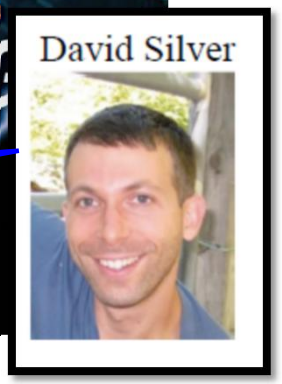
计算机前沿技术

行为主义（强化学习）

Deep Reinforcement Learning（深度强化学习）

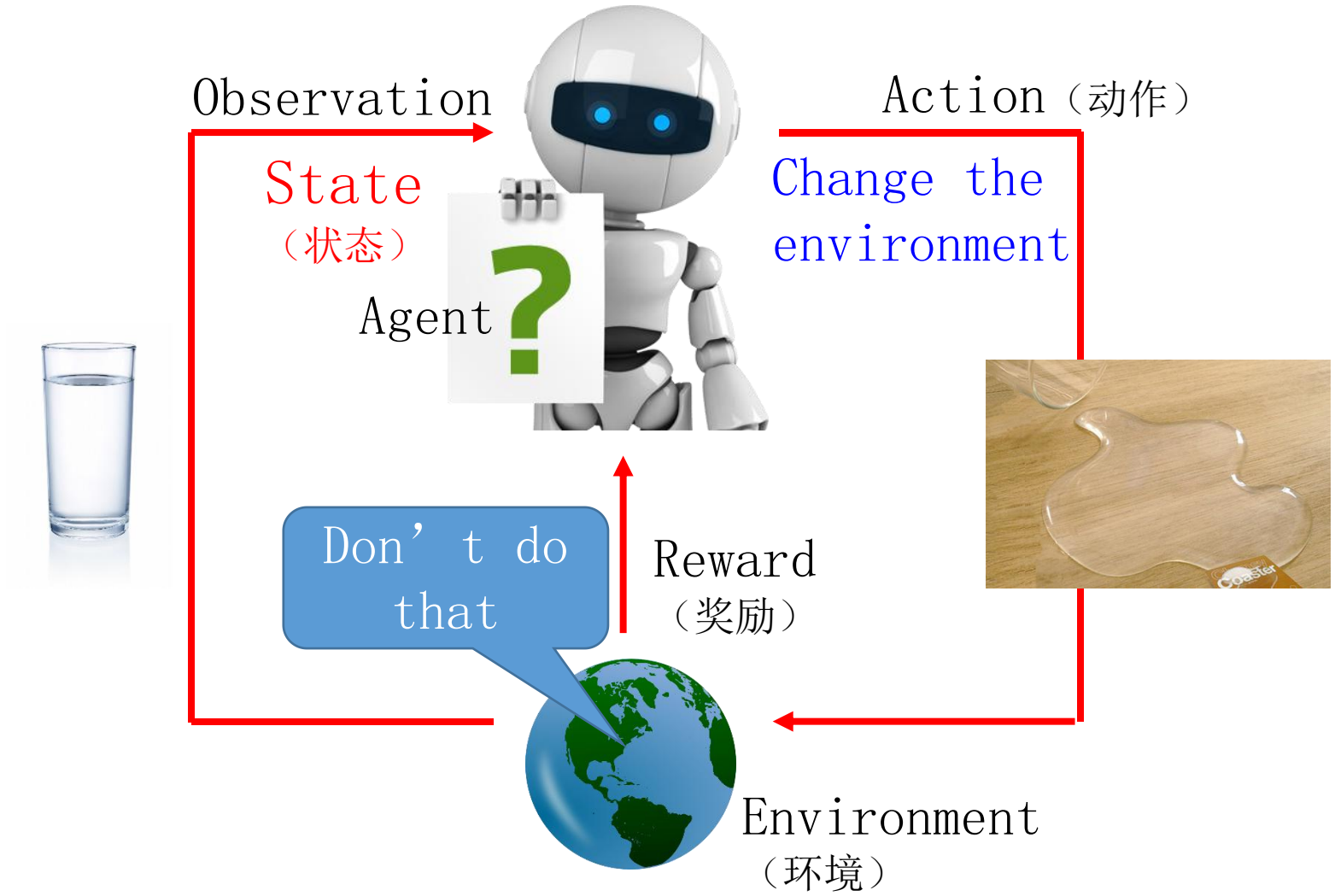
Deep Reinforcement Learning: AI = RL + DL



