6名/48名（全球包括香港大学等）；

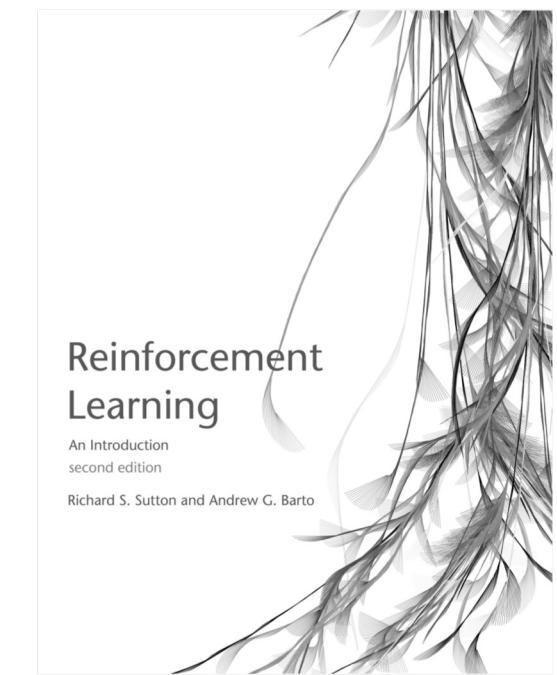
# 课程简介

国外相关课程，作为部分讲义参考，仅用于教学和交流。

- 1 Sutton & Barto, 1998/2018, “**Reinforcement Learning: An Introduction**”
- 2 强化学习圣经，1998年第一版开山之作，2020年更新为第二版，系统讲述了强化学习问题，查表法（多臂老虎机，有限马尔科夫决策过程，动态规划，蒙特卡洛方法，时间差分，n步自举，规划和学习）、函数逼近法（同策略的预测和控制，异策略，资格迹，策略梯度），和前沿问题（和心理学、神经科学交叉的前沿，AlphaGo等的典型应用成果，以及一些前沿探索讨论）。
- 3 <http://incompleteideas.net/book/the-book-2nd.html>
- 4 电子版网上可下载，网页上同时提供了授课课件，课后作业相关的代码可参考，等等。
- 5 建议：可反复阅读学习，并和作业的程序编写调试结合起来，以加深理解。



Andrew G. Barto Richard S. Sutton



# 课程简介

国外相关课程，作为部分讲义参考，仅用于教学和交流。

- **David Silver**, University College London Course on Reinforcement Learning
    - <https://www.davidsilver.uk/teaching/>
    - 10 lectures\*1h30min+1 project
  - Emma Brunskill, Stanford CS234 Reinforcement Learning
    - 16 lectures\*1h20min+3 assignments+1 project
  - Sergey Levine, UC Berkeley CS 294 Deep Reinforcement Learning
    - 28 lectures\*1h30min+5 homeworks+1 project
  - CMU 10703: Deep RL and Control
    - 30 lectures\*1h20min+3 homeworks+1 project
  - 李宏毅，机器学习
- 1 .....



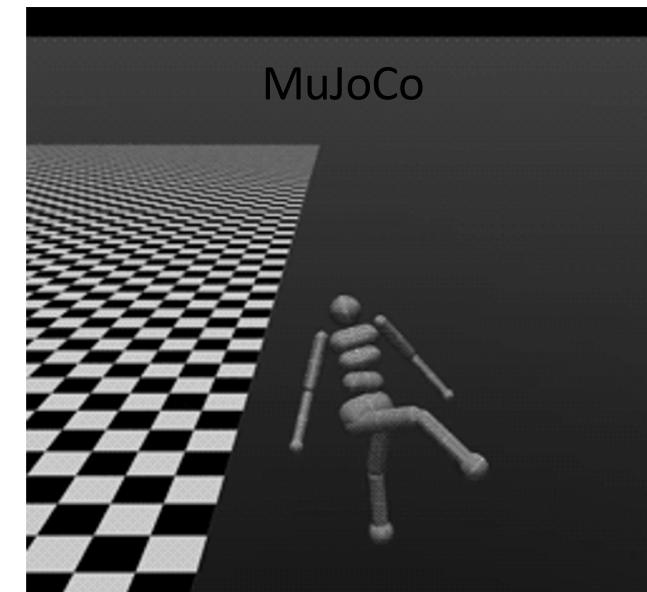
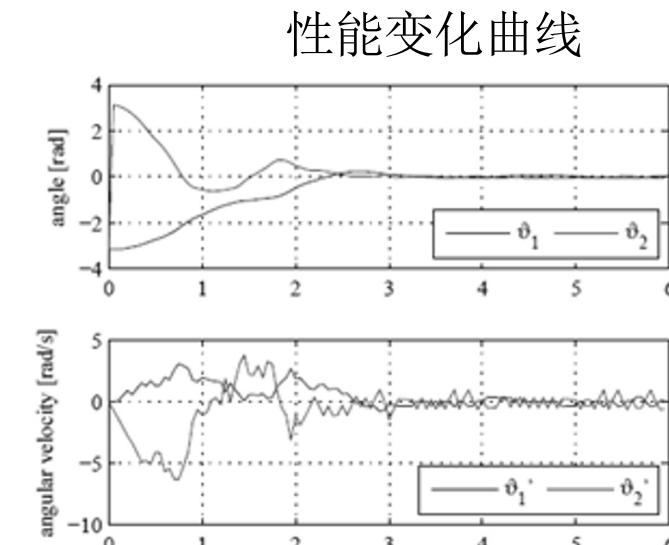
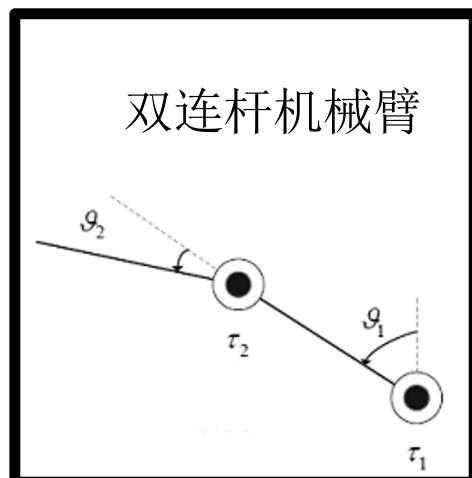
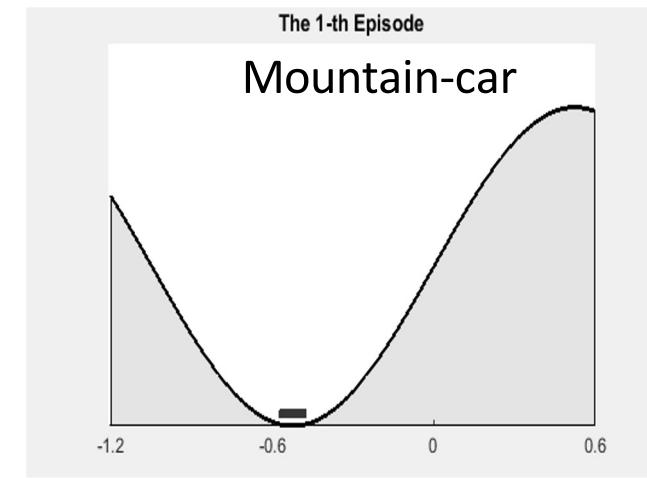
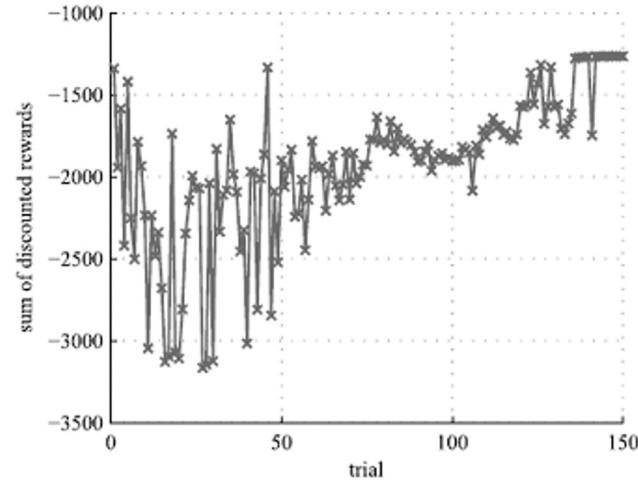
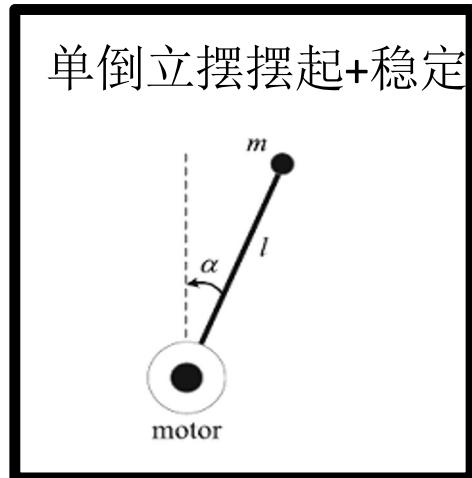
David Silver

# 课程内容

1. 强化学习概述
2. 马尔可夫过程+ 第1次作业安排(基本强化学习方法)
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-策略梯度方法
8. 策略梯度方法2+基于博弈理论的强化学习+小组研讨课1
9. 第2次作业+强化学习基线算法+华为基座和MindSpore介绍
10. 逆强化学习+离线强化学习
11. 深度强化学习与游戏AI
12. 深度强化学习与智能驾驶
13. 小组研讨课2

# 课程内容

## ➤作业1：倒立摆的优化控制问题及一些常规的RL benchmark算法测试



# 课程内容

## ➤作业2的要求

### ✓ 提交材料

- 3~5人一个小组，以组为单位提交一份大报告
- 华为MindSpore程序代码，基本介绍：
- <https://gitee.com/mindspore/reinforcement/tree/master>
- 可演示的demo

### ✓ 要求：

- 需要在大报告里体现每个人分工完成的部分
- 必须涉及到强化学习/深度强化学习方法
- 大报告可以以会议论文格式，或者以PPT汇报的形式
- 建议进行不同算法的对比分析
- 可提前和老师确认好非指定的作业任务，如比赛内容

严格禁止抄袭，如有发现，大作业成绩全组计零！

# 强化学习介绍

## 强化学习与多学科交叉

## 相关领域

## 工程（自动化/电气）

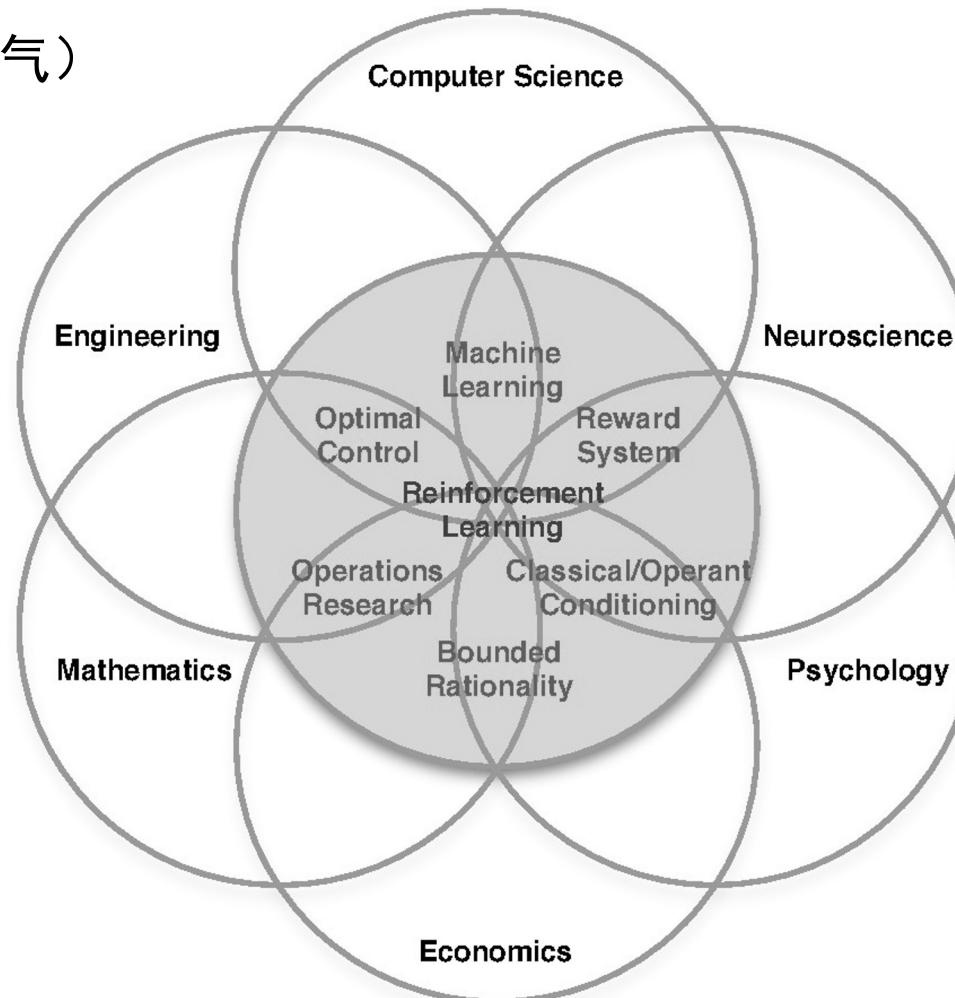
计算机科学

## 神经系统科学

数学

心理学

## 经济学



# 与现有课程体系的结构关系

# 机器学习

最优控制

运筹学

线性代数

概率统计

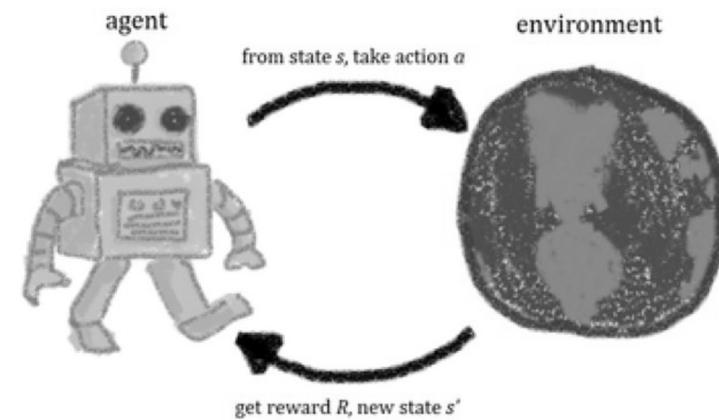
## 随机过程

计算机编程

--from David Silver, Reinforcement Learning Course, UCL

# 强化学习 (Reinforcement Learning)

- 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



# 生物学的启示

## ■ 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

# 强化学习过程

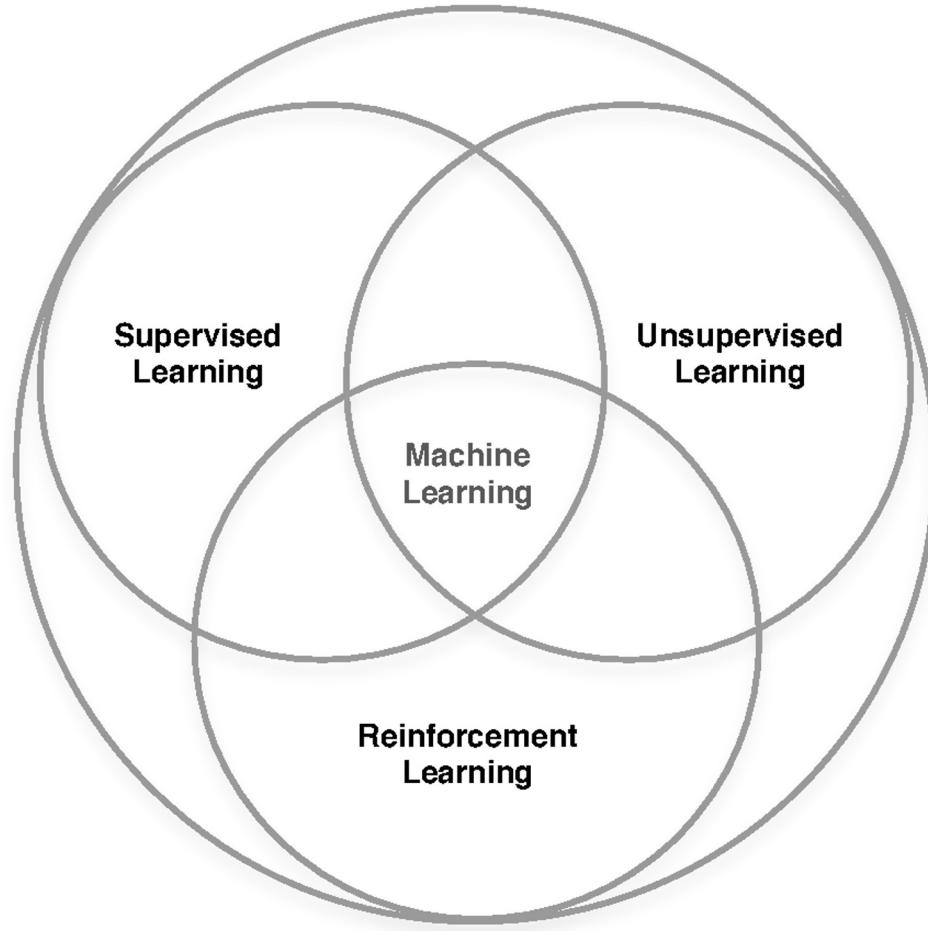
- 根据环境反馈的奖励，调整智能体的行为策略，提升智能体实现目标的能力。
  - 没有明确告诉采取哪些动作是可以实现目标
  - 通过间接的奖励信号反映完成目标的情况(稀疏)
  - 例如：下棋输赢 +1/-1，汽车行驶碰撞 +1/-1，机器人离目标点的距离
  - 好处：简单，便宜

# 强化学习过程

- 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，避免负反馈的策略
- 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

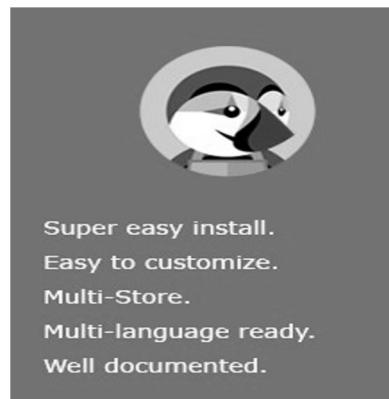
# 机器学习分支 Machine Learning (ML)



## ■ 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

This item: Structure and Interpretation of Computer Programs - 2nd Edition (MIT Electrical Engineering and... by Harold Abelson Paperback \$50.50 + The Pragmatic Programmer: From Journeyman to Master by Andrew Hunt Paperback \$32.59 Total price: \$83.09 Add both to Cart Add both to List

Customers Who Bought This Item Also Bought

The Little Schemer - 4th Edition \$39.00 ✓Prime Instructor's Manual for Structure and Interpretation of Computer Programs... \$28.70 ✓Prime The Pragmatic Programmer: From Journeyman to Master \$32.59 ✓Prime Introduction to Algorithms, 3rd Edition (MIT Press) \$70.00 ✓Prime An Introduction to Functional Programming Through Lambda Calculus \$66.32 ✓Prime Purely Functional Data Structures \$40.74 ✓Prime Code: The Hidden Language of Computer Hardware and Software \$17.99 ✓Prime The Little Prover (MIT Press) \$31.78 ✓Prime

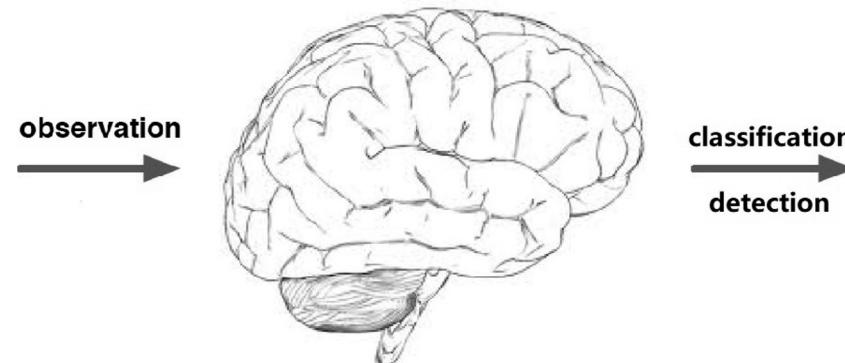
推荐系统

# 序贯决策过程 (sequential decision making)

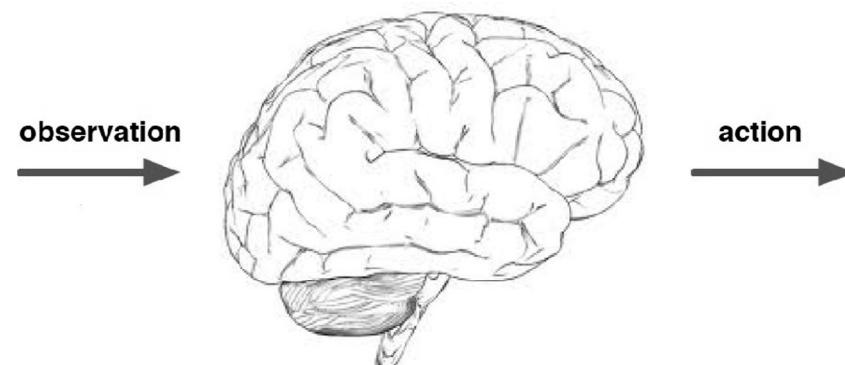
- 智能体处在特定的环境中产生一系列的动作，而这些动作改变智能体的状态。
- 举例
  - 1 遥控直升飞机的特技表演
  - 2 打败围棋世界冠军
  - 3 管理股票证券
  - 4 发电厂调控
  - 5 控制人型机器人双足行走
  - 6 视频游戏上超越人类
- 强化学习考虑的是 序贯决策过程

# 大脑的功能

#### ■ 感知：识别或估计观测的内容



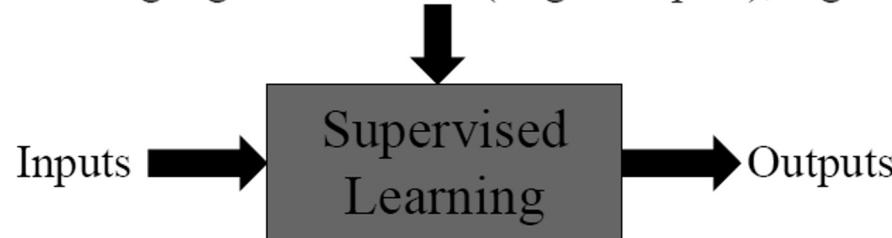
#### ■ 决策：根据观测做出行为



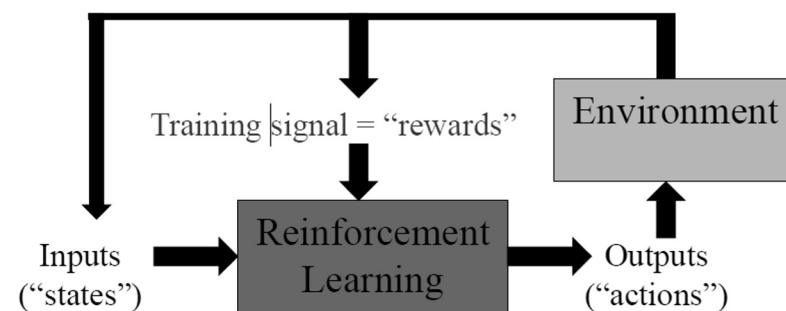
# 大脑的功能

## ■ 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



# 强化学习与其它机器学习的不同

## ■ 强化学习：

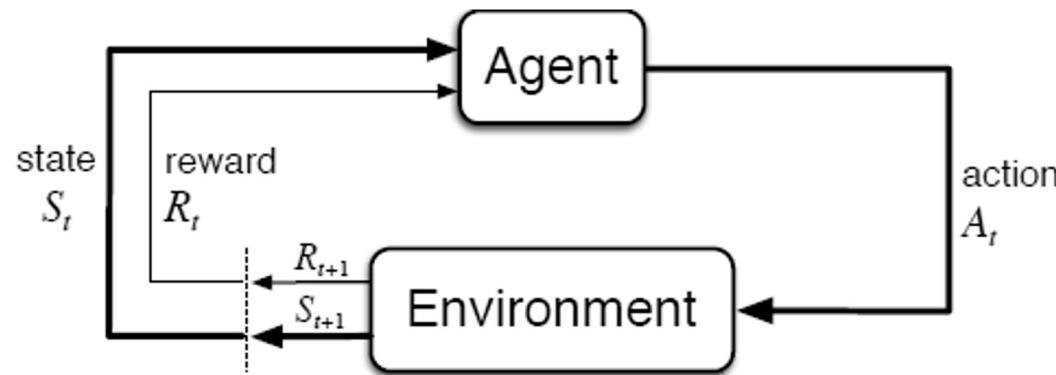
- 1 产生的结果（动作）能够改变数据的分布（状态）
- 2 最终的目标可能要很长时间才能观察到/奖励稀疏 (e.g. 下棋)
- 3 没有明确的标签 (label) 数据
- 4 根据当前的奖励，最终实现长远的目标

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

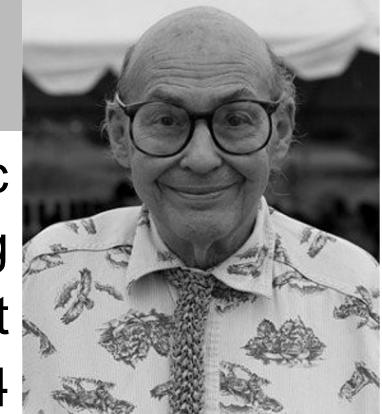
- 1 产生的结果（输出）不会改变数据的分布
- 2 结果是瞬时的/输出误差
- 3 要么有明确的标签数据 (SL)
- 4 要么完全没有任何标签数据 (USL)

# 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



- **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  $\pi$
- 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

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

0.22	0.25	0.27	0.30	0.34	0.38	0.34	0.30	0.34	0.38
→	↑	↑	→	↑	↓	↑	↑	↑	↓
0.25	0.27	0.30	0.34	0.38	0.42	0.38	0.34	0.36	0.42
→	→	→	→	→	↓	→	→	→	↓
0.21					0.46				0.46
↑					↓				↓
0.20	0.22	0.25	-0.78		0.52	0.57	0.64	0.57	0.52
→	↑	↓	↑		→	→	↓	↑	↑
0.22	0.25	0.27	0.25		0.08	-0.36	0.71	0.64	0.57
↑	↑	↓	↑		↓	→	↓	↑	↑
0.25	0.27	0.30	0.27		1.20	0.08	0.79	-0.29	0.54
↑	↑	↓	↑		↑	↓	↓	↑	↓
0.27	0.30	0.34	0.30		1.01	0.97	0.87	-0.21	0.57
↑	↑	↓	↑		↑	↑	↑	↑	↓
0.31	0.34	0.38	-0.58		0.01	-0.19	0.71	0.71	0.64
↑	↑	↓	↑		↑	↑	↑	↑	↑
0.34	0.38	0.42	0.46	0.52	0.57	0.64	0.71	0.64	0.51
→	→	→	→	→	→	→	↑	↑	↑
0.31	0.34	0.38	0.42	0.46	0.52	0.51	0.61	0.51	0.51

# 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。

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

# Richard Ernest Bellman 1920-1984

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

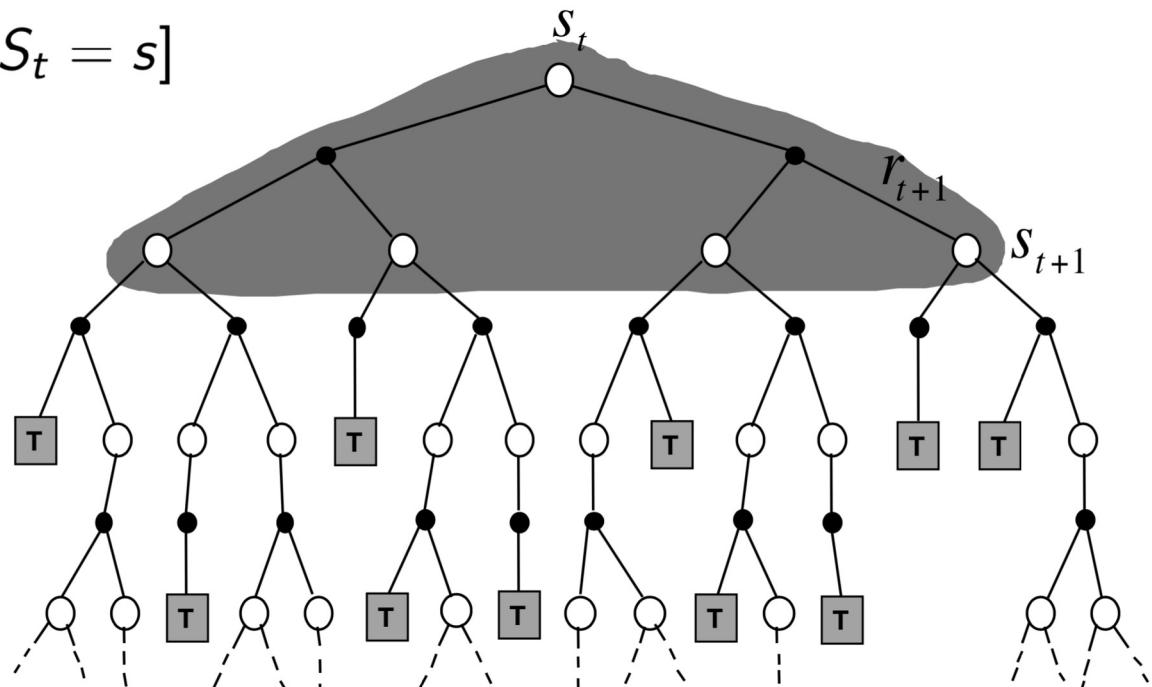
$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi$$

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

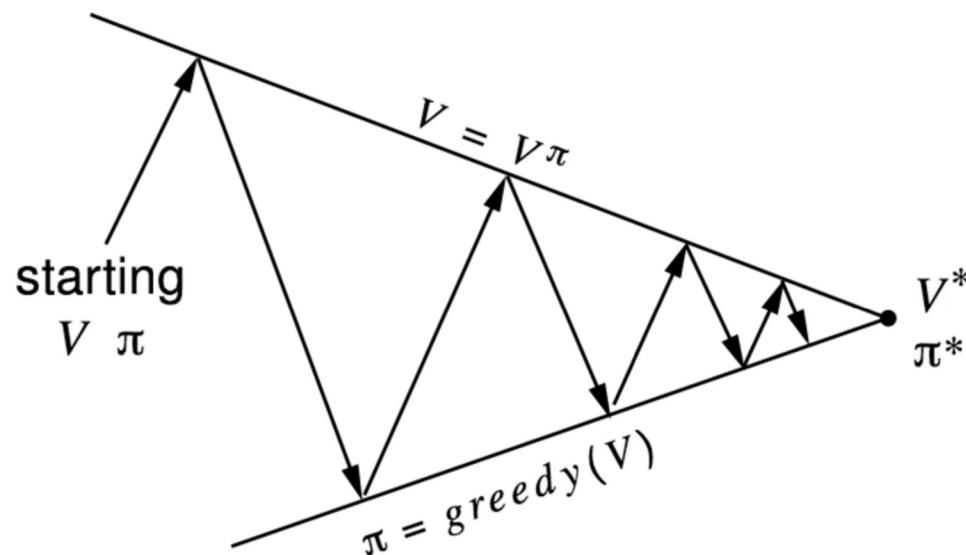
维数灾：离散状态、动作空间大

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

策略迭代、值迭代



## 策略迭代/值迭代（第3讲）



## 策略迭代

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

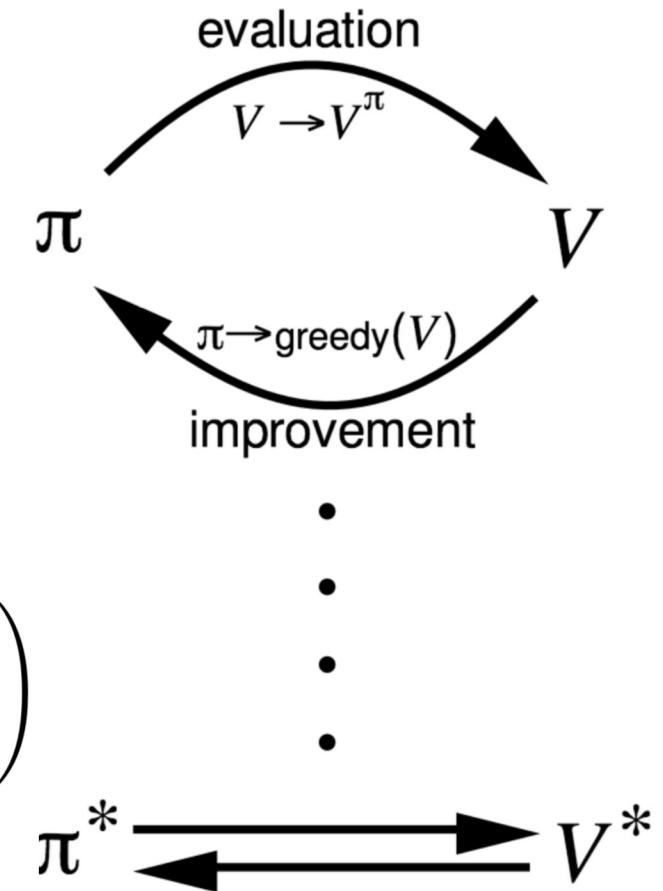
$$\mathbf{v}^{k+1} = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi \mathbf{v}^k$$

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

## 值迭代

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

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

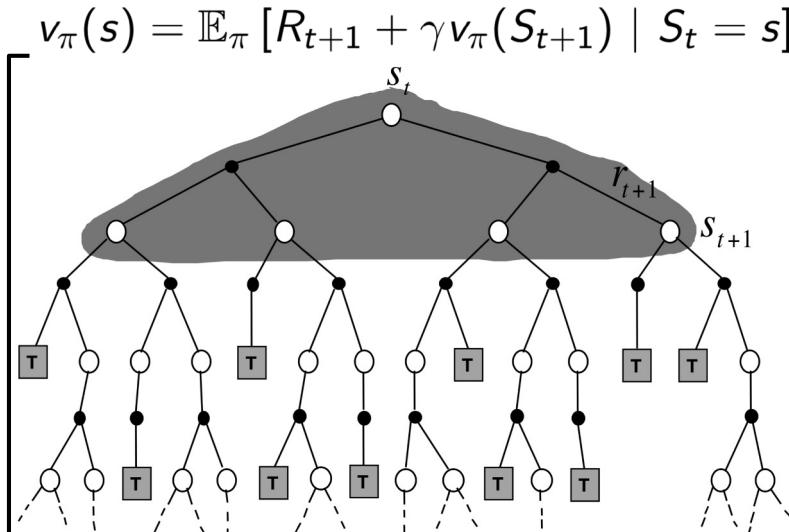


# 蒙特卡洛算法和时间差分学习算法（第4讲）



Richard S. Sutton  
Temporal-Difference  
(TD) 1988

动态规划

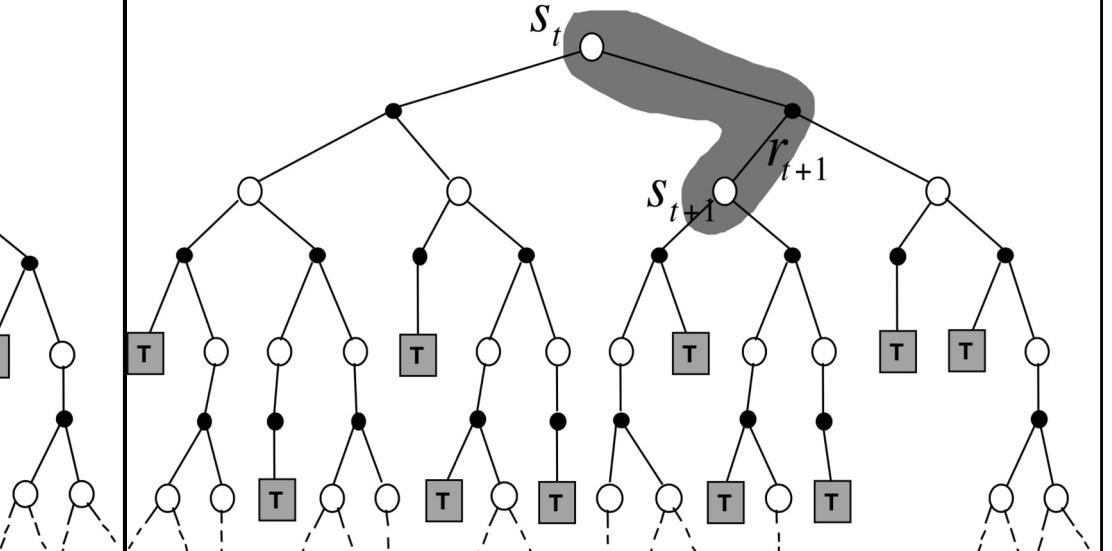
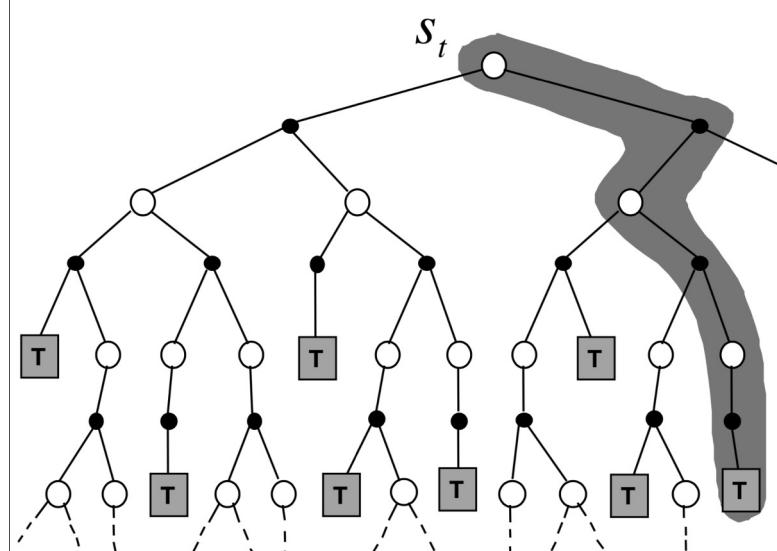


蒙特卡洛算法

时间差分学习算法

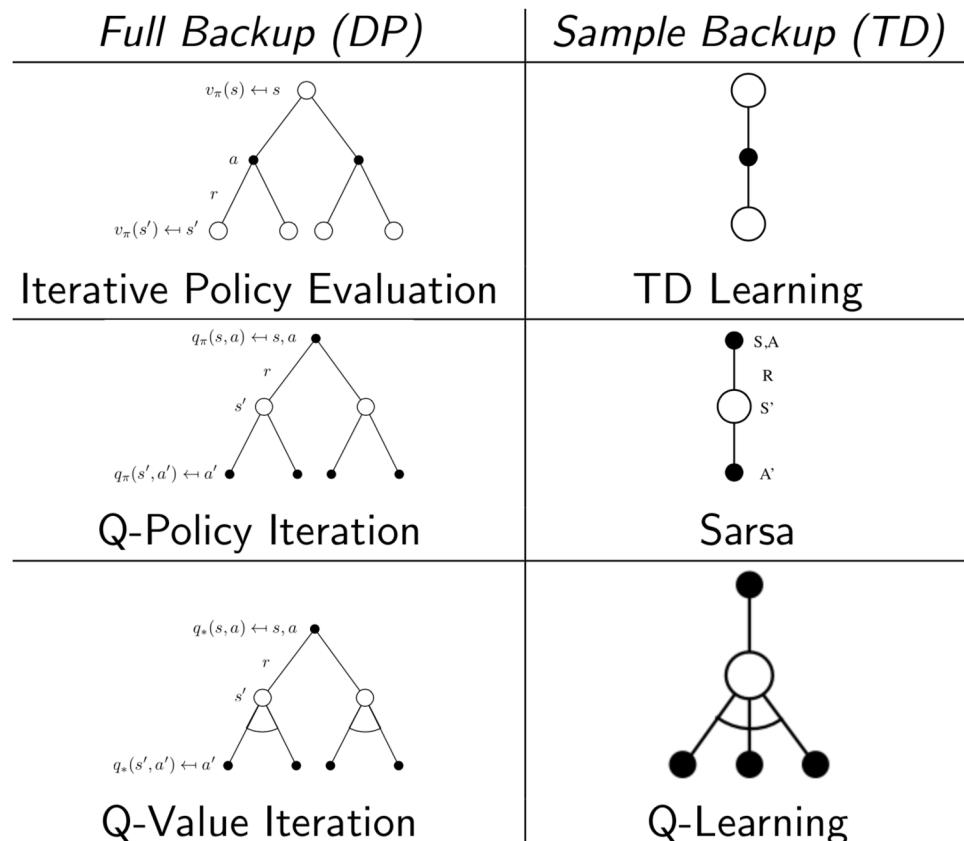
$$V(S_t) \leftarrow V(S_t) + \alpha (G_t - V(S_t))$$

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



# 无模型学习控制：Sarsa和Q学习（第5讲）

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



*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

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

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)

$$\tilde{v}_k(s) = \max_{a \in \mathcal{A}} \left( \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \hat{v}(s', \mathbf{w}_k) \right)$$

Adaptive DP(ADP)

$\hat{J}(t+1)$



Paul Werbos  
误差反传 1974

ADP 1977

脑科学  
能源管理

$\underline{X}(t)$

$\hat{\mathbf{R}}$



$\underline{R}(t)$

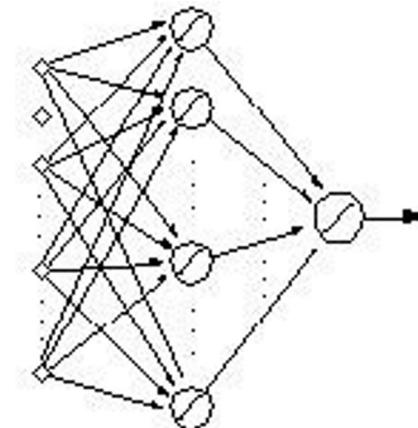
Red Arrows:  
Derivatives  
Calculated By  
Generalized  
Backpropagation

Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook*, 22(12):25–38.