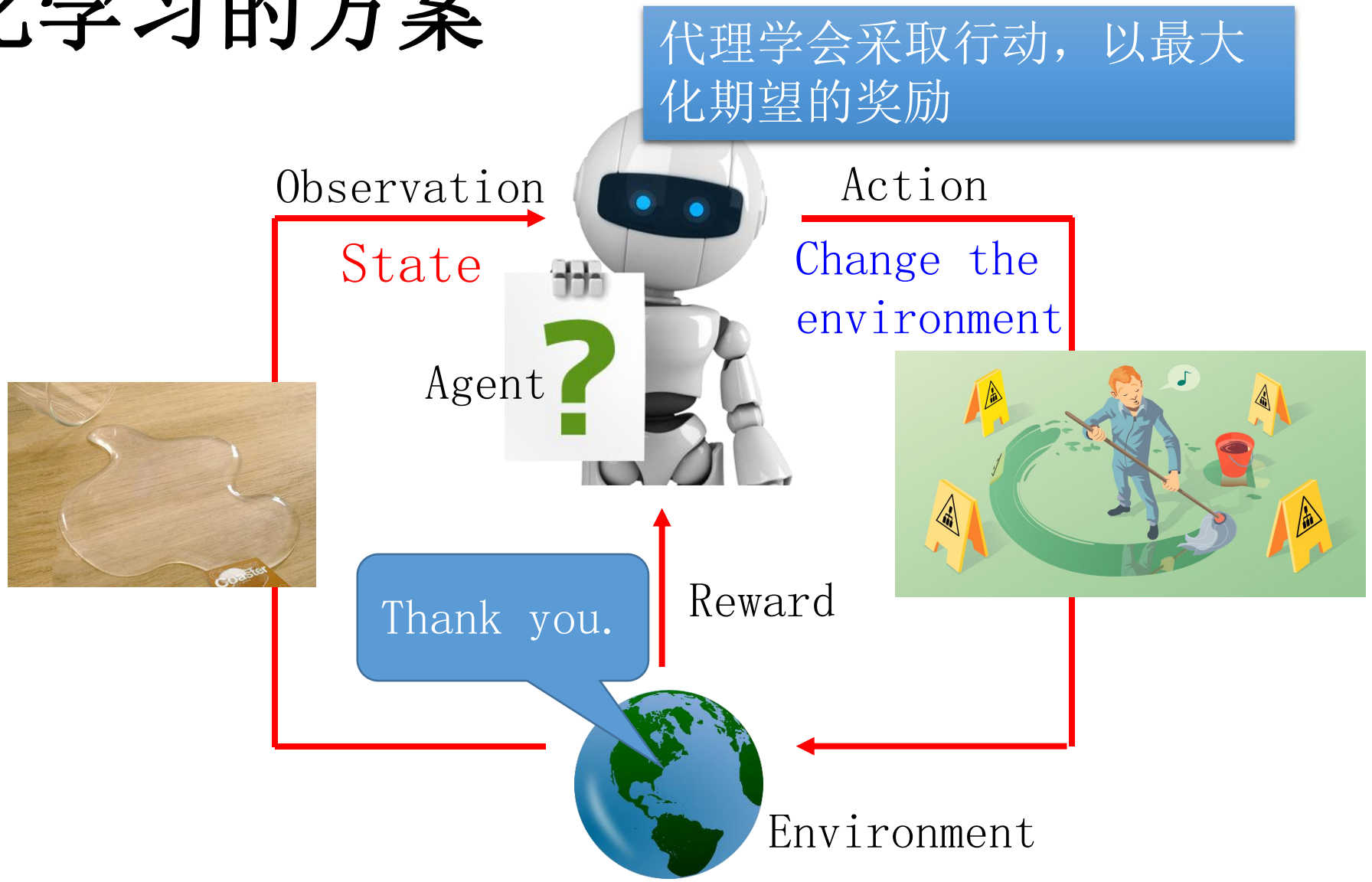
参考资料

- Textbook: Reinforcement Learning: An Introduction
  - http://incompleteideas.net/sutton/book/the-book.html
- Lectures of David Silver
  - http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Teaching.html (10 lectures, around 1:30 each)
  - http://videolectures.net/rldm2015_silver_reinforcement_learning/ (Deep Reinforcement Learning)
- Lectures of John Schulman
  - https://youtu.be/aUrX-rP_ss4