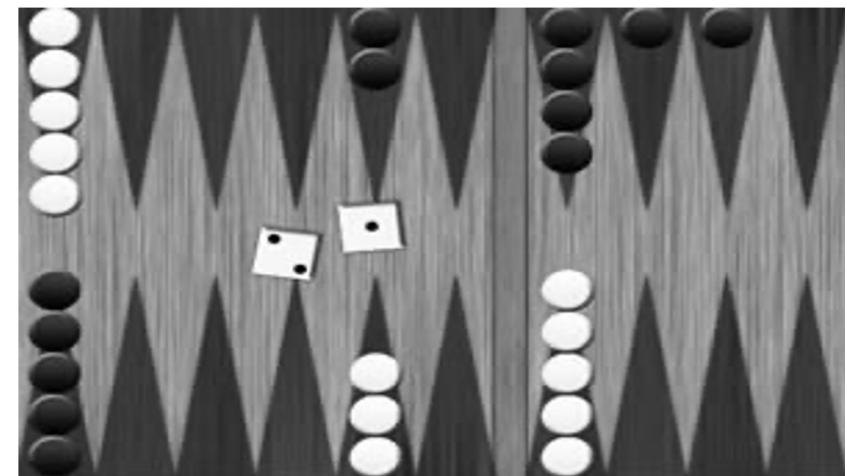
# TD-Gammon

- 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

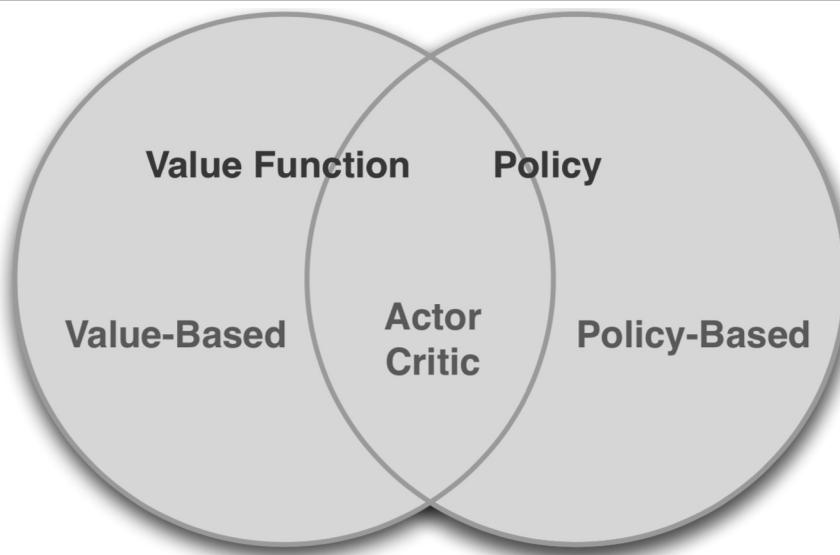


Gerald Tesauro  
Backgammon 1992  
IBM Watson  
Deep Blue  
.....



# 策略梯度 (第7讲)

Ronald J. Williams  
REINFORCE 1992



$$J(\theta) = \mathbb{E}_{\pi_\theta} [r]$$

$$= \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \pi_\theta(s, a) \mathcal{R}_{s,a}$$

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

$$= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) r]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) v_t]$$

REINFORCE

$$= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) Q^w(s, a)]$$

Q Actor-Critic

$$= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) A^w(s, a)]$$

Advantage Actor-Critic

$$= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) \delta]$$

TD Actor-Critic

$$= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a) \delta e]$$

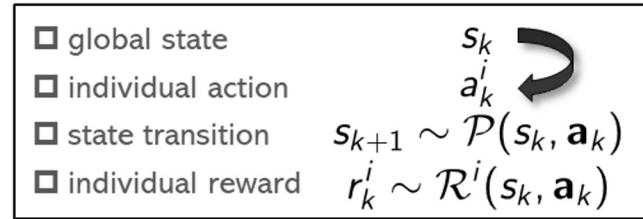
TD( $\lambda$ ) Actor-Critic

Natural Actor-Critic

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



**马尔可夫博弈**  $\mathcal{MG} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, \{\mathcal{R}^i\}_{i \in \mathcal{N}}, \mathcal{P}, \gamma, \rho_0)$

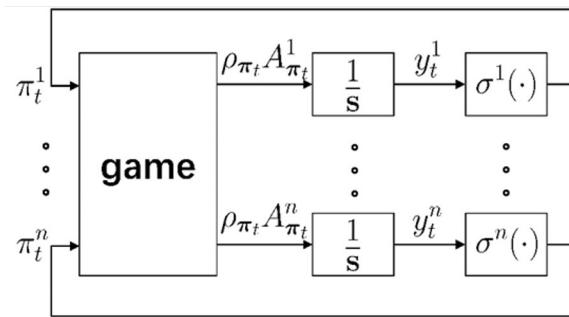


**策略 policy**  
 $\pi : \mathcal{S} \times \mathcal{A}^i \rightarrow [0, 1]$   
 $a_k^i \sim \pi(s_k)$

**纳什均衡策略**  $\pi_* = (\pi_*^i)_{i \in \mathcal{N}}$   $\leftrightarrow V^i(\pi_*^i, \pi_*^{-i}) \geq V^i(\pi^i, \pi_*^{-i}), \forall \pi^i \in \Pi^i$

**贝尔曼最小最大方程**  $V_*(s) = \max_{\pi^1(s) \in \Pi^1} \min_{a^2 \in \mathcal{A}^2} \sum_{a^1 \in \mathcal{A}^1} \pi^1(a^1|s) \left[ \mathcal{R}(s, a^1, a^2) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a^1, a^2) V_*(s') \right]$

**连续时间动力学模型**



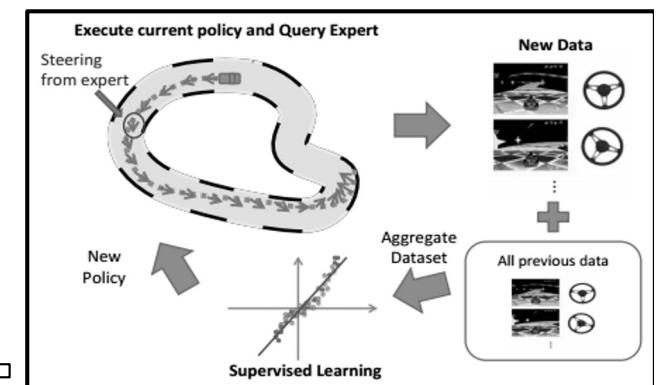
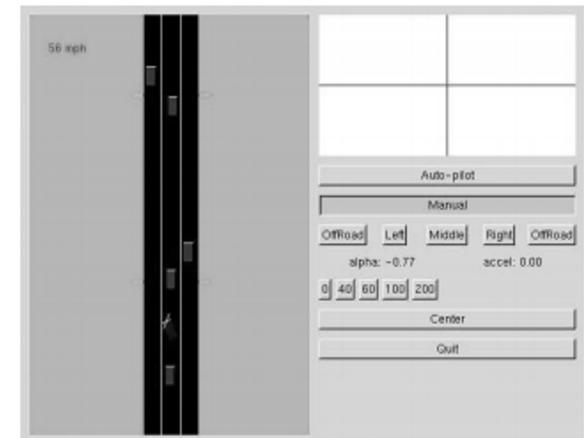
Three-stage process

- 1. **Assessment Stage:** 将玩家在当前博弈中的策略性能  $\rho_{\pi_t} A_{\pi_t}^i$  累积到  $y_t^i$  中
- 2. **Choice Stage:** 根据  $y_t^i$  利用  $\sigma^i(\cdot)$  更新出玩家新的策略  $\pi_t^i = \sigma^i(y_t^i)$
- 3. **Game Stage:** 观察各玩家使用最新的策略博弈时的表现  $\rho_{\pi_t}, A_{\pi_t}^i$

# 逆强化学习（第10讲）

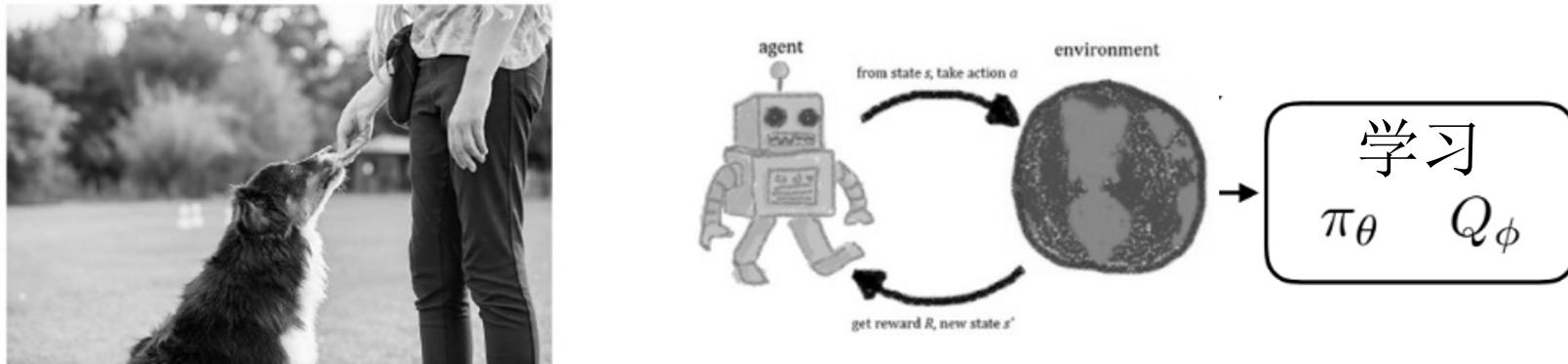
## 奖赏信号：难以人为事先给定

- 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。
- 2023年，牛津大学发现人类偏好与人类行为之间的真正关系比目前 IRL 中使用的任何模型都要复杂得多，从理论上对 IRL 模型进行了数学分析，获AAAI 2023杰出论文奖。

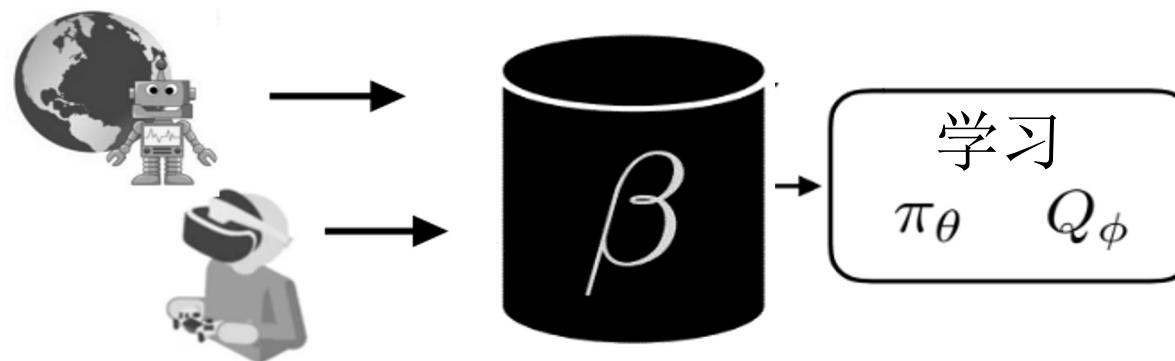


# 离线强化学习（第10讲）

- Online RL: 依赖与仿真器在线交互试错学习



- Offline RL: 仅依赖离线交互数据进行学习



离线数据集  $\mathcal{D} = \{(s, a, s', r)_j\}$

课程将详细介绍以下最新工作

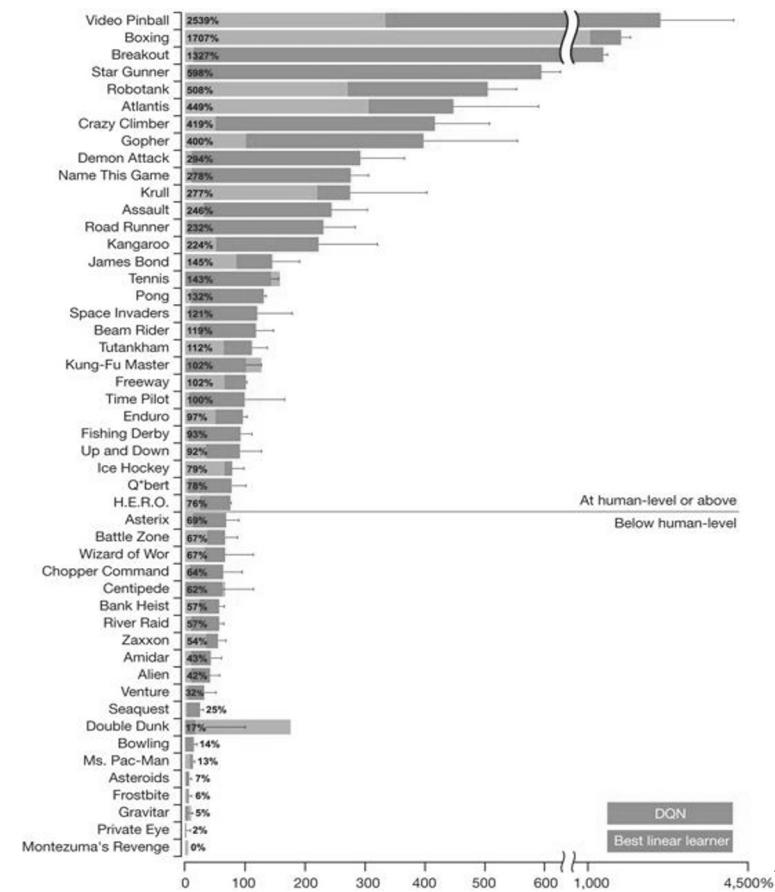
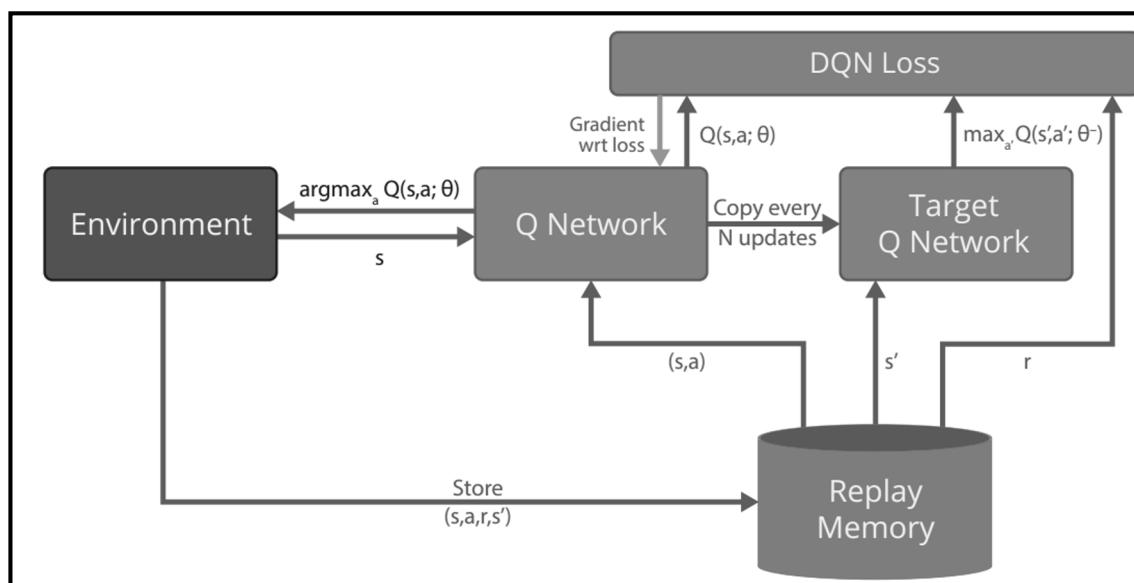
- BCQ (ICML 2019)
- CQL (NeurIPS 2020)
- TD3+BC(NeurIPS 2021)
- IQL (ICLR 2022)
- POR (NeurIPS 2022)

# DQN (第11讲)

Volodymyr Mnih  
多伦多大学  
DQN, A3C

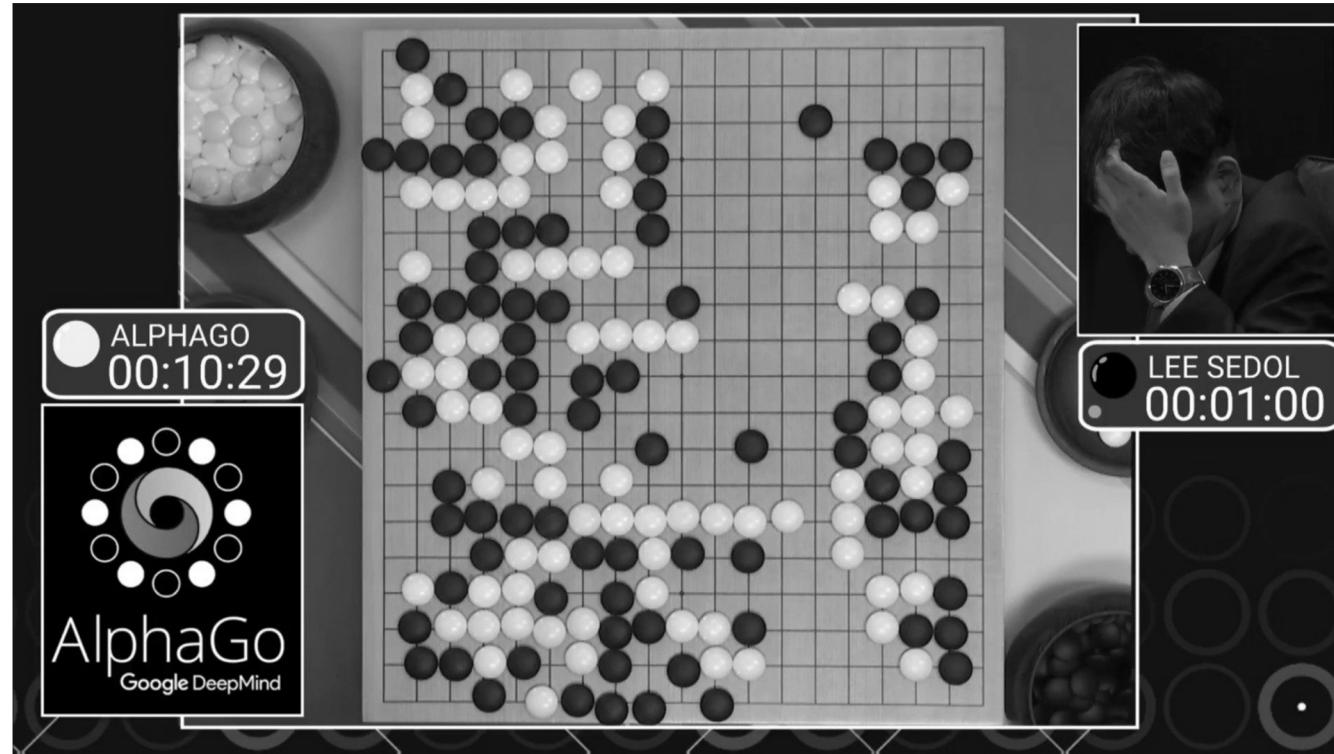


DeepMind, 2015年2月 Nature  
DQN 算法，将卷积神经网络和 Q 学习结合，并集成了经验回放技术，在 57 款 Atari 游戏上超过了人类水平。



# AlphaGo (第11讲)

David Silver  
UCL教授  
Alpha系列

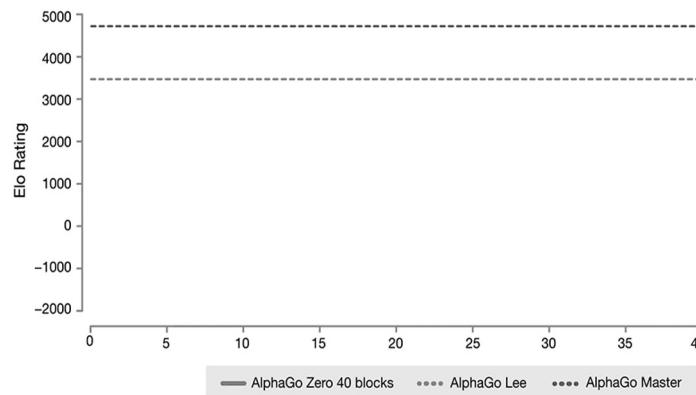


- 2016 年 3 月, DeepMind 开发的 AlphaGo 围棋程序 4-1 战胜 Lee Sedol(前世界排名第一)
- DeepMind, *Nature*, 2016

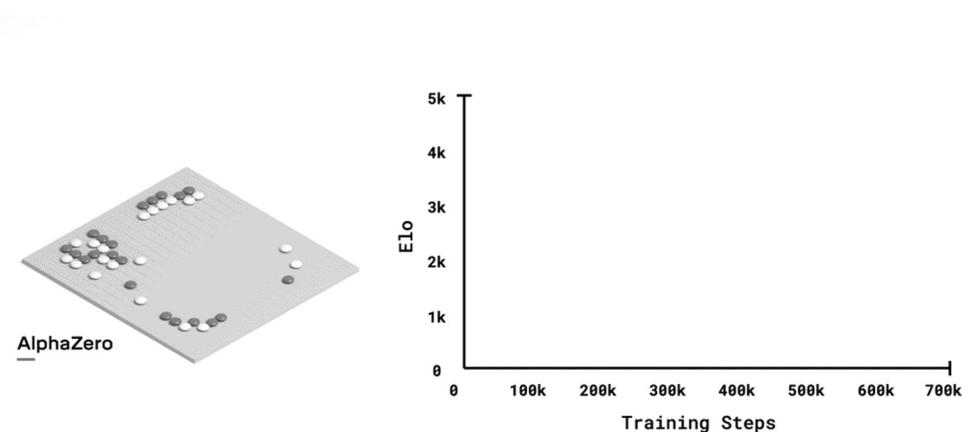
赵, 邵, 朱, 李, 陈, 王, 刘, 周, 王。深度强化学习综述: 兼论计算机围棋的发展, 控制理论与应用, 2016. (入选科技部F5000, 本学科前1%高被引论文, 《控制理论与应用》年度优秀论文)

# AlphaGo Zero/AlphaZero

- Deepmind AlphaGo Zero masters Go without human knowledge (*Nature*, 2017)