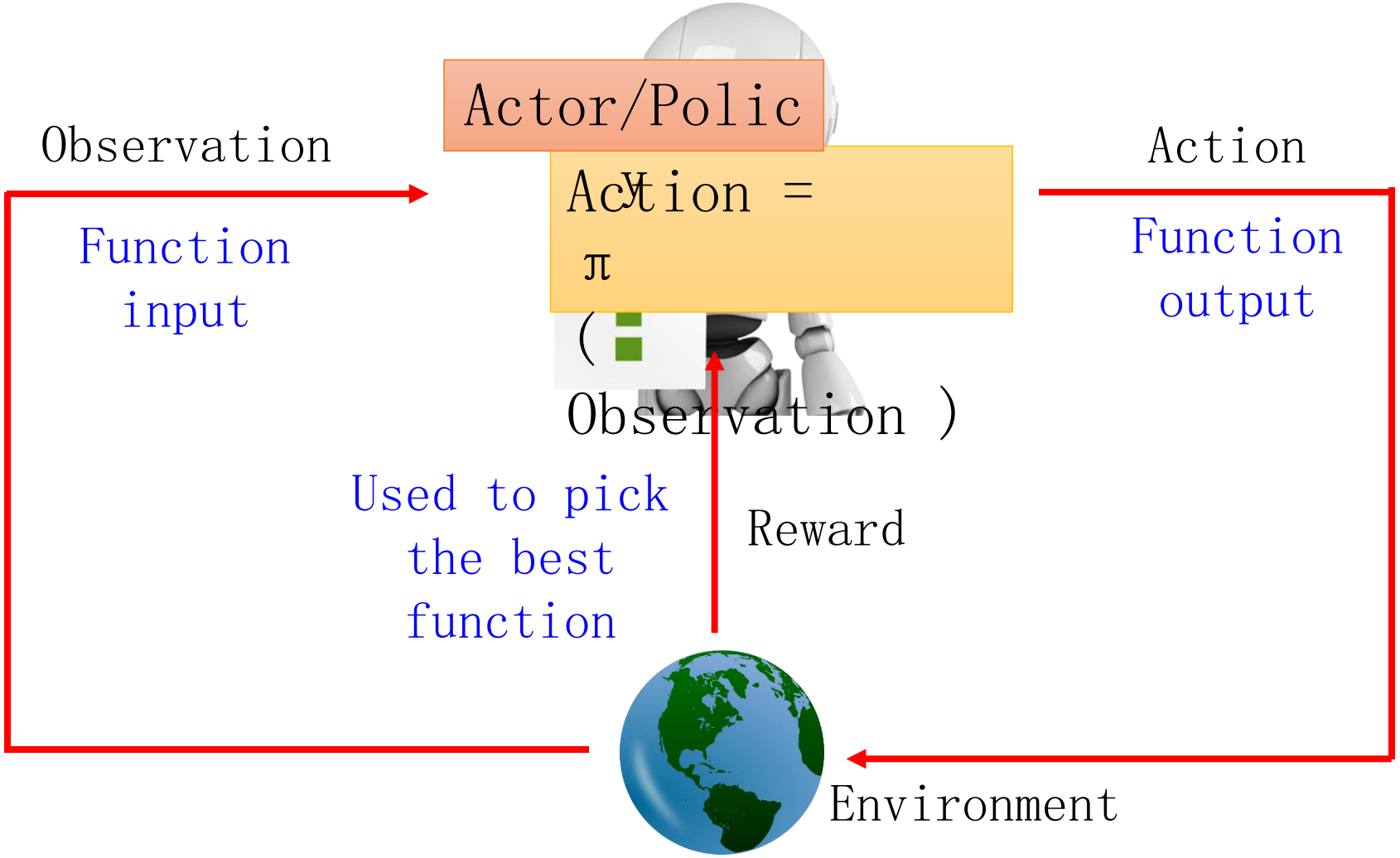
强化学习的方案


Observation
Action（动作）
State
（状态）
Change the environment
Agent
?
Don’t do that
Reward
（奖励）
Environment
（环境）