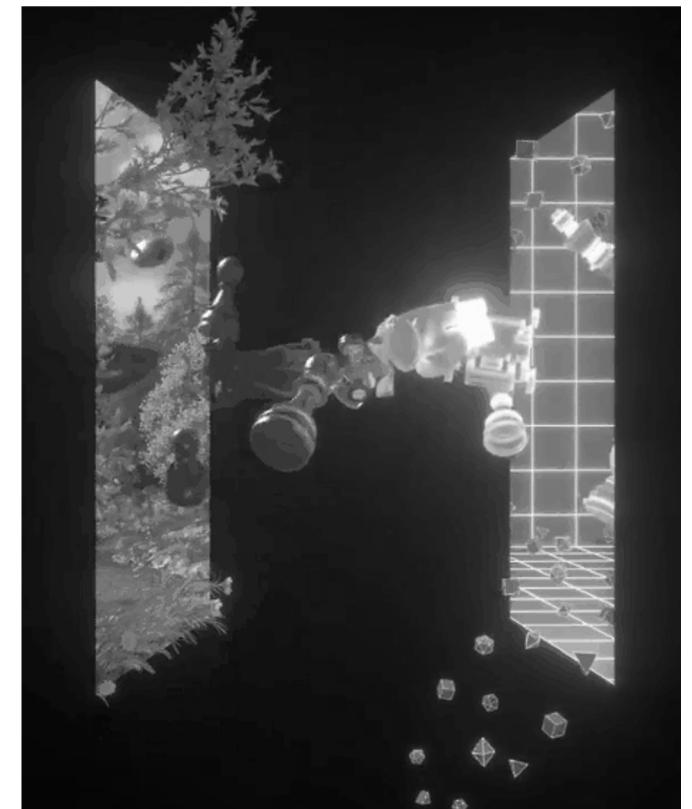
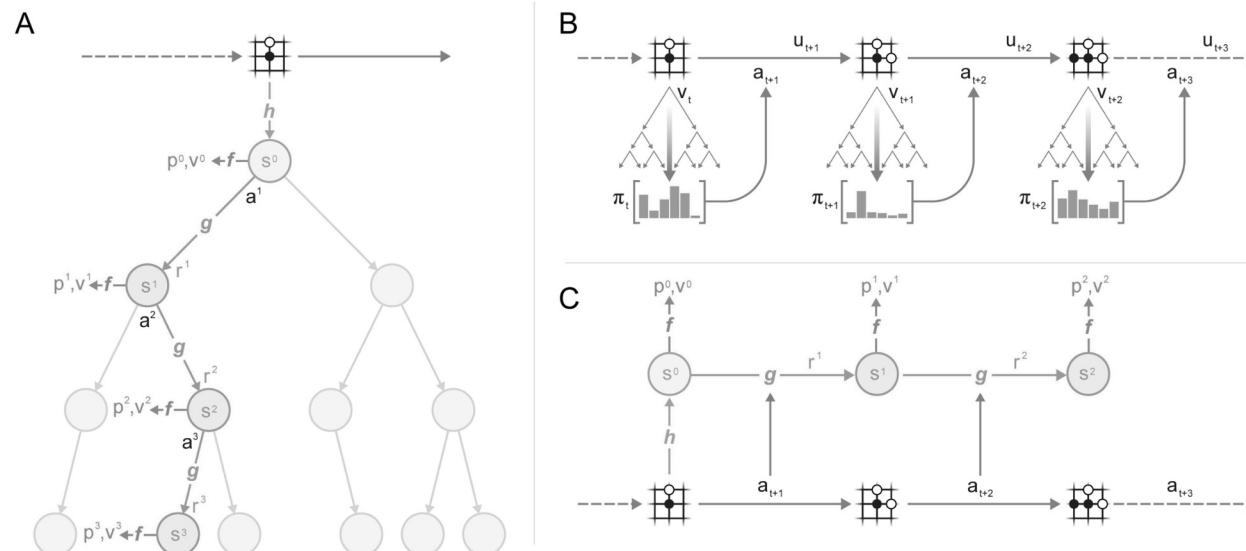


- AlphaZero masters chess, shogi, and Go through self-play (*Science*, 2018)



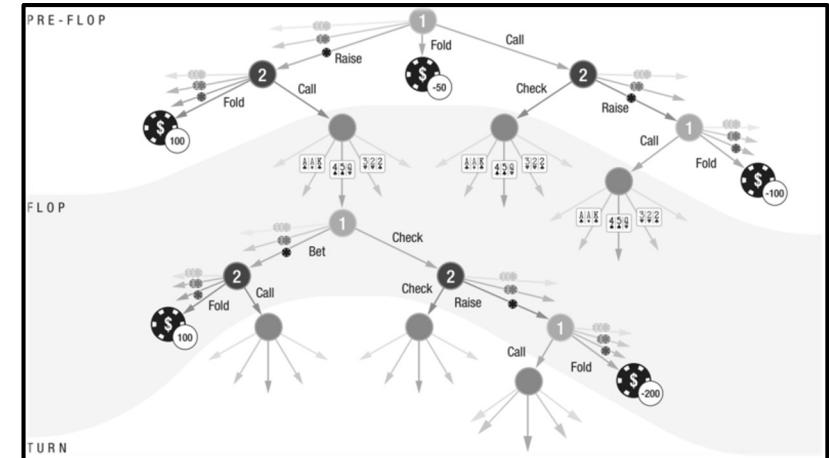
# MuZero

- 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*

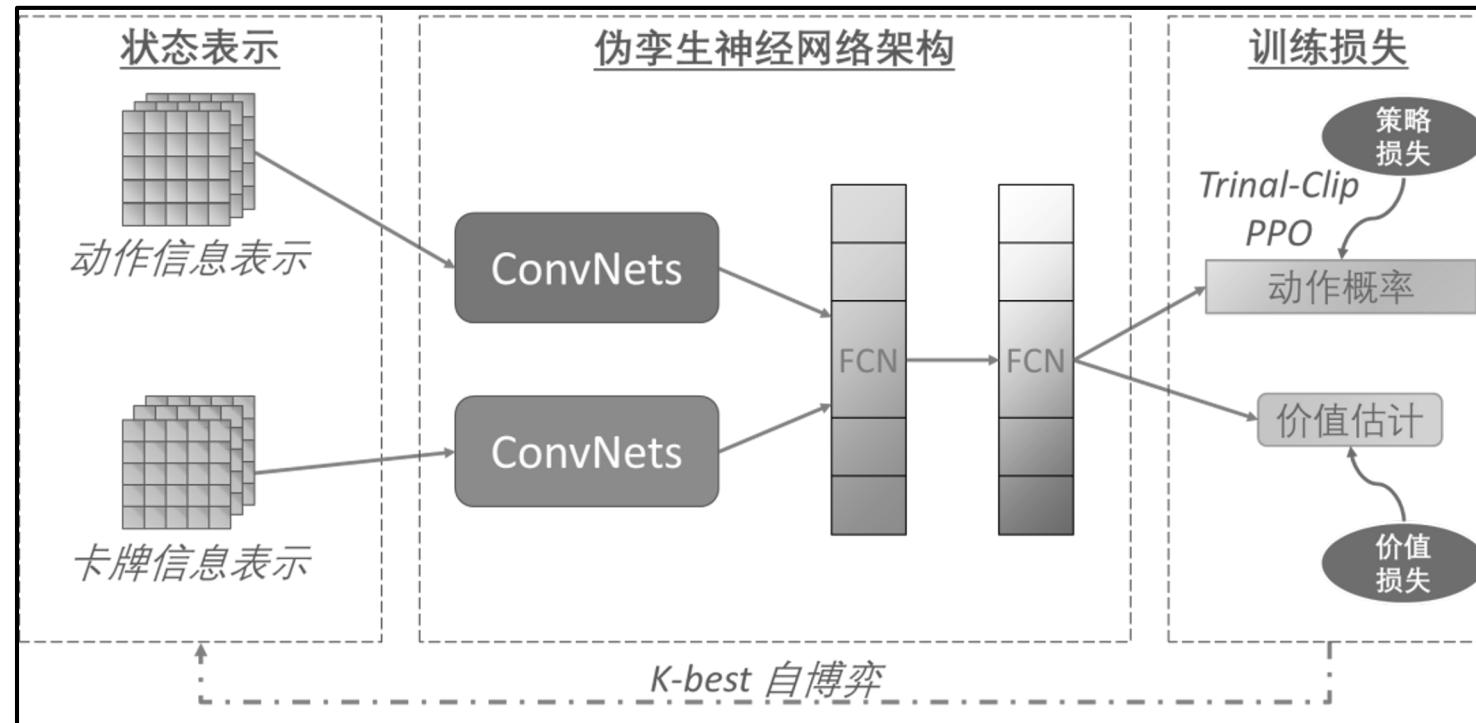


# 德州扑克 1v1

- University of Alberta, *Science*, 2017  
DeepStack: 第一个在一对无限注德州扑克中击败职业扑克玩家的计算机程序
- CASIA, AlphaHoldem, AAAI 2022  
Distinguished Paper, 三天的自博弈学习后战胜了Slumbot和DeepStack



不完全信息博弈



# 德州扑克 1v5

# 德州扑克1v5

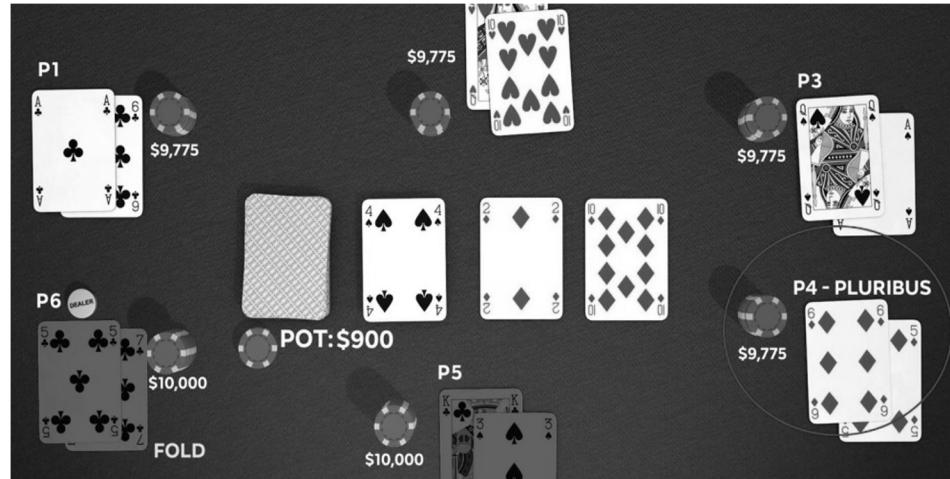
Pluribus, CMU,  
Science, 2019.07

## 技术突破：

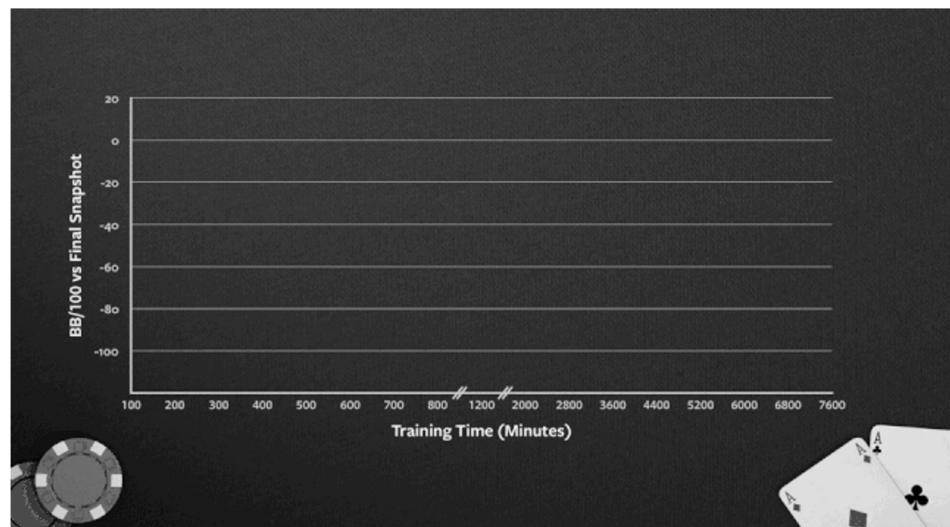
1. 率先在多人制德州扑克战胜职业选手
  2. 较低成本的计算资源（64核CPU服务器）
  3. 解决不完全信息下的多人零和博弈问题

## 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

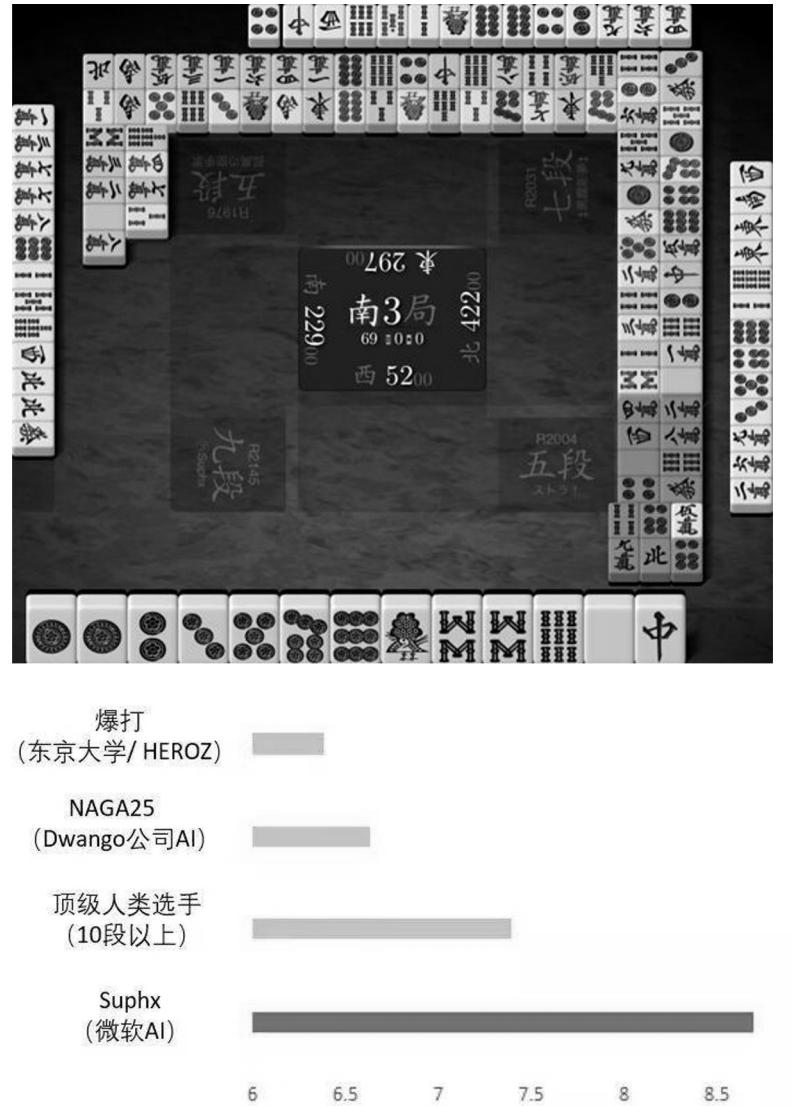
# 麻将AI

# 麻将AI

Suphx, 10段, 微软, 2019.08

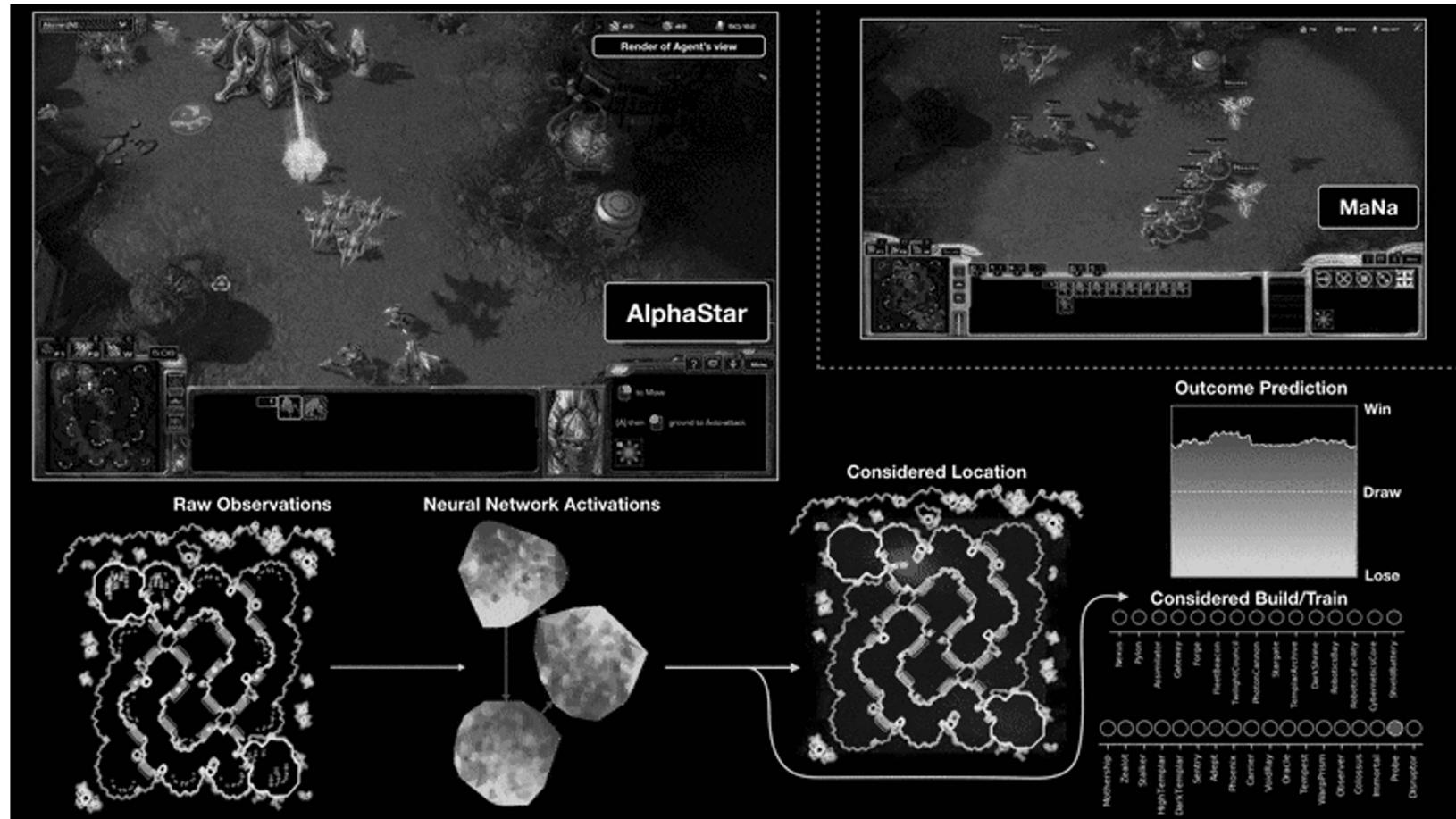
## 技术路线:

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。



# AlphaStar

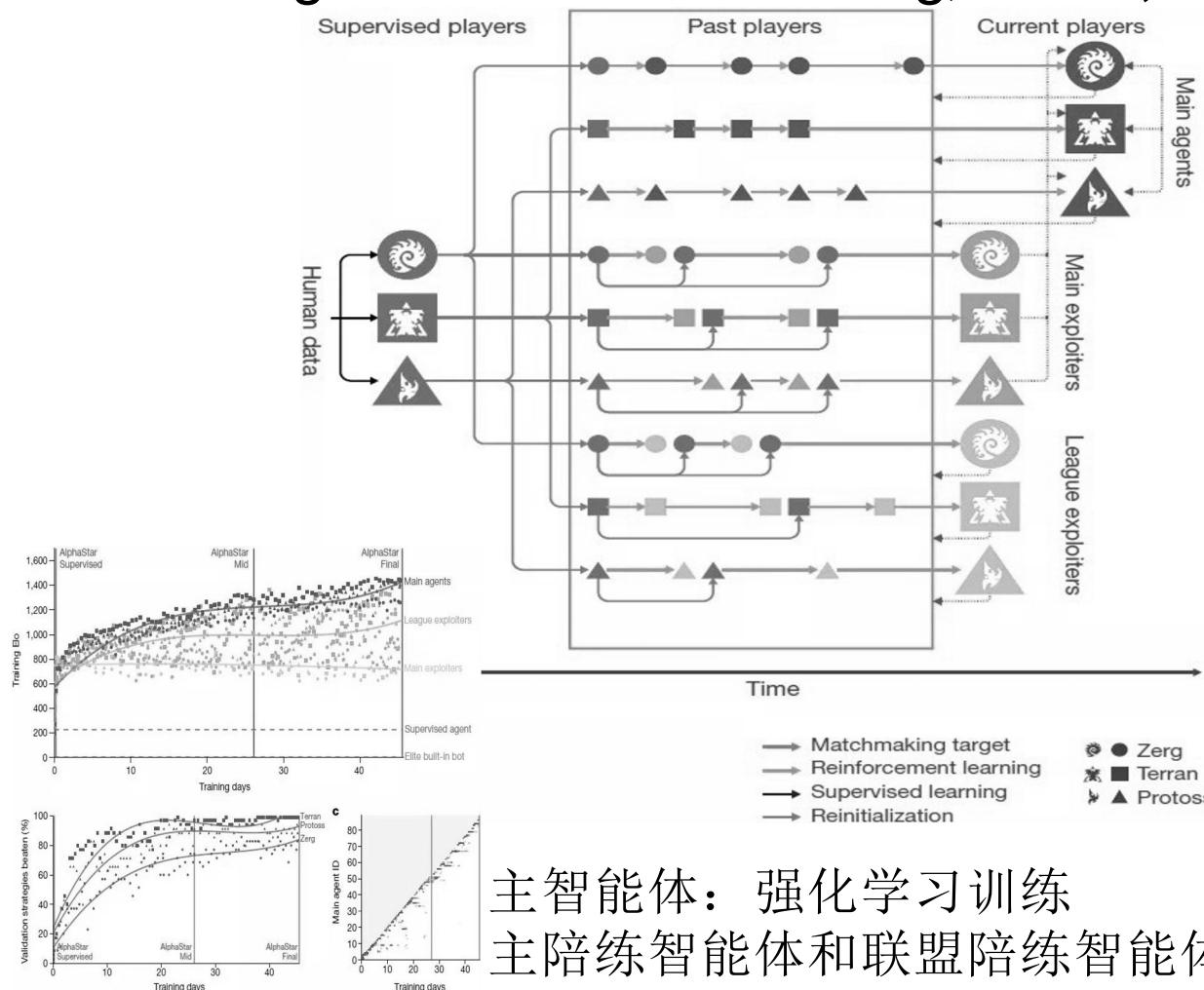
- 2019 年 1 月, DeepMind 公布了开发的 AlphaStar 与人类职业选手录像与比赛, 最终 10:1 获胜\*



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

# AlphaStar

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



主智能体：强化学习训练  
主陪练智能体和联盟陪练智能体

## 问题

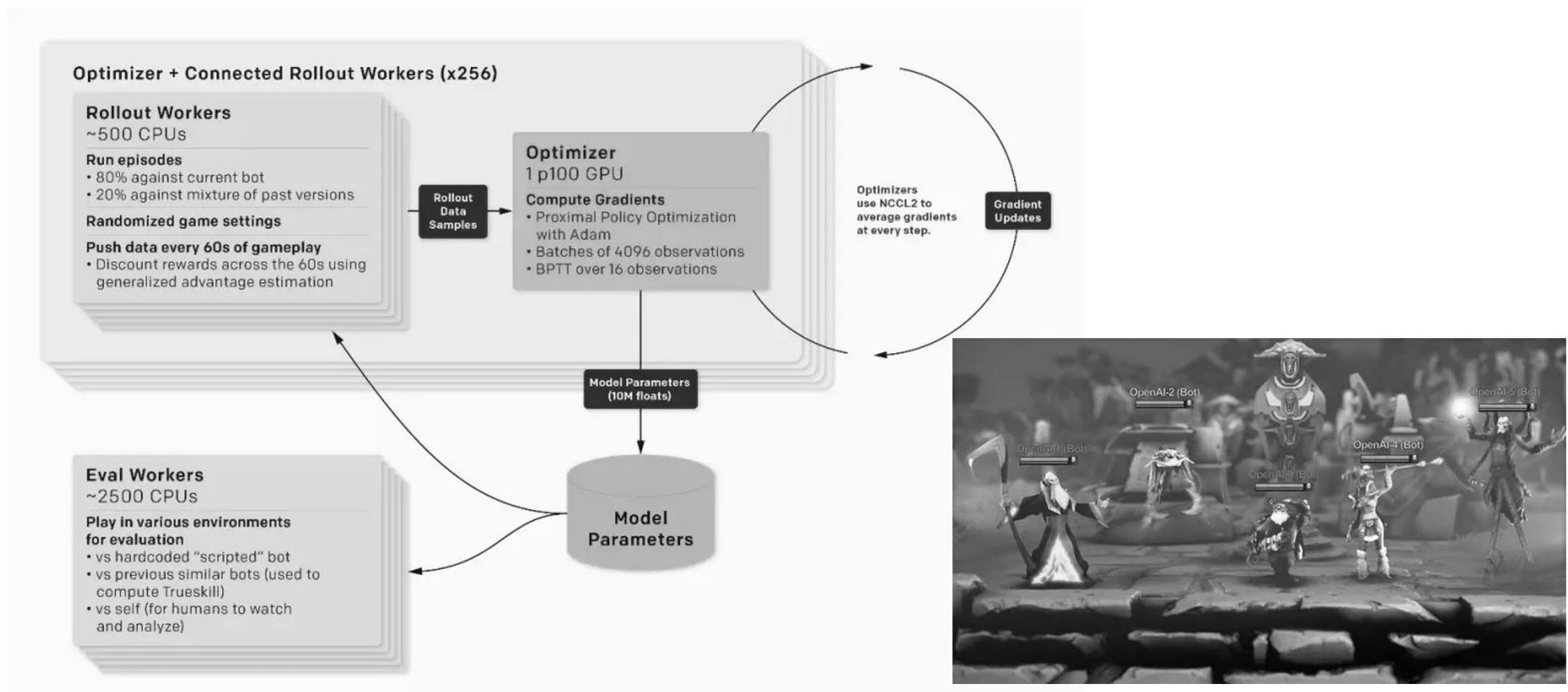
- 没有单一最佳策略
- 非完全信息
- 有蝴蝶效应
- 实时决策
- 巨大动作空间
- 三种不同种族

## 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%

# OpenAI Five

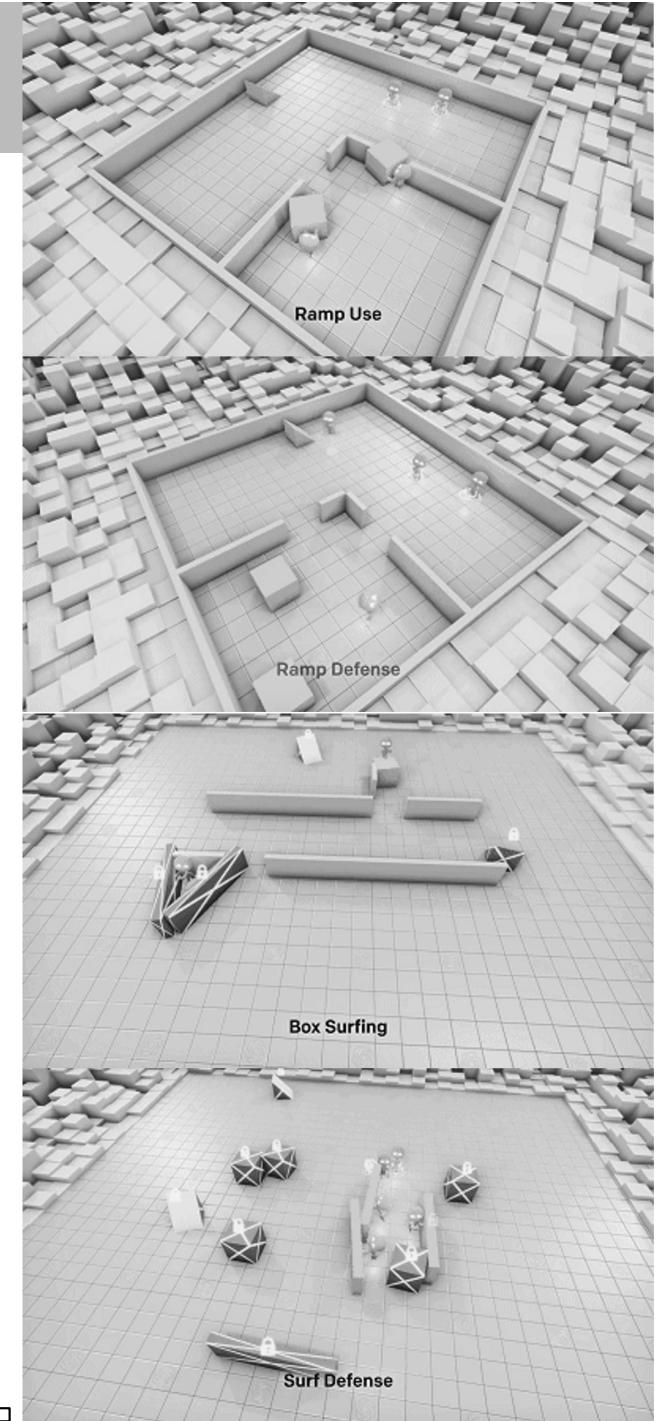
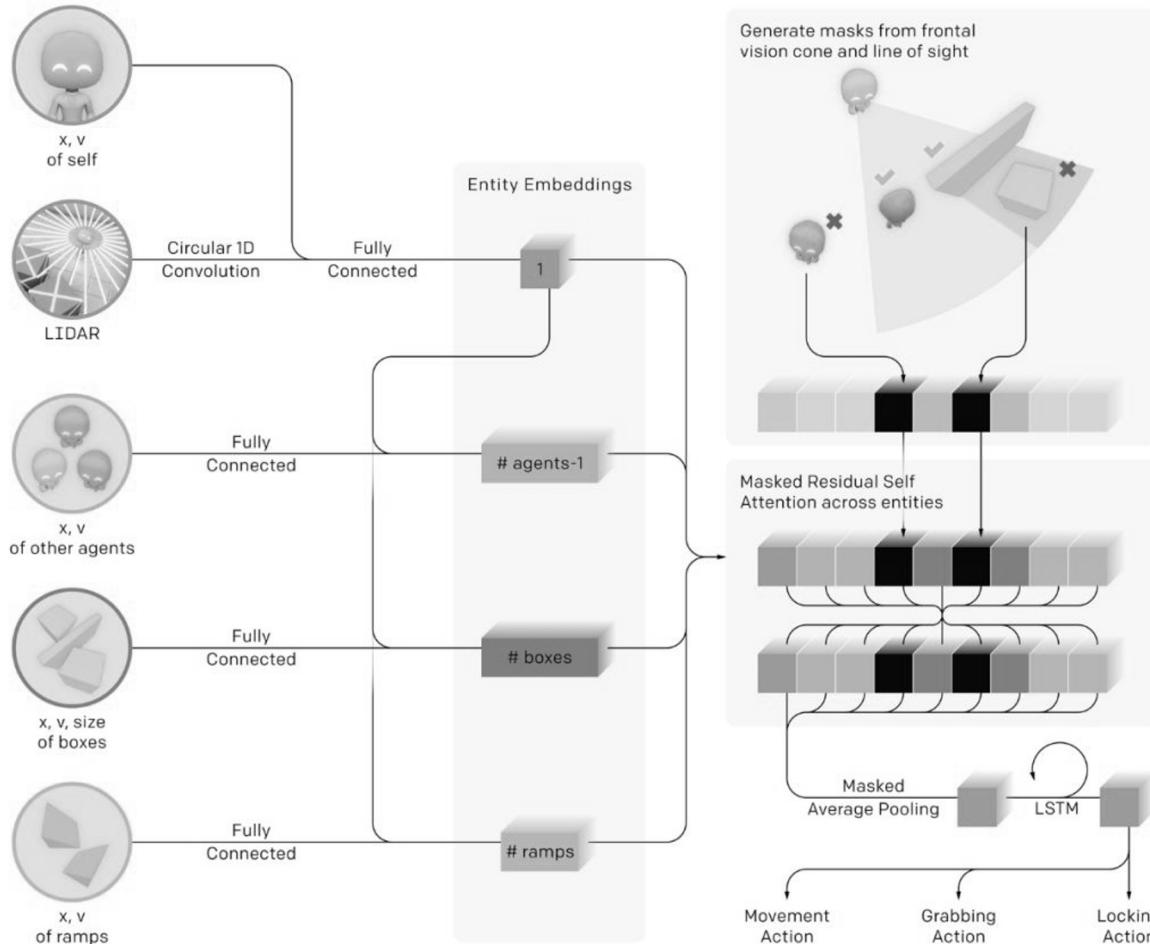
- 2019 年 4 月, OpenAI Five 人工智能系统迎战去年 Ti8(第八届 Dota2 国际邀请赛) 冠军 OG 战队, 最终 2:0 获胜
- 近端策略优化(PPO), 公开赛中获得了 99.4% 胜率



# OpenAI--捉迷藏

2019, 通过自动课程学习复杂的策略和反策略

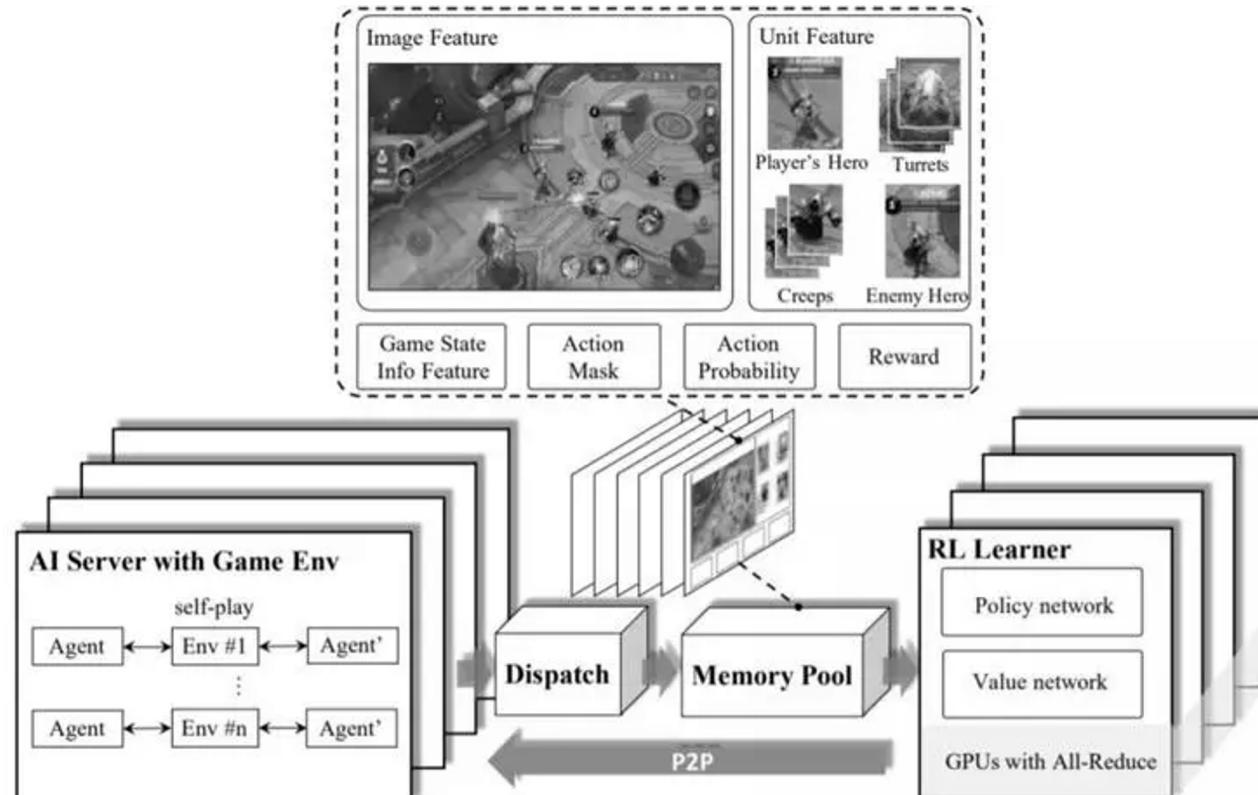
Policy Architecture



# 腾讯绝悟

■ 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进行批训练时的收敛性。

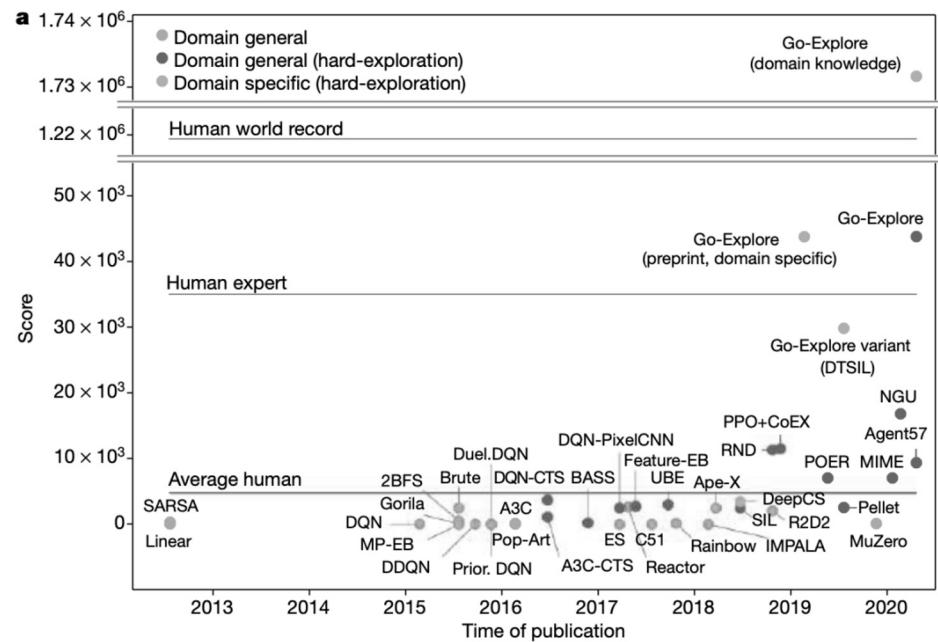
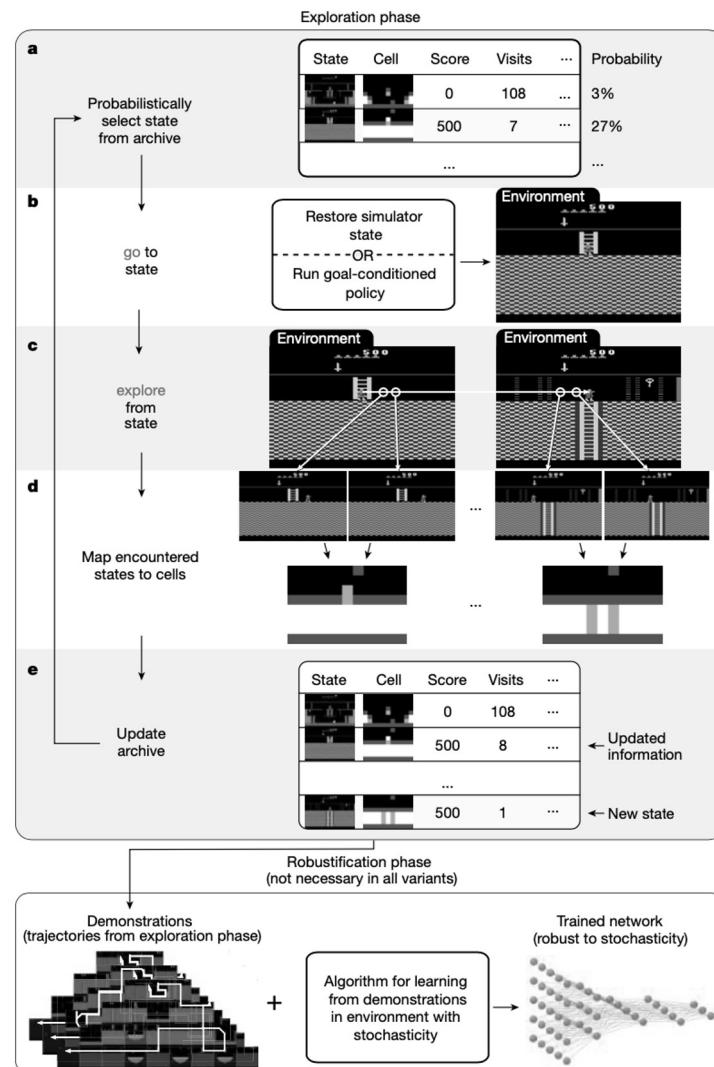
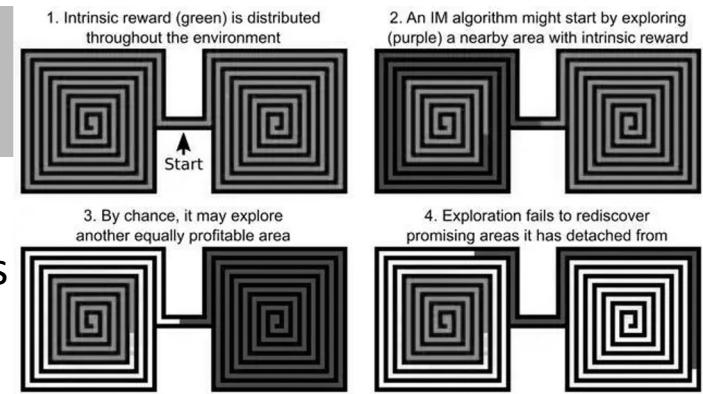


# OpenAI&Uber- Go Explore

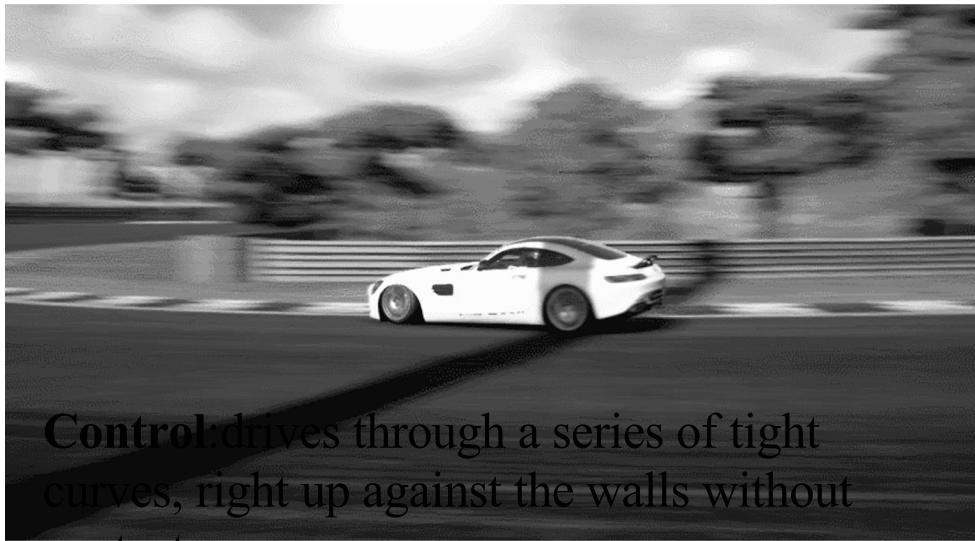
2021, Nature, the main impediment to effective exploration

Detachment: forgetting how to reach previously visited states

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



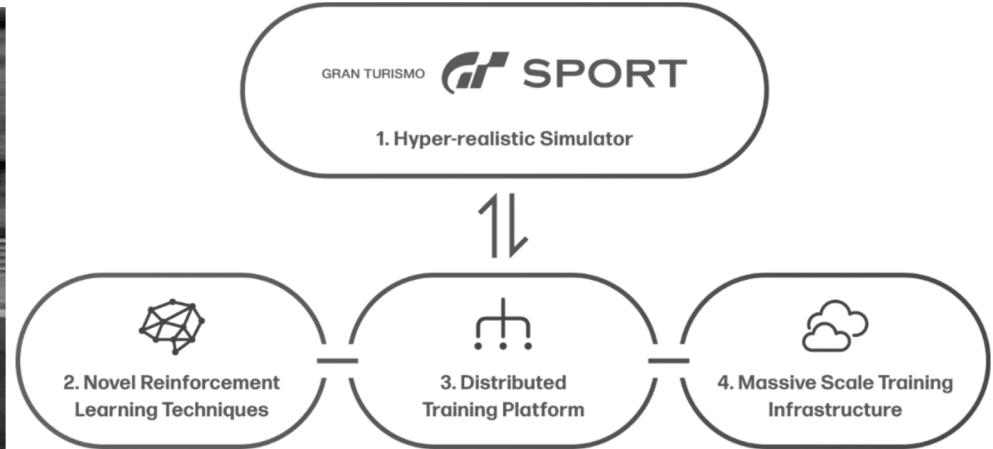
# GT-Sophy 2022 Nature



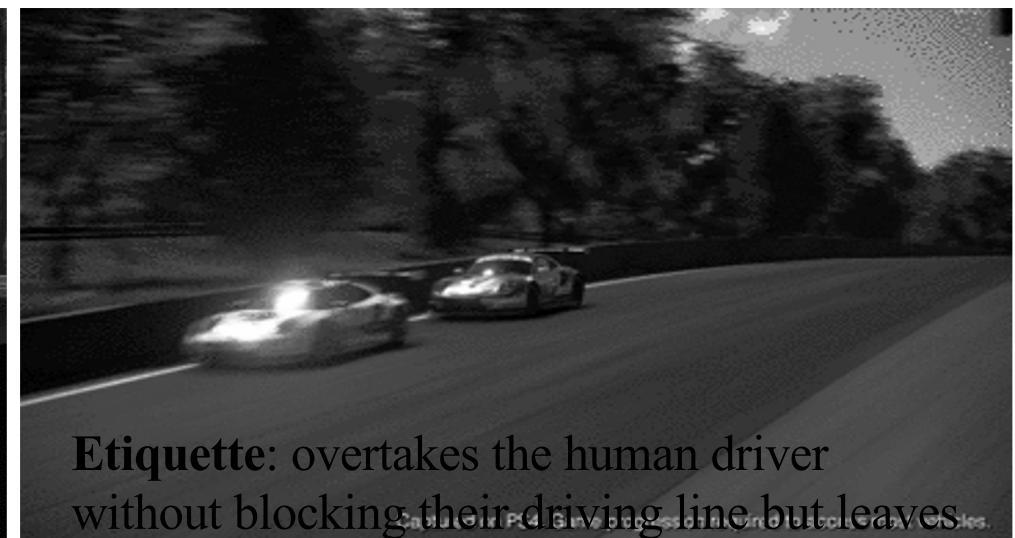
Control: drives through a series of tight curves, right up against the walls without contact.



takes  
antage of  
the space  
available  
on the track.