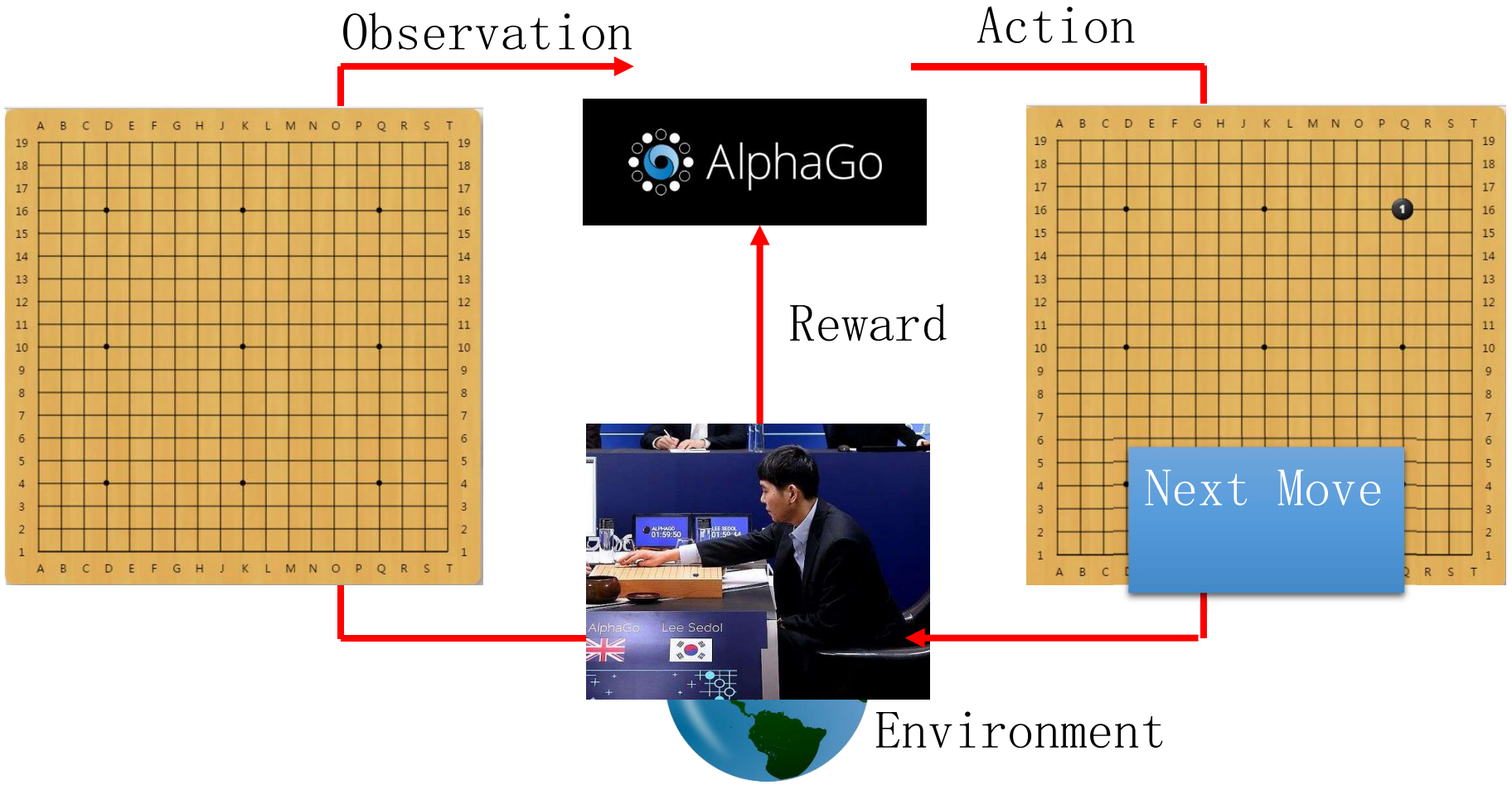
强化学习的方案


代理学会采取行动，以最大化期望的奖励
Observation
State
Agent
Action
Change the environment
Thank you.
Reward
Environment

机器学习
≈ 寻找一个函数

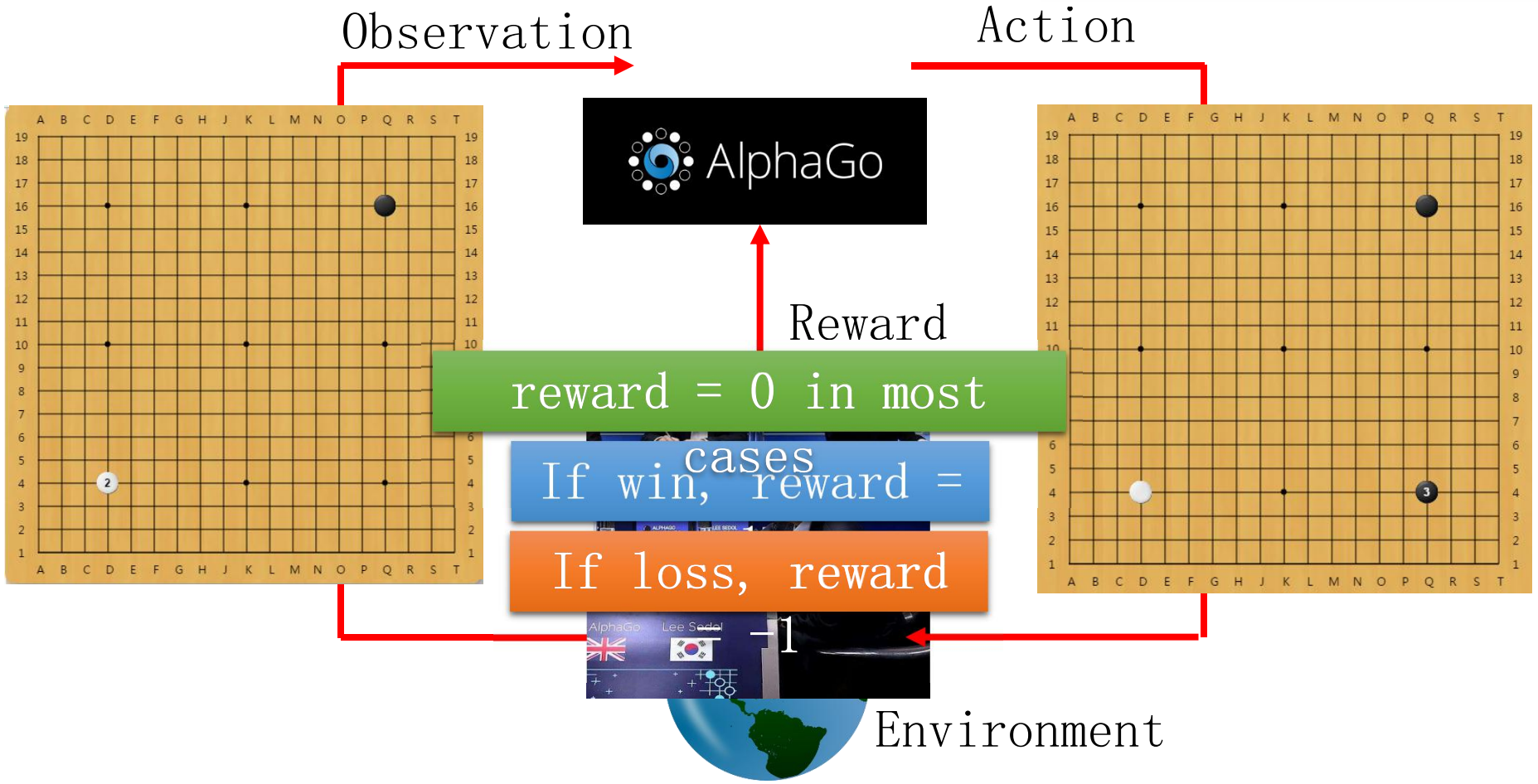

Actor/Polic
Observation
Action
Function input
Function output
Action =
π
(
Observation)
Used to pick the best function
Reward
Environment

学习下围棋



学习下围棋

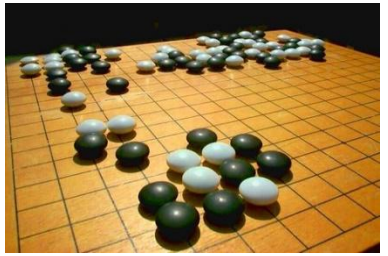
代理学会采取行动，以最大化期望的奖励



学习下围棋

- 监督学习：Learning from teacher

Next move:
“5-5”



Next move:
“3-3”

- 强化学习：Learning from experience

First move → …… many moves …… → Win!