Quantile-Regression Soft Actor-Critic (QR-SAC)

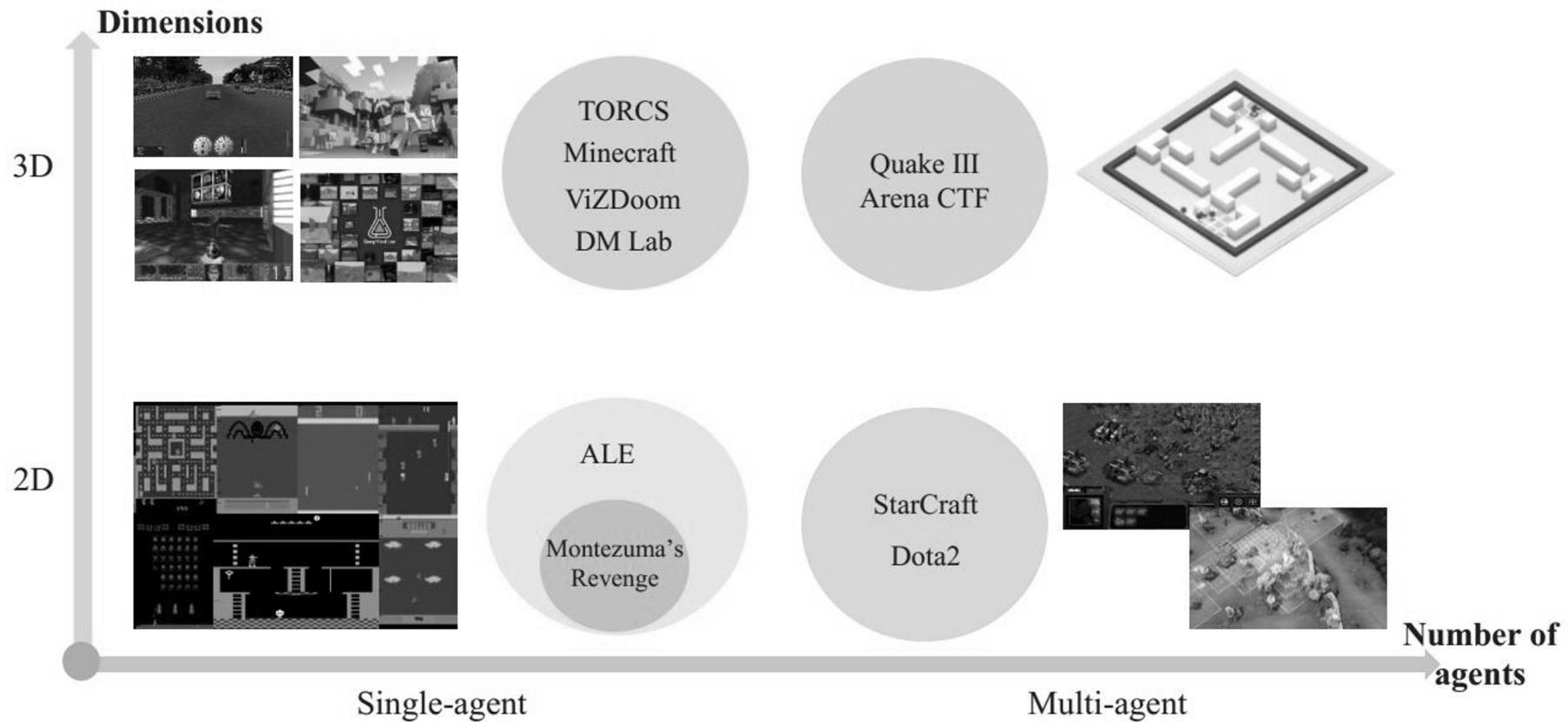


Etiquette: overtakes the human driver without blocking their driving line but leaves spaces

# 强化学习典型应用

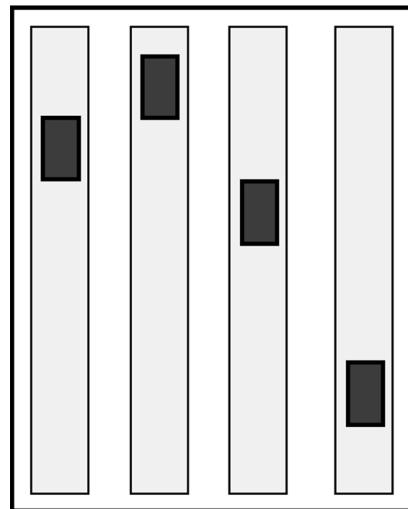
# 游戏AI（第9讲作业1+第11讲）

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



# Elevator Dispatching

- 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)

# 小狗机器人

## ■ 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



# 双足机器人

- 双足机器人行走 (actor-critic + eligibility trace)<sup>\*</sup>



# 遥控直升机

遥控直升机倒立飞行 \*

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



# 乒乓球机器人

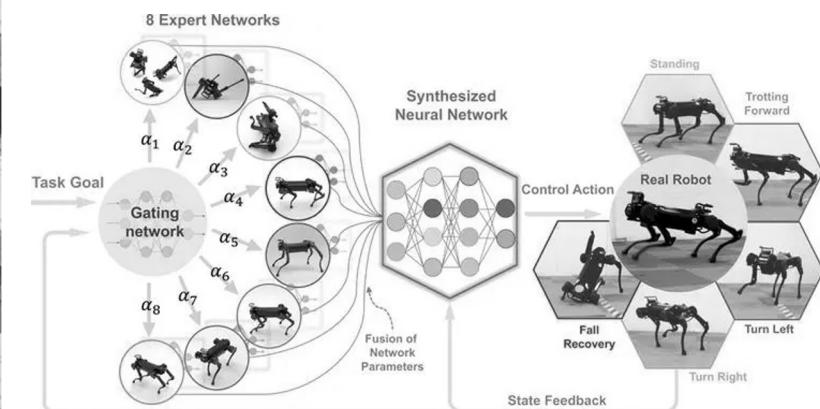
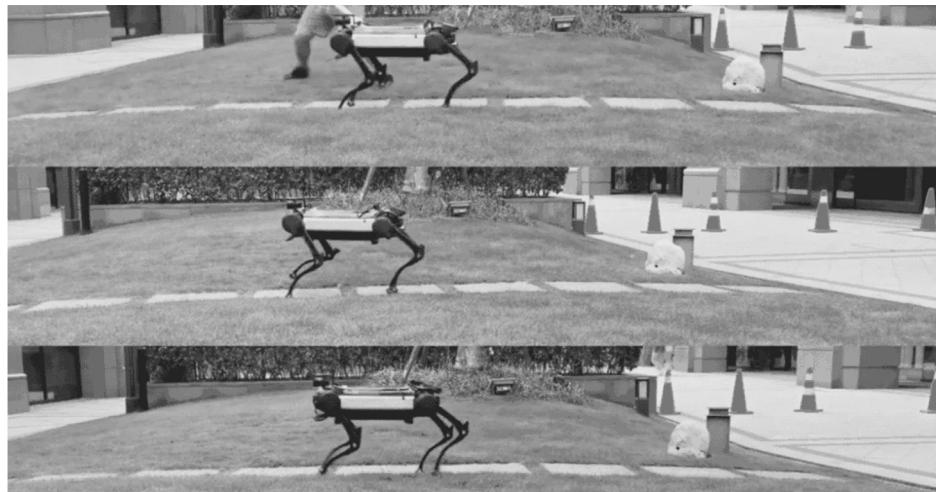
a Barrett WAM arm uses the mixture of motor primitives (MoMP) algorithm to learn successful hitting movements in table tennis using imitation and reinforcement Learning.\*



# 绝影—爱丁堡&浙江大学

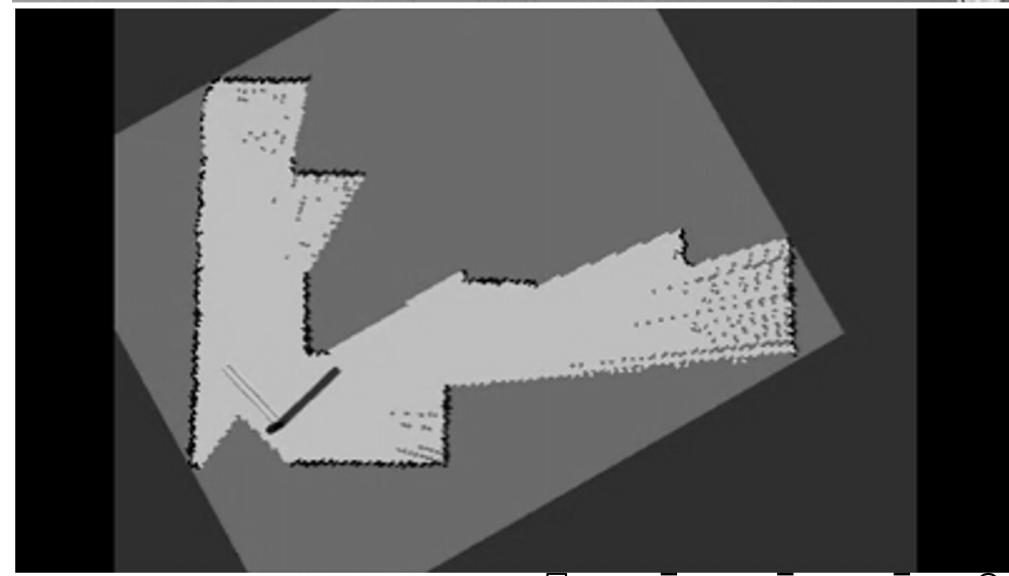
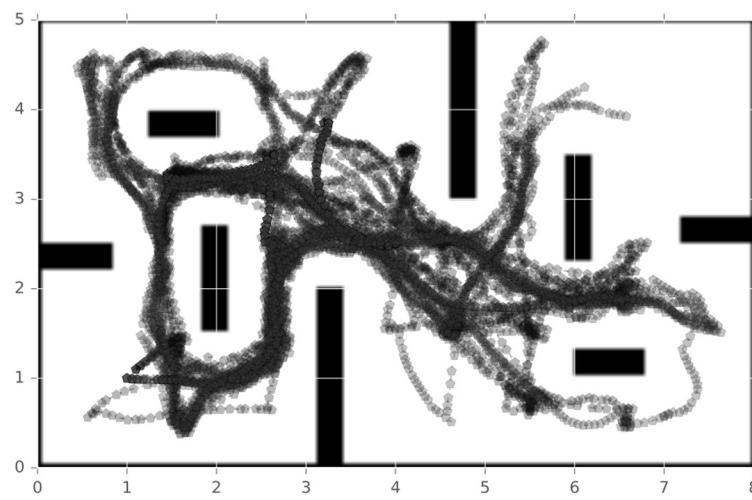
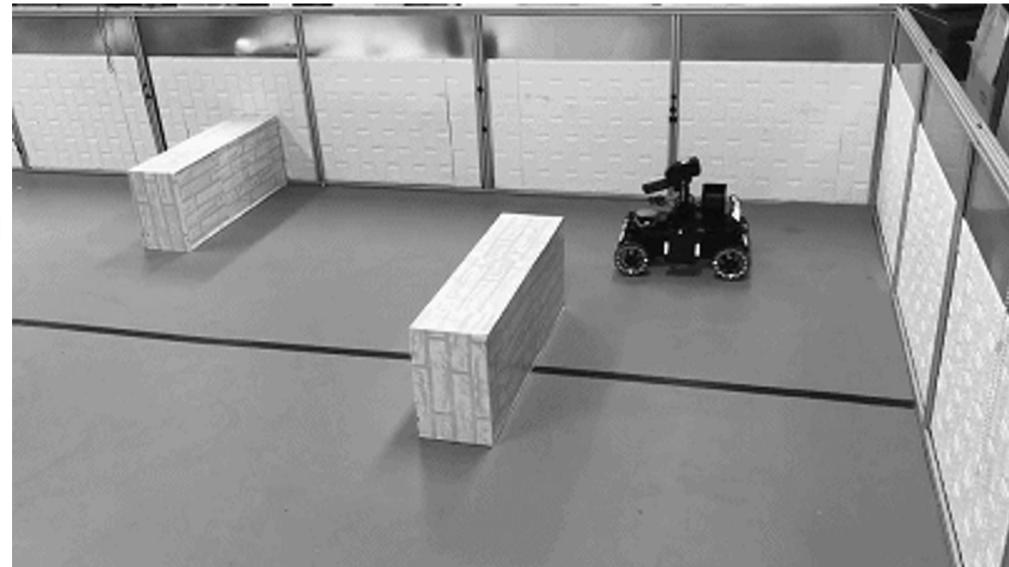
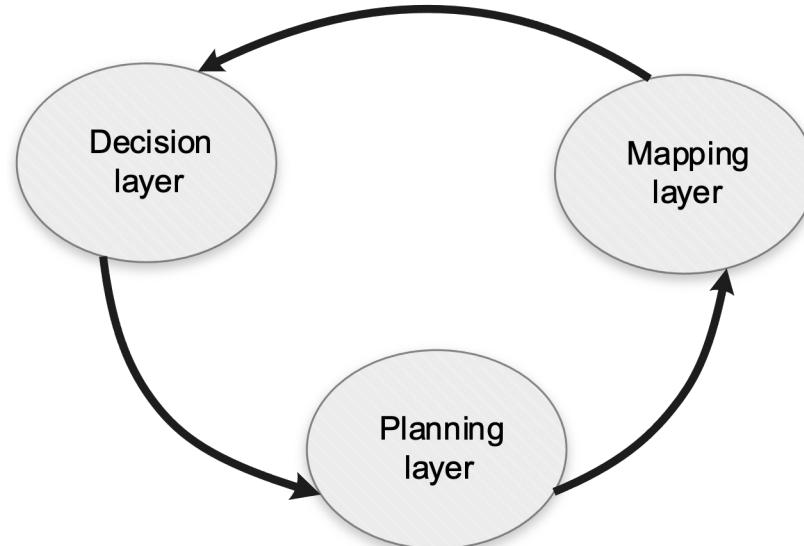
# MELA: Multi-expert learning of adaptive legged locomotion (Science Robotics 2020 Cover)

- ✓ First initialized by a distinct set of pretrained experts, each in a separate deep neural network (DNN).
  - ✓ A hierarchical deep reinforcement learning the combination of these DNNs using a gating neural network (GNN)



# 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。



# Robomaster机器人对抗赛(第9讲作业2)

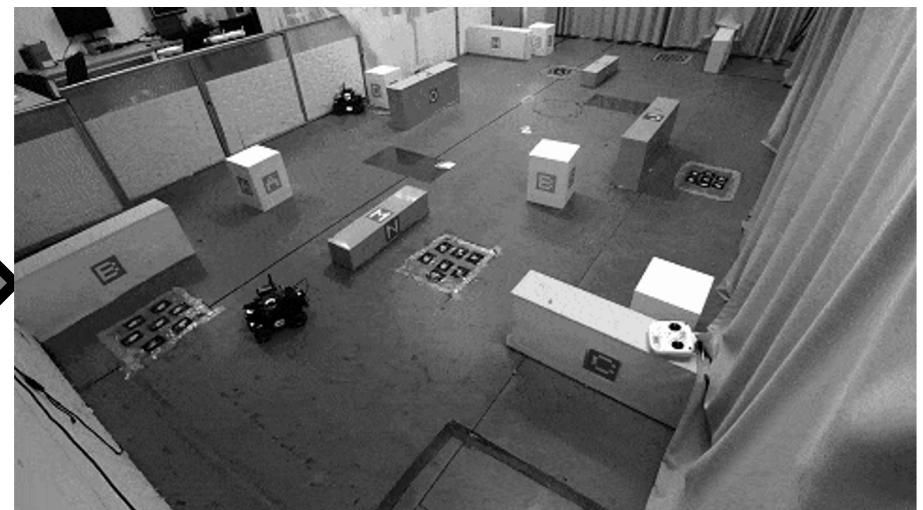
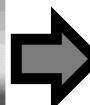
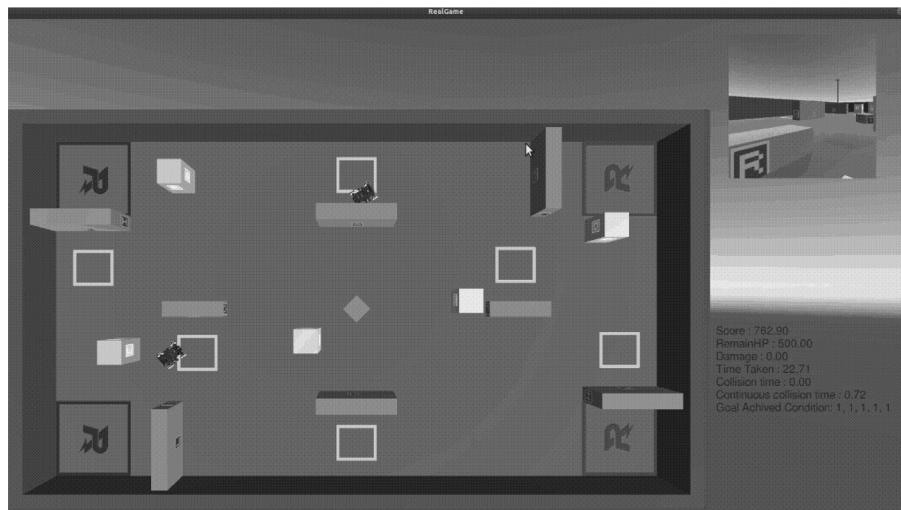
## RoboMaster Sim2Real Challenge

### □ Track 1: complete information-based task

- The position and the direction of the EP robot and the defensive robot;
- The position of five goals, and image, lidar and some other necessary information.