(Two agents play with each other.)

Alpha Go 是监督学习 + 强化学习.

更多应用

- 直升机
  - https://www.youtube.com/watch?v=0JL04JJjocc
- 自动驾驶
  - https://www.youtube.com/watch?v=0xo1Ldx3L5Q
- 机器人
  - https://www.youtube.com/watch?v=370cT-OAzzM
- 谷歌用DeepMind驱动的人工智能削减巨额电费
  - http://www.bloomberg.com/news/articles/2016-07-19/google-cuts-its-giant-electricity-bill-with-deepmind-powered-ai
- 文本生成
  - https://www.youtube.com/watch?v=pbQ4qe8EwLo

例子: Playing Video Game

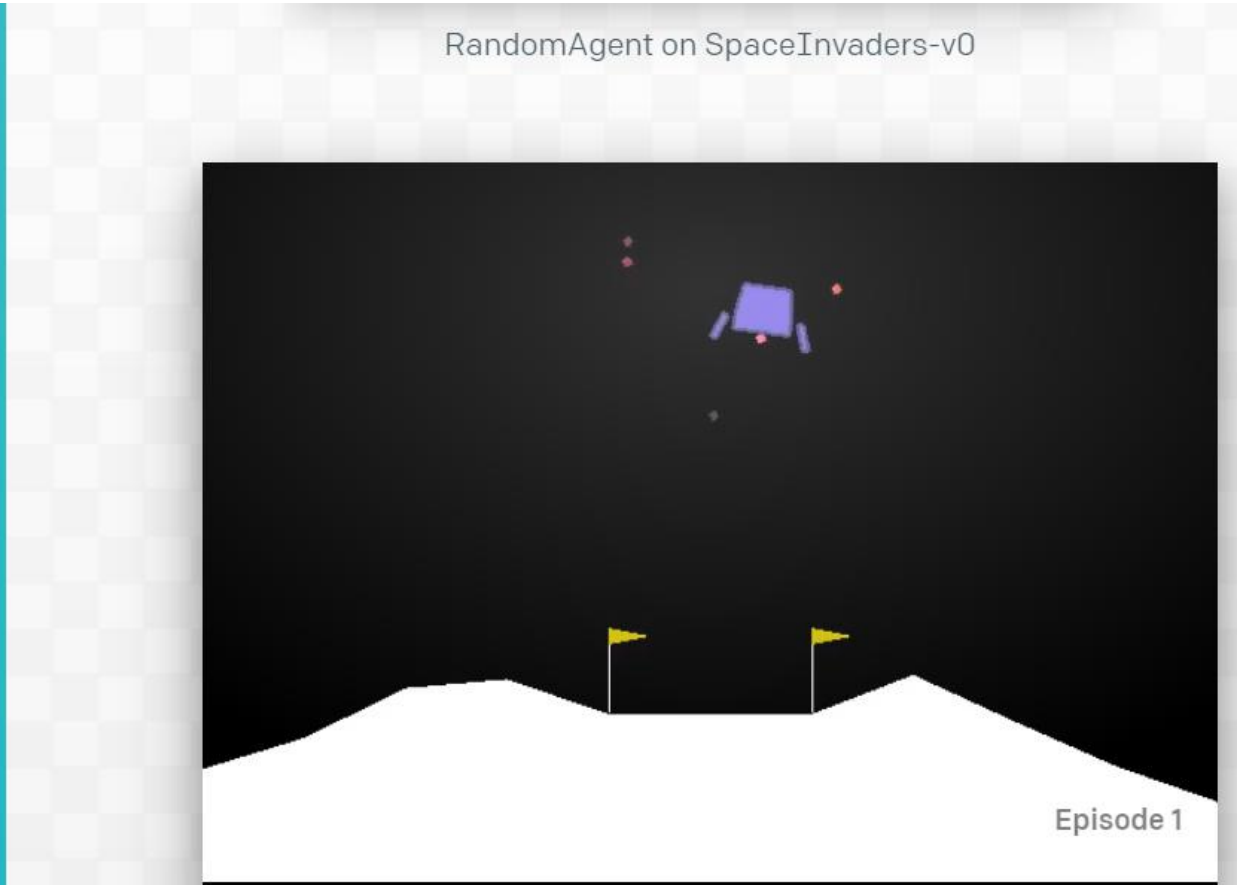
机器学习学习像人类玩家一样打游戏。

- 机器观察到的是像素
- 机器学会自己采取适当的行动

例子: Playing Video Game

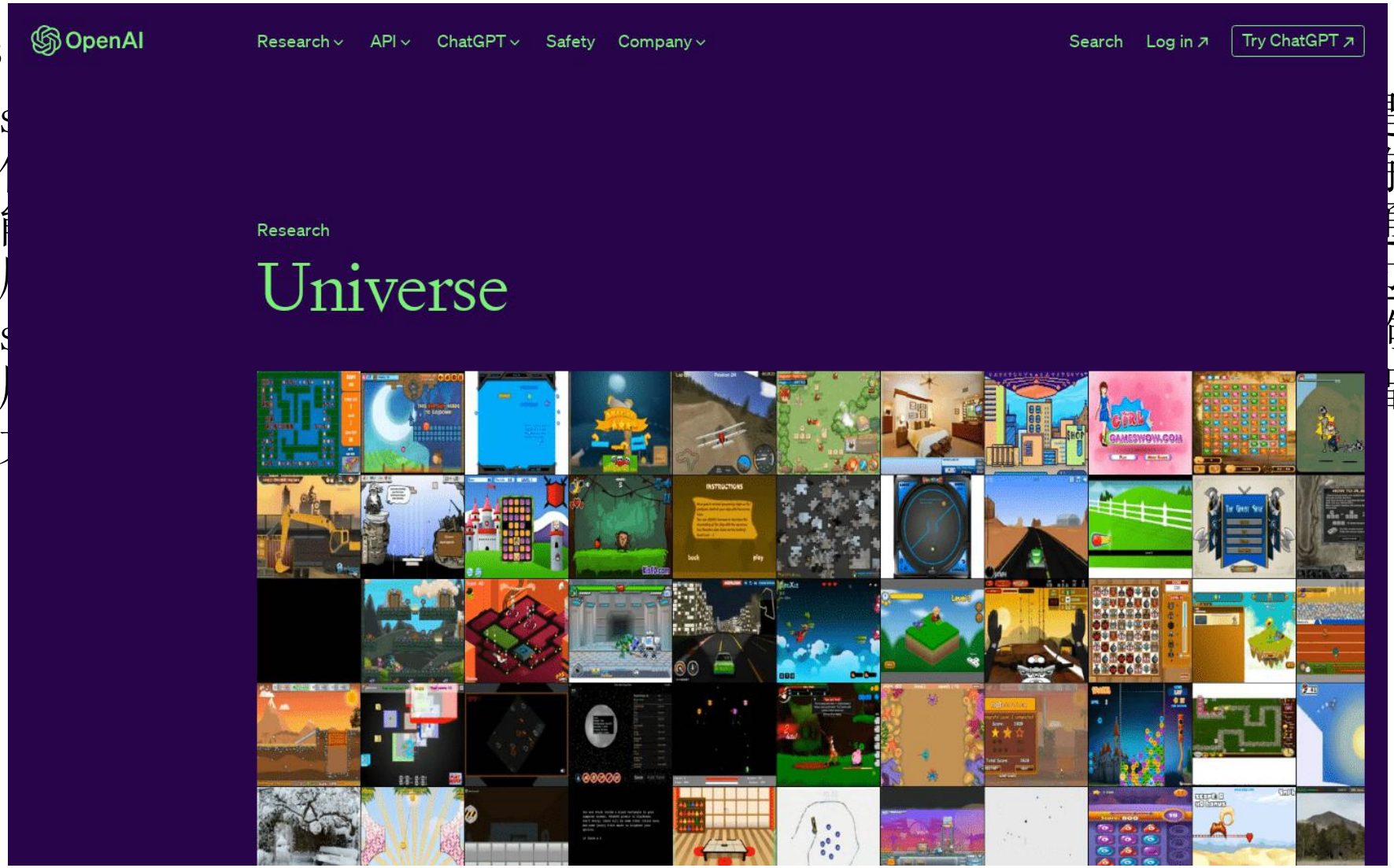
Gym

Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents everything from walking to playing games like Pong or Pinball.



例子：Playing Video Game

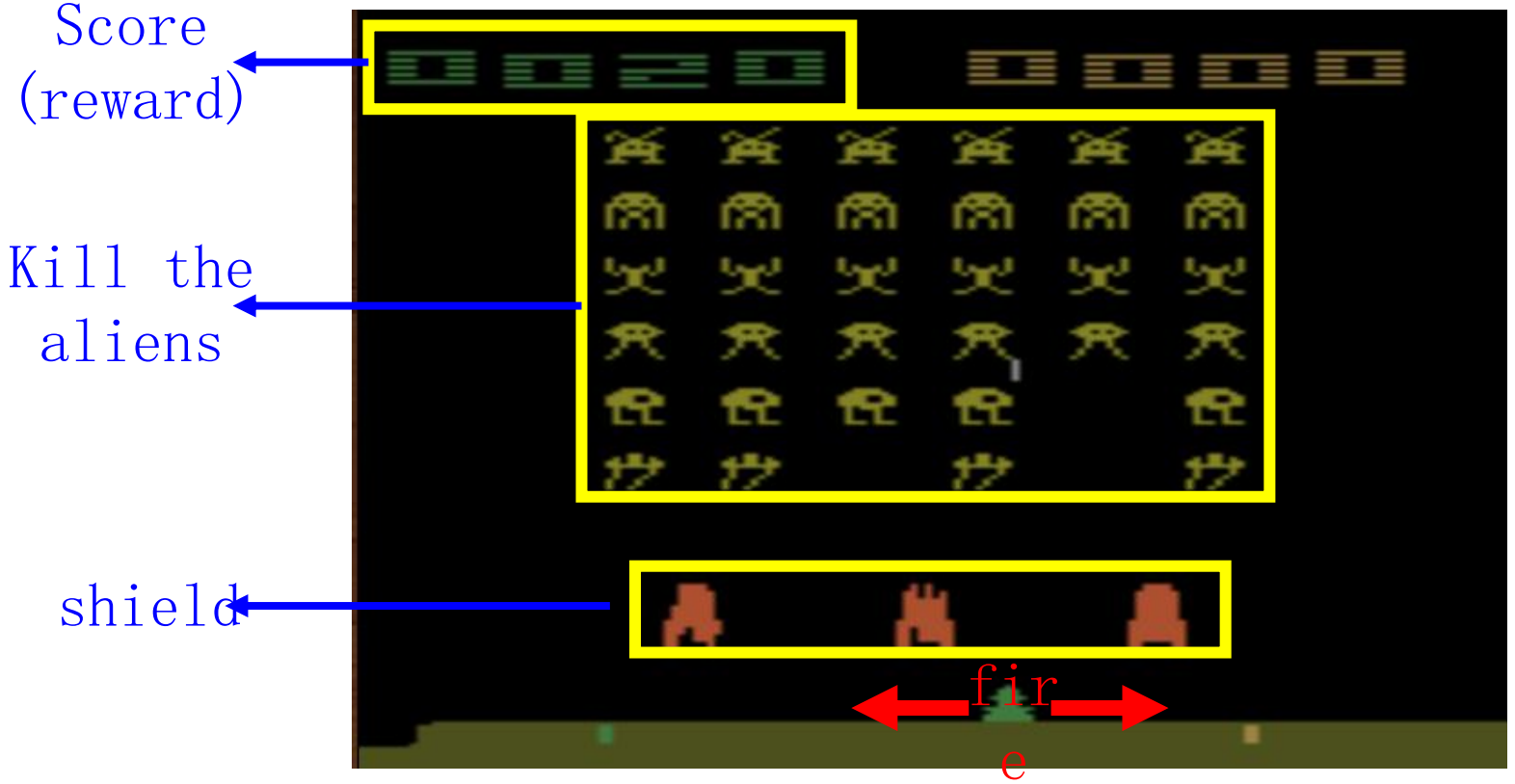
- Widely s
  - Univers
    一个软
    人工智
    过观察
    括 Flas
    灵活应
    能的重