### □ Track 2: image-based task

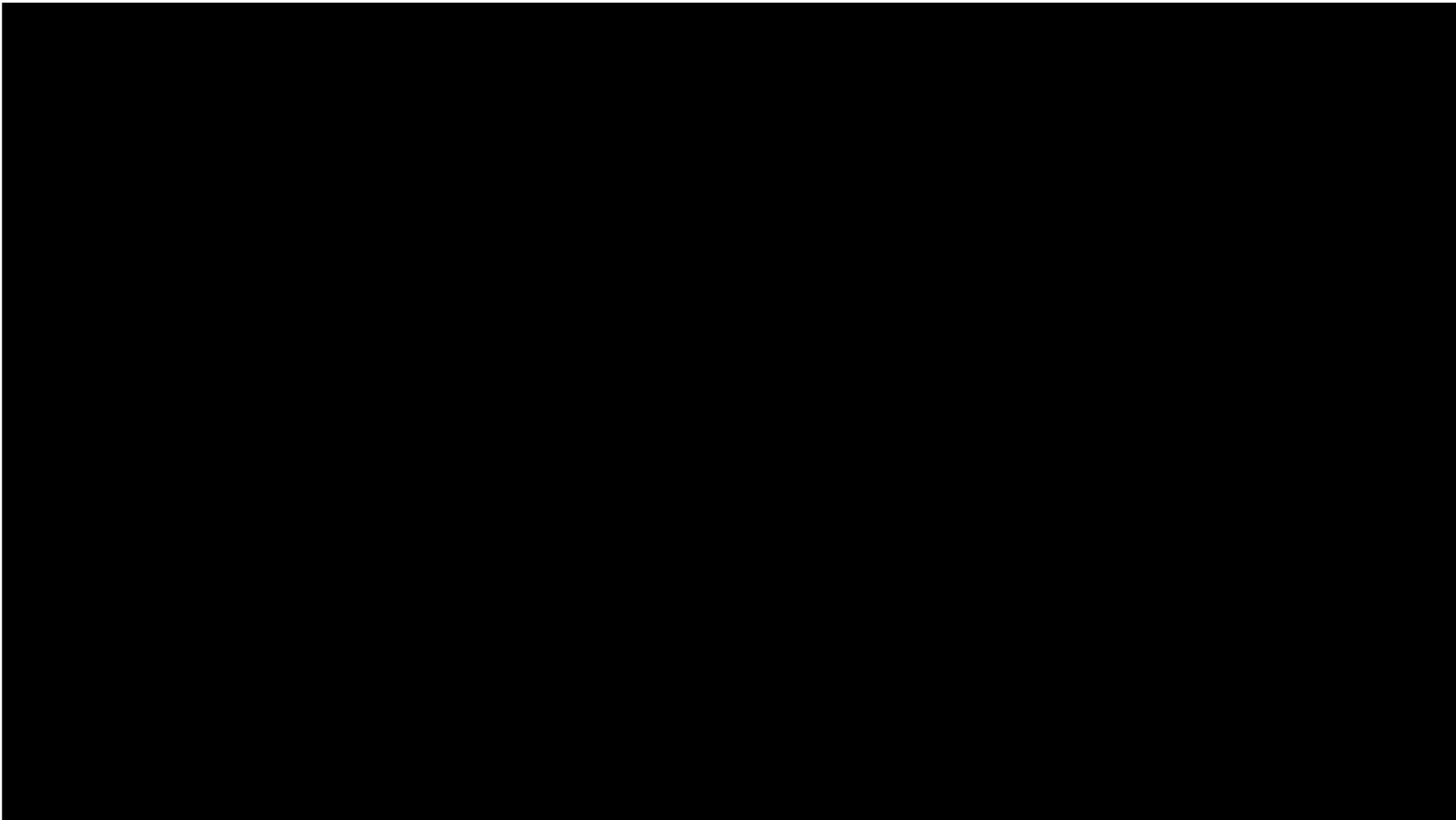
- No position and direction of the robots, but other information as image and lidar.
- ✓ Output the speed (x-direction, y-direction, yaw) and fire command to control the robot.



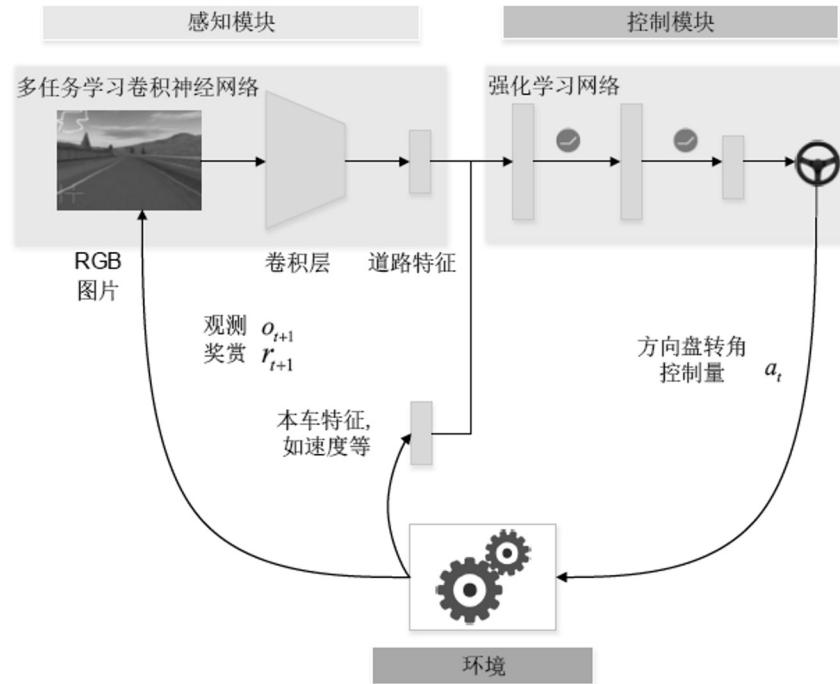
[https://ieee-cog.org/2022/cog\\_sim2real/index.html](https://ieee-cog.org/2022/cog_sim2real/index.html)

# 智能驾驶 - 实车

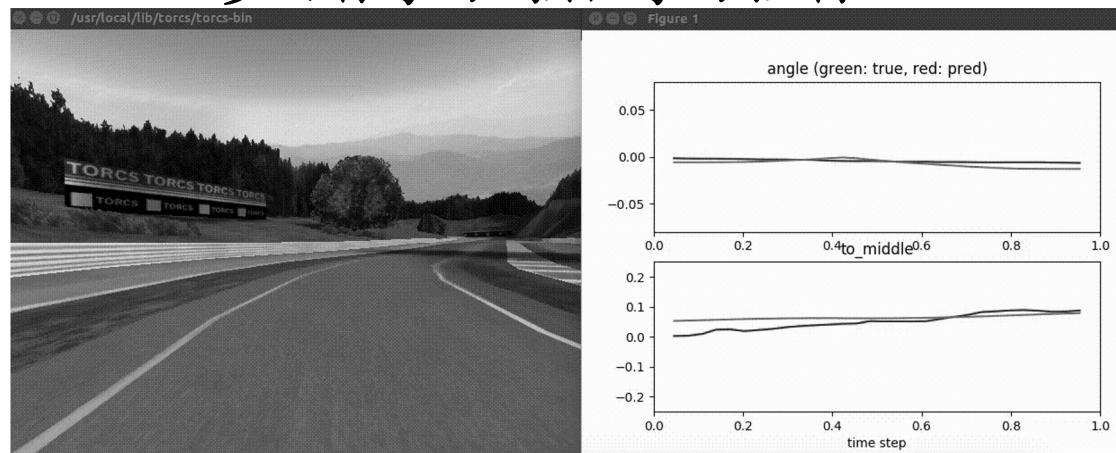
剑桥大学创业公司 wayve—The first example of reinforcement learning on-board an autonomous car<sup>\*</sup>



# 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

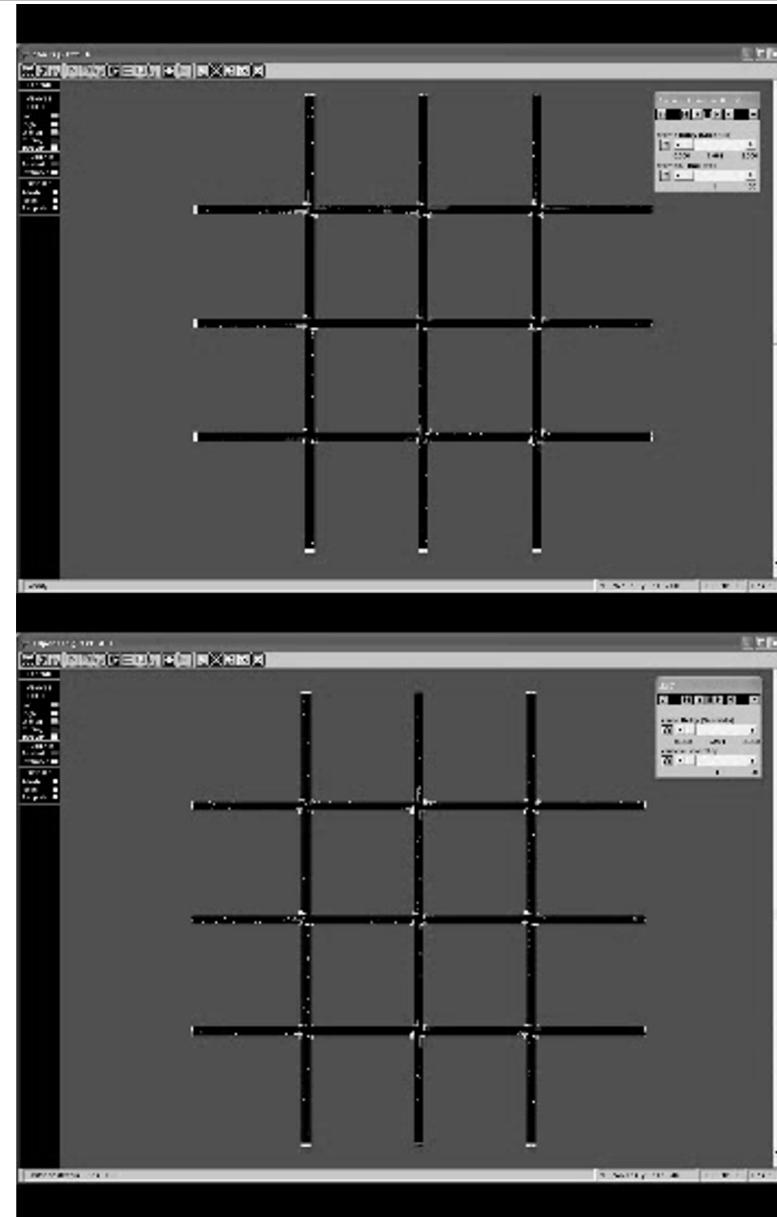
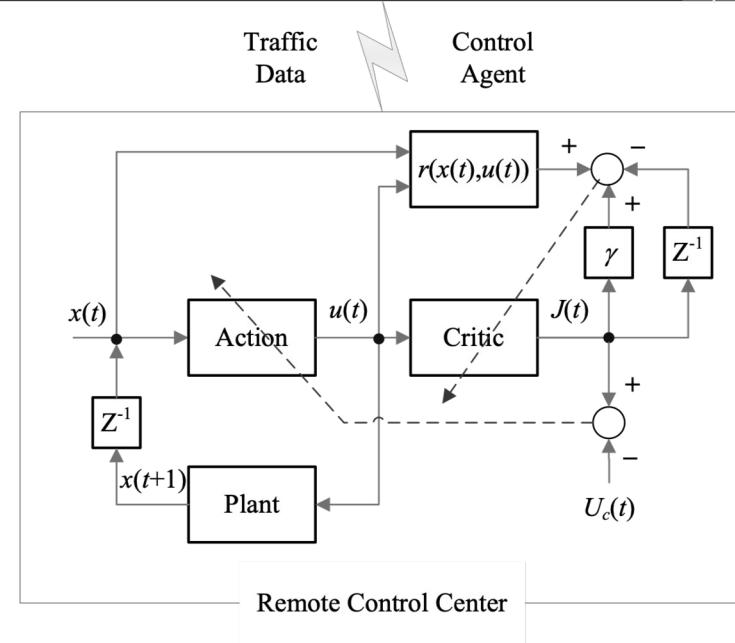
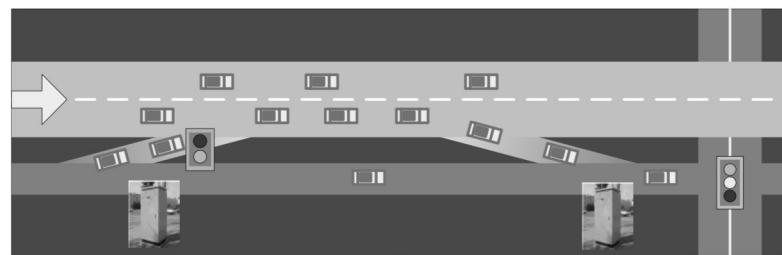


## 传统控制方法

## ADP 方法

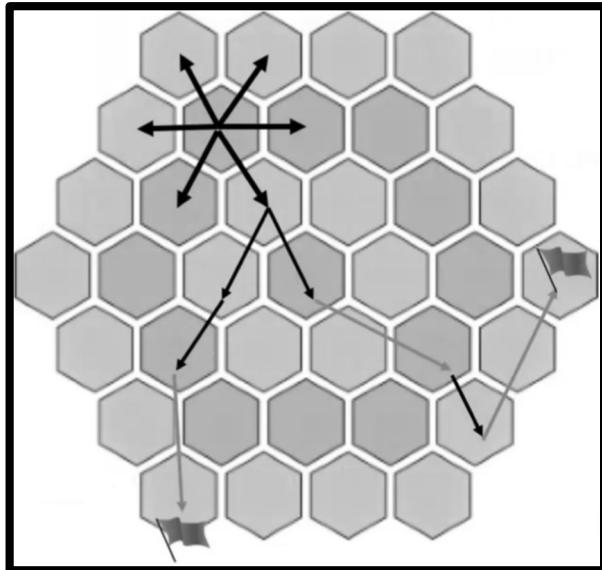
# 交通信号控制

➤ 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；

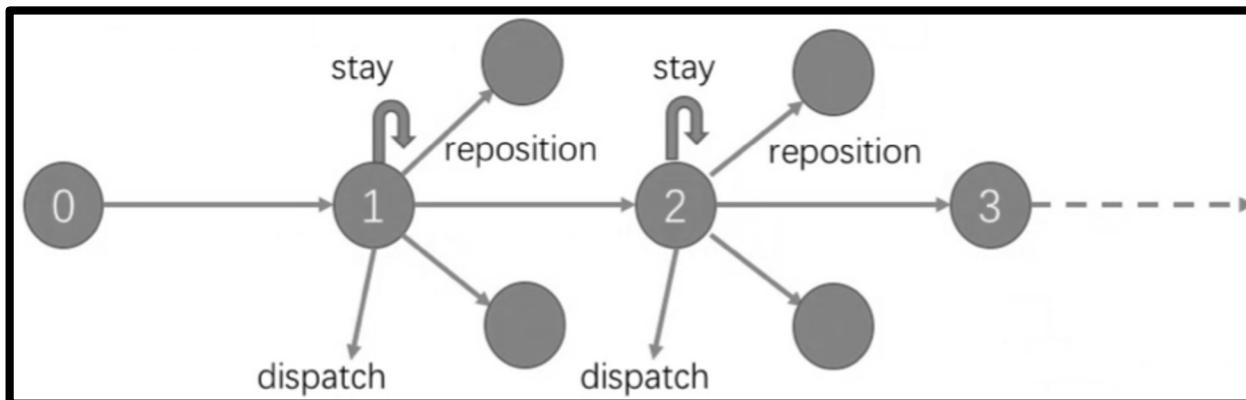


# 智慧城市

➤ 应用需求：派车、调度、物流、供应链、交通管理、智能电网等



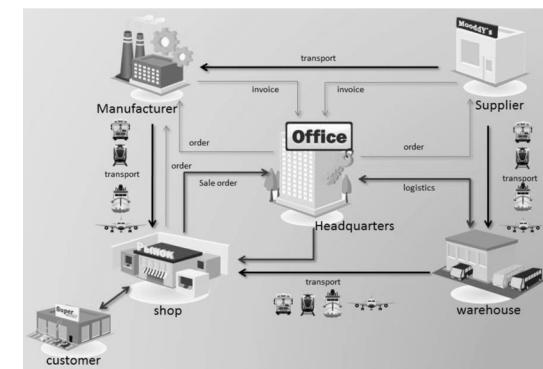
状态空间：六角形格  
动作空间：当前格或相邻格的目的地，任意格的目的地（长距离任务），转换到派单任务  
环境模型：如下图  
特点：供需随机性变化，大规模智能体协调



滴滴城市网约车调度



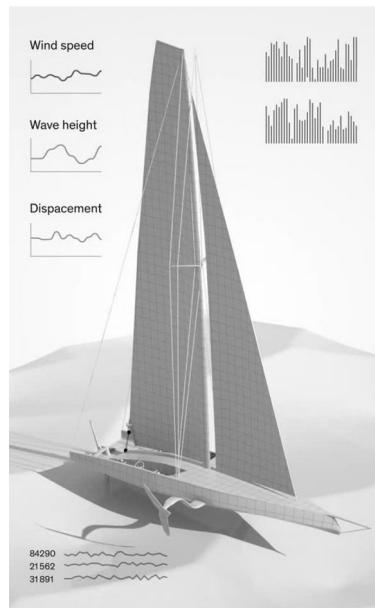
场景	Well Define Solver	Data Driven	结果
仓内拣选	Batching	Learn to Define Optimization (Embedding)	拣选时间降低10%
Last Mile	DVRP	Can we learn the dispatching rule?	易于接受
包材推荐	Bin Packing	Multi-task selected learning	包材成本降低4.5%到6.6%
AI大脑	MIP	Offline Training + Online Prediction	荷兰109万月；法国、波兰等362万月
智能调度	Heuristic	Adaptive	Improved convergence speed and quality of solutions



# Sailing Boat Design

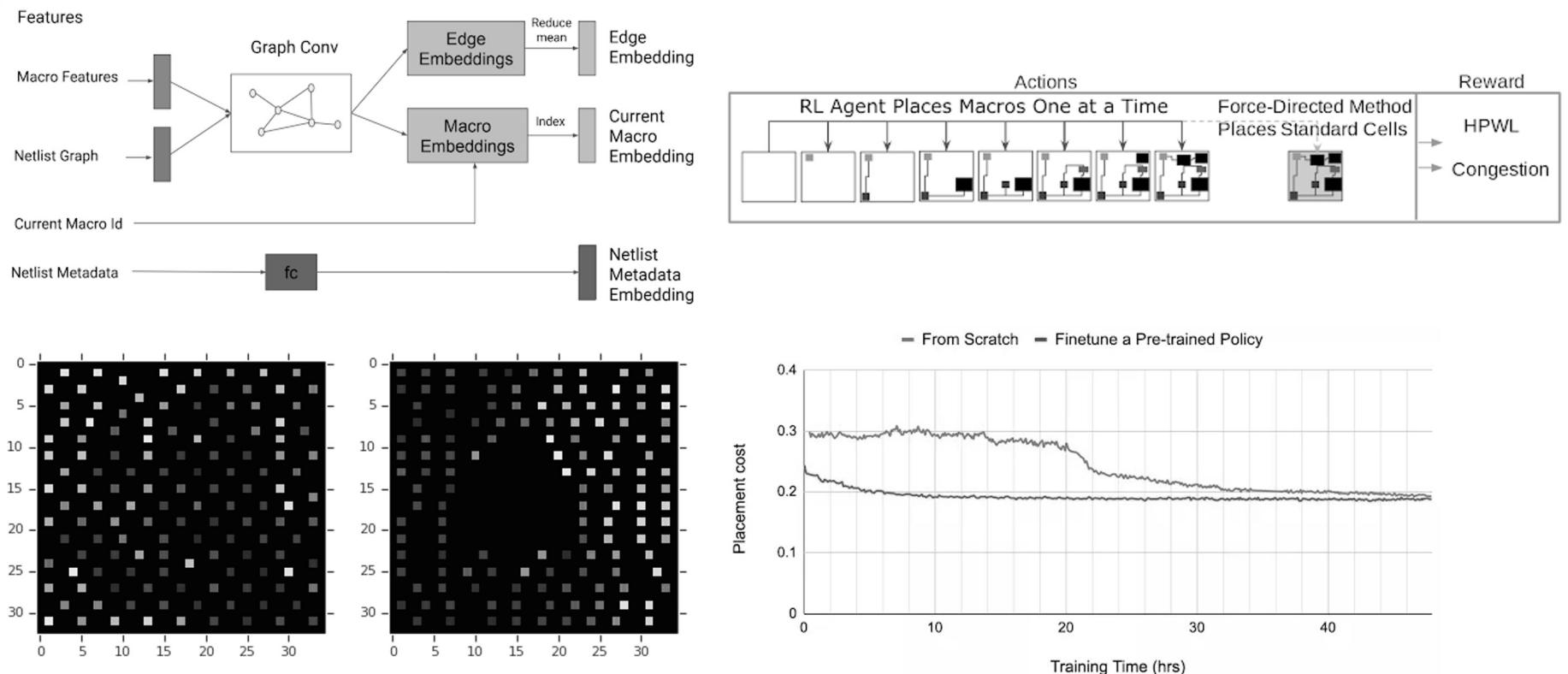
The world's best sailor in 2021 America's Cup, March 2021.

- ✓ Deep reinforcement learning accumulates experience and continually refines its skill.
- ✓ It responds to varying conditions, adjusts 14 different boat controls accordingly, and has to understand the trade-offs between immediate and long-term goals.
- ✓ A network of more than 1,000 AI agents running in parallel, the agents quickly reach a level of mastery to outperform world-champion sailors in the simulator and begin testing design concepts for the team.



# 芯片设计

- ✓ Nature 2021
- ✓ Ground representation learning in the supervised task of predicting placement quality, to enable the RL policy to generalize to unseen blocks.
- ✓ An end-to-end method generates placements in under 6 hours, whereas the strongest baselines require human experts in the loop and take several weeks.



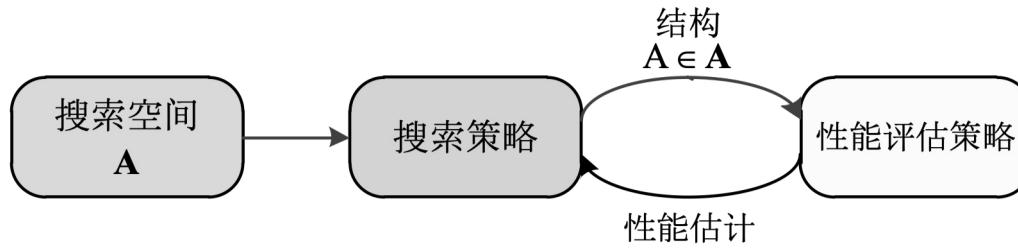
- 量化投资：采用统计、计算机、人工智能等技术，来实现复杂的金融市场的量化决策，选择合适的投资目标，提升投资成效。

- 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)
- .....

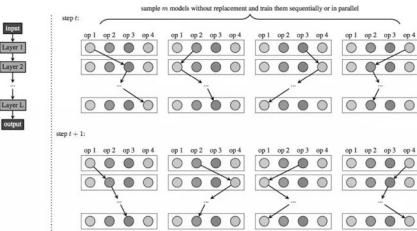


# AutoML - 神经架构搜索

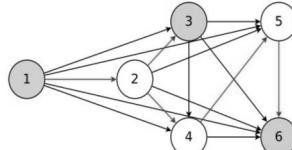


权重加权求和  
梯度求解  
1块卡1天

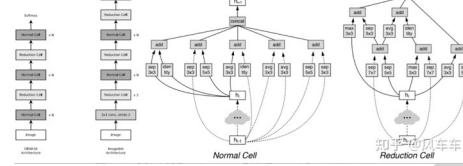
训练超网络



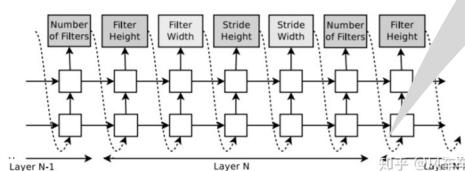
Weights sharing  
1块卡0.5天



500块GPU跑了4天  
Cifar10/ImageNet



800块GPU跑了快一个月



NAS  
Google  
ICRL'17

ENAS

Cell  
Google  
ICML'18

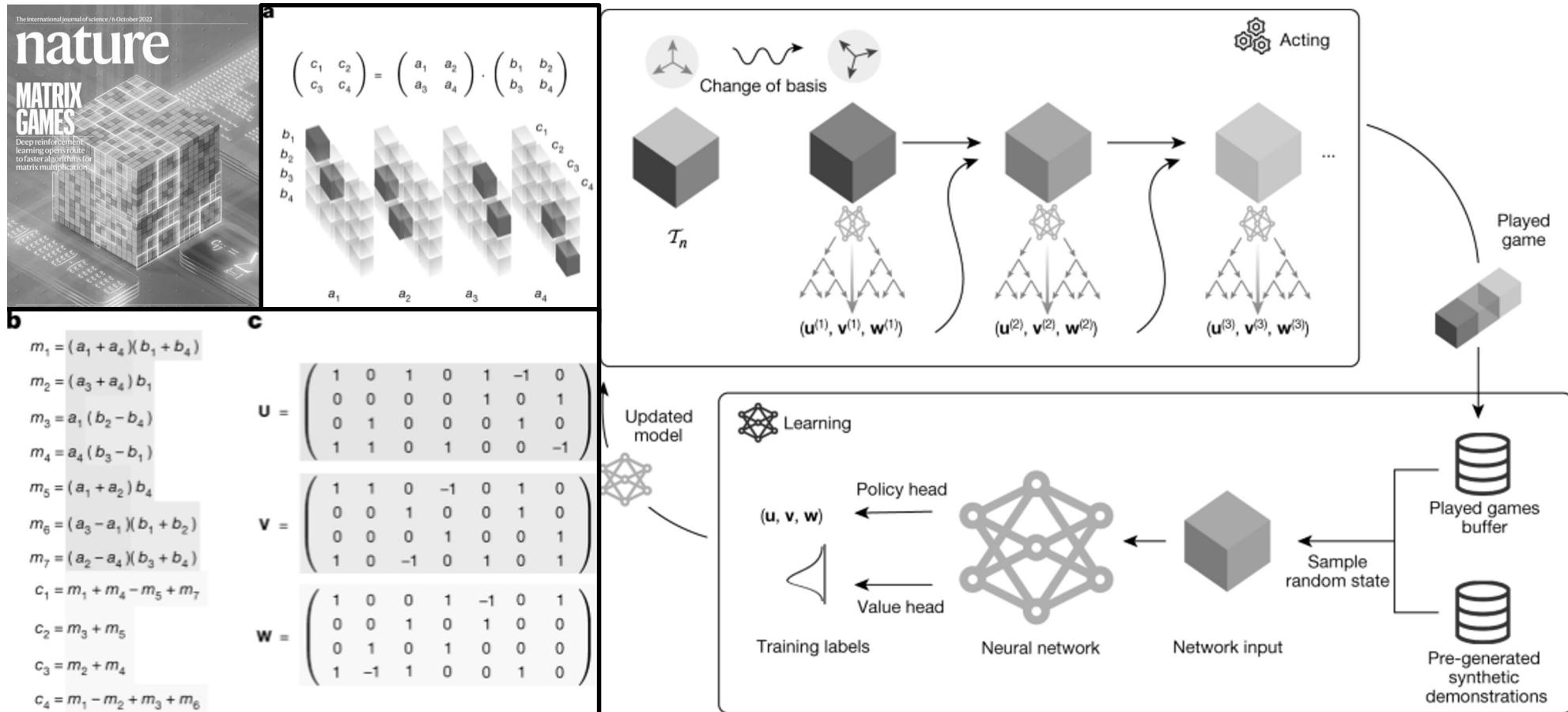
DARTS  
CMU+Google  
ICLR'19

One-shot

旷世、小米等

# 矩阵运算

## DeepMind: AlphaTensor@Nature Oct.

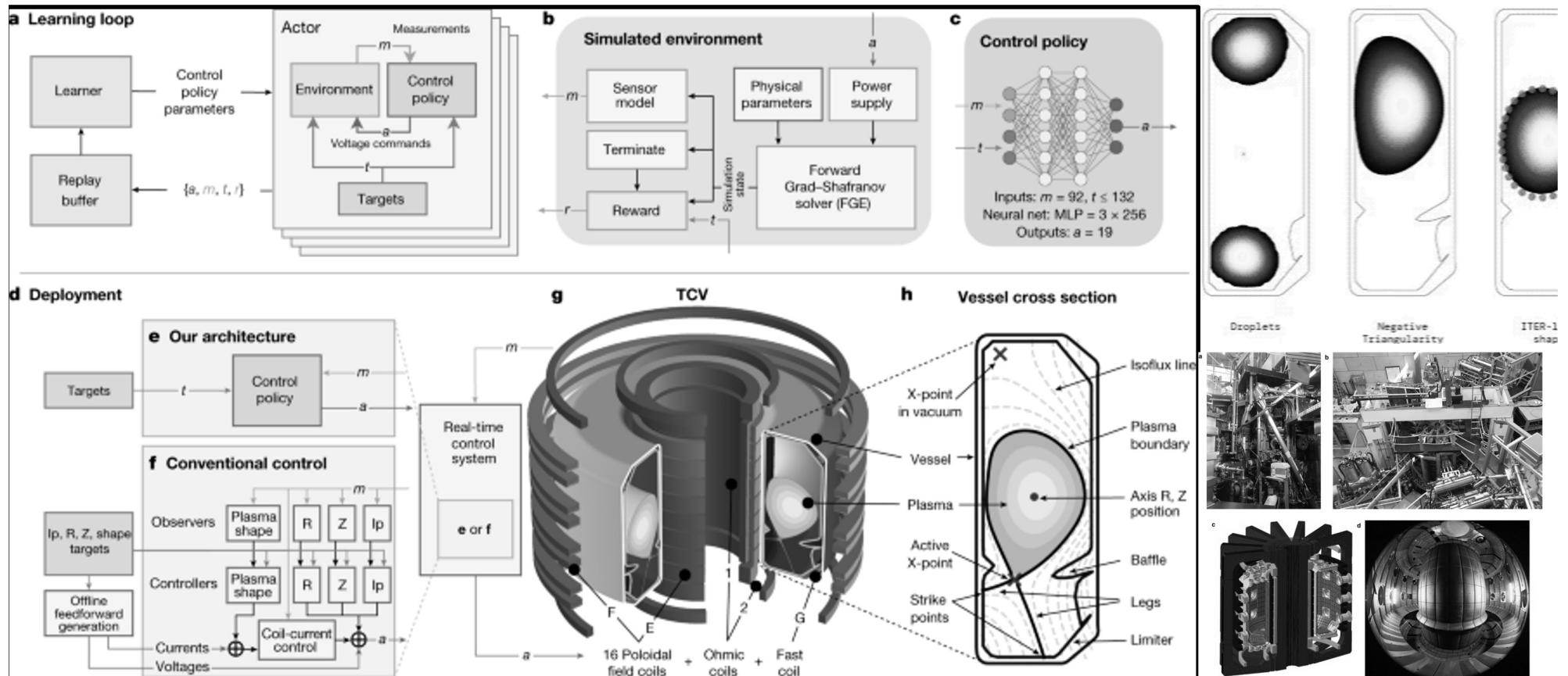


Amazing applications of AlphaZero!  
To save about 10% computation resources.

## 核聚变等离子体控制

2022, Nature

- ✓ 超过1亿摄氏度的环境下加热氢成等离子体的状态
  - ✓ 一次控制全部19个线圈，并精确操纵等离子体自主呈现各种形状



# OpenAI: ChatGPT@Dec. 2022

## Reinforcement Learning from Human Feedback (RLHF)

DeepMind Sparrow: reinforcement learning based on people's feedback, using the study participants' preference feedback to train a model of how useful an answer is.

Step 1

Collect demonstration data and train a supervised policy.

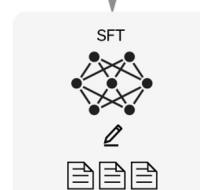
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



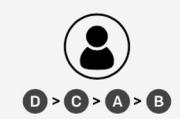
Step 2

Collect comparison data and train a reward model.

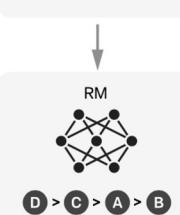
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



[www.oncoscience.us](http://www.oncoscience.us)

Oncoscience, Volume 9, 2022

Research Perspective

## Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective

ChatGPT Generative Pre-trained Transformer<sup>2</sup> and Alex Zhavoronkov<sup>1</sup>

<sup>1</sup>Insilico Medicine, Hong Kong Science and Technology Park, Hong Kong

<sup>2</sup>OpenAI, San Francisco, CA 94110, USA

Correspondence to: Alex Zhavoronkov, email: alex@insilico.com

Keywords: artificial intelligence; Rapamycin; philosophy; longevity medicine; Pascal's Wager

Received: December 14, 2022

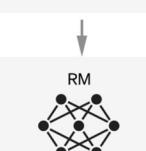
Accepted: December 15, 2022

Published: December 21, 2022

Copyright: © 2022 Zhavoronkov. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#) (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The policy generates an output.

Once upon a time...



The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

$r_k$



From algorithm design to theory analysis, from simulation games to real applications

# 强化学习基本元素

# 状态 State

- 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 Action

- 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 Policy

- 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)

# 状态转移 State Transition

- 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 Reward

- 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

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

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

# 奖励举例

- 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：

■ 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  $r = +1$

■ 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： $r = -\text{dist}(\text{robot}, \text{target})$

■ 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 Return

- **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
  - $\gamma$  越接近 0，回报越是“目光短浅”
  - $\gamma$  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 Value

- 价值：智能体在当前状态下回报的期望  $V$

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

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

# 最优策略和最优价值

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

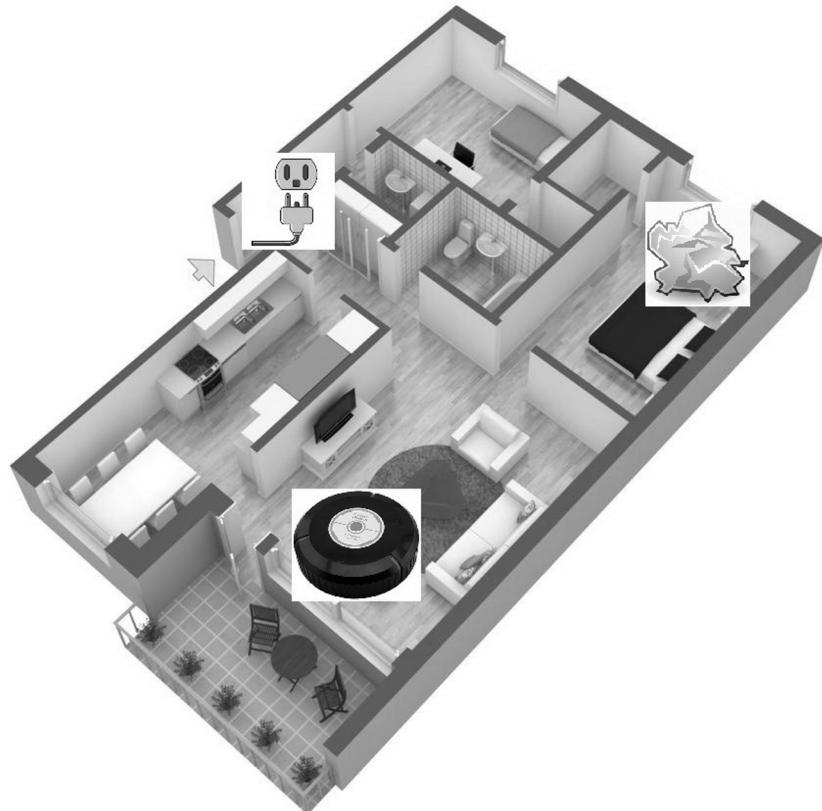
$$V^*(s) = \max V(s) = \max \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \dots]$$

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

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

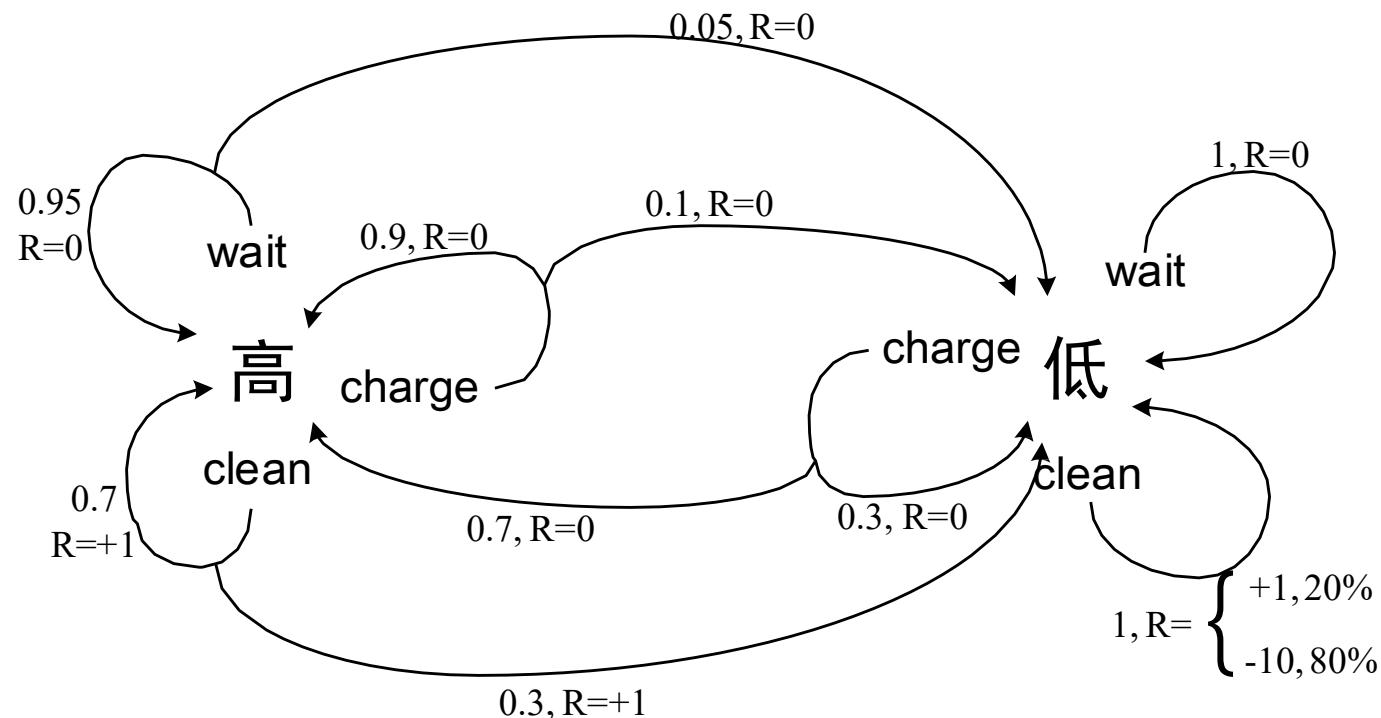
- 对每个马尔可夫决策问题，有且只有一个最优价值
- 但最优策略不一定是唯一的

# 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



充电不带来奖励，却是保证机器人长期运行的重要动作

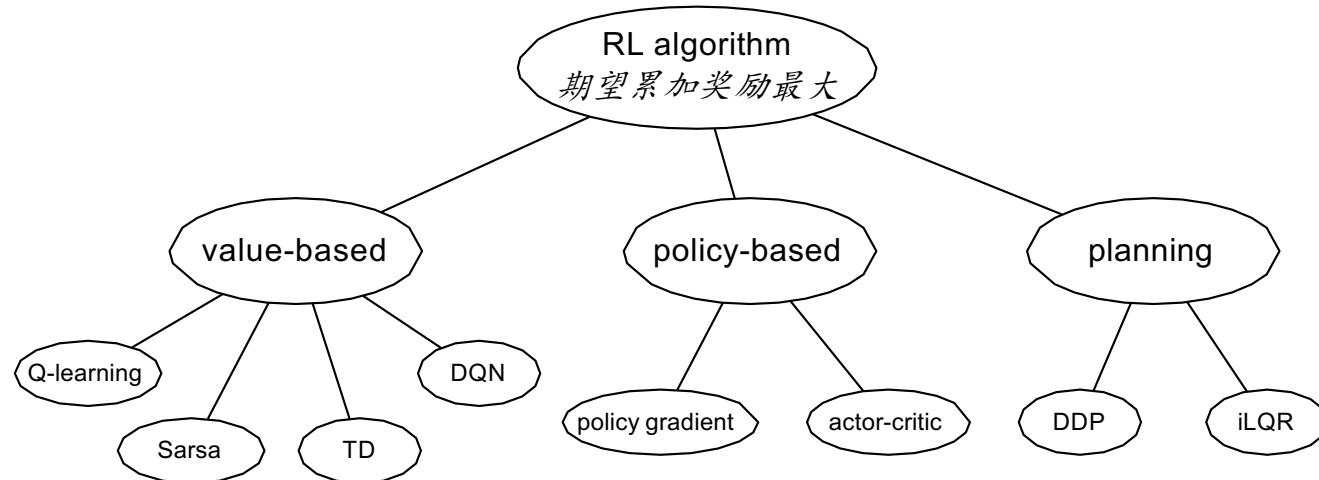
- 策略 1: 电量高时 wait, 电量低时 charge (lazy)
- 策略 2: 电量高时 clean, 电量低时 clean (poor experience)
- 策略 3: 电量高时 clean, 电量低时 charge (work too hard)
- 策略 4: 电量高时 50%clean, 50%wait, 电量低时 charge (very good)
- 回报: 机器人在未来一段时间获得的奖励

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

- 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/离线学习

- 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/不基于模型

- 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

# 回顾

课程简介（MindSpore完成作业）

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<sup>1†</sup>, 邵 坤<sup>1</sup>, 朱圆恒<sup>1</sup>, 李 栋<sup>1</sup>, 陈亚冉<sup>1</sup>, 王海涛<sup>1</sup>

(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<sup>2</sup>, 周 彤<sup>3</sup>, 王成红<sup>4</sup>

(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

2017年底发表，下载5000余次，年度第1

第34卷第12期

2017年12月

控制理论与应用

Control Theory & Applications

Vol. 34 No. 12

Dec. 2017

DOI: 10.7641/CTA.2017.70808

## 深度强化学习进展: 从AlphaGo到AlphaGo Zero

唐振韬, 邵 坤, 赵冬斌<sup>†</sup>, 朱圆恒

(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; 中国科学院大学, 北京 100190)

**摘要:** 2016年初, AlphaGo战胜李世石成为人工智能的里程碑事件。其核心技术深度强化学习受到人们的广泛关注和研究, 取得了丰硕的理论和应用成果。并进一步研发出算法形式更为简洁的AlphaGo Zero, 其采用完全不基于人类经验的自学习算法, 完胜AlphaGo, 再一次刷新人们对深度强化学习的认知。深度强化学习结合了深度学习和强化学习的优势, 可以在复杂高维的状态动作空间中进行端到端的感知决策。本文主要介绍了从AlphaGo到AlphaGo Zero的深度强化学习的研究进展。首先回顾对深度强化学习的成功作出突出贡献的主要算法, 包括深度Q网络算法、A3C算法、策略梯度算法及其它算法的相应扩展。然后给出AlphaGo Zero的详细介绍和讨论, 分析其对人工智能的巨大推动作用。并介绍了深度强化学习在游戏、机器人、自然语言处理、智能驾驶、智能医疗等领域的应用进展, 以及相关资源进展。最后探讨了深度强化学习的发展展望, 以及对其他潜在领域的人工智能发展的启发意义。

**关键词:** 深度强化学习; AlphaGo Zero; 深度学习; 强化学习; 人工智能

中图分类号: TP273 文献标识码: A

## Recent progress of deep reinforcement learning: from AlphaGo to AlphaGo Zero

The screenshot shows a mobile news application interface with six news cards. Each card has a timestamp at the top left (e.g., 2018-01-30, 2017-10-21) and a category at the top right (e.g., Control Theory & Applications, CASIA). The news cards are as follows:

- 【深度】自动化所解读“深度强化学习”：从AlphaGo到AlphaGoZero** (2017-10-21, 新智元)  
Image: A hand playing chess on a board.
- 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, 人工智能学家)  
Image: A hand playing chess on a board.
- 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)  
Image: A hand playing chess on a board.
- 【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)  
Image: A hand playing chess on a board.
- 【团队新作】深度强化学习进展：从AlphaGo到AlphaGo Zero** (2018-01-31, 中国科学院自动化研究所)  
Image: A hand playing chess on a board.
- 论文报道|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)  
Image: A hand playing chess on a board.

# 交流方式

群聊: 强化学习 22-23 春-课  
程交流群



**深度强化学习@CASIA**



该二维码 7 天内 (2 月 27 日前) 有效, 重新进入将更新