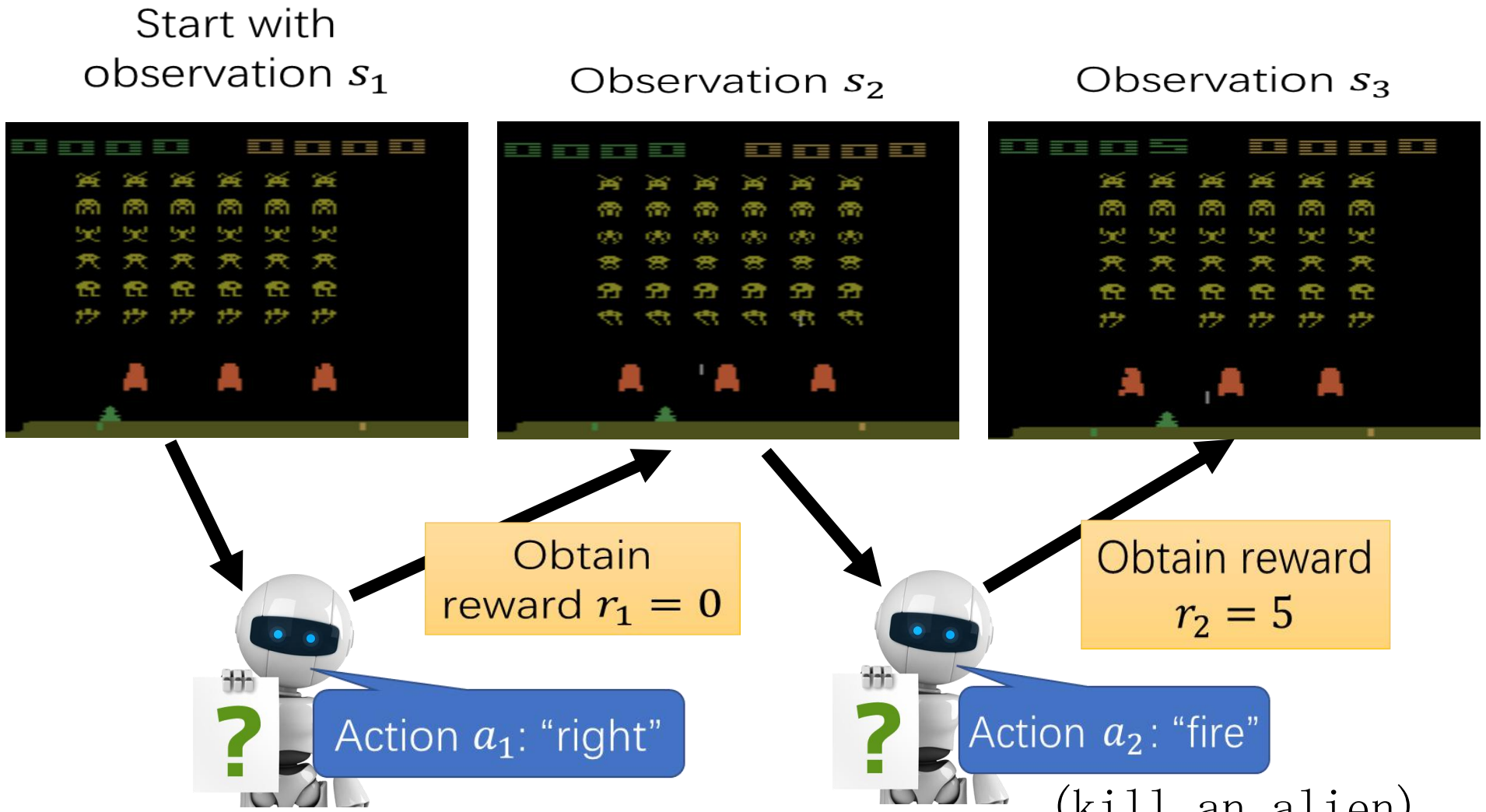


RL玩游戏的例子，左上方是已获得分数，中间是还没打完的怪，下方则是你可操作的动作，包括向左移动、向右移动以及开火

- Space invader

Termination: 所有怪物被杀死，或者你的飞船坠毁。



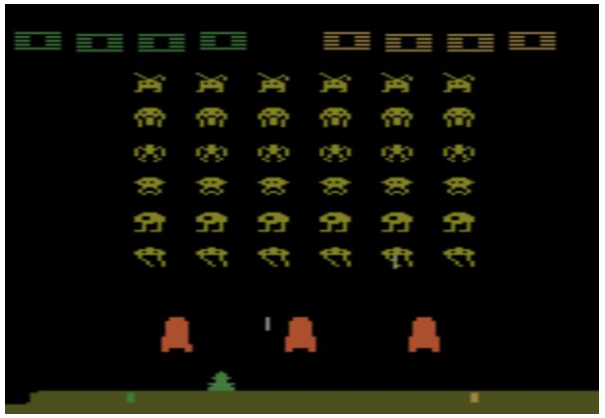


(kill an alien)

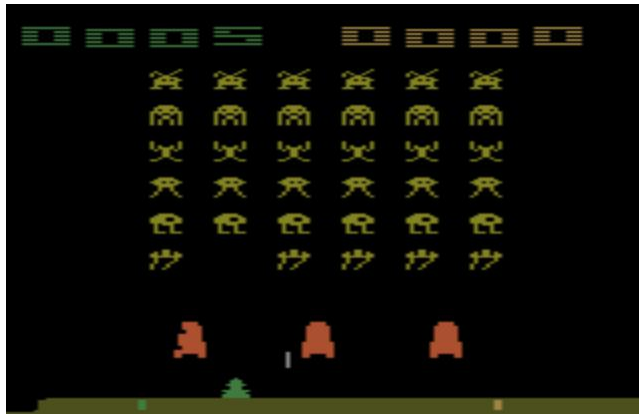
首先机器看到最左边的画面（state s1），接着采取行动（action a1）向右走一步，得到回馈reward（r1 = 0），然后再接收状态资讯（state s2），接着再选择开火（action a2），然后环境给予他的回馈奖励（r2 = 5），

Start with observation s_1

Observation s_2



Observation s_3



After many turns

Obtain reward r_T

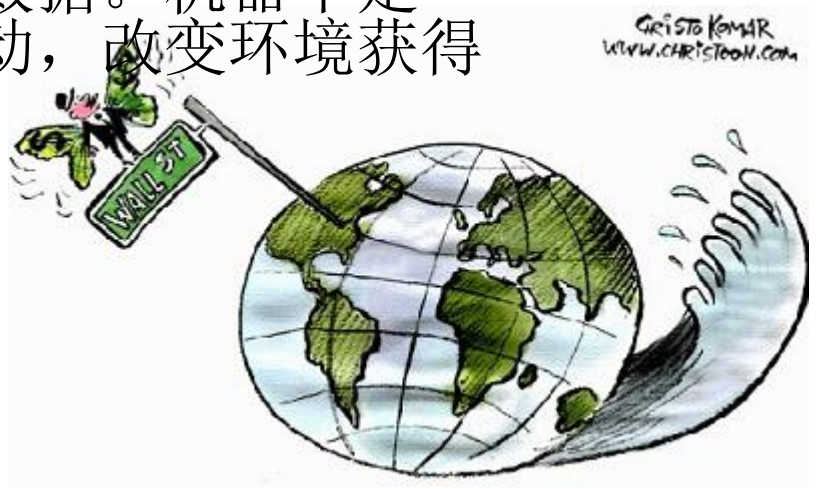
Action a_T

This is an episode

直到游戏结束，整个过程会得一个累积的奖励，游戏会以整个情节的奖励为目标，并按照目标最大化原则调整行为。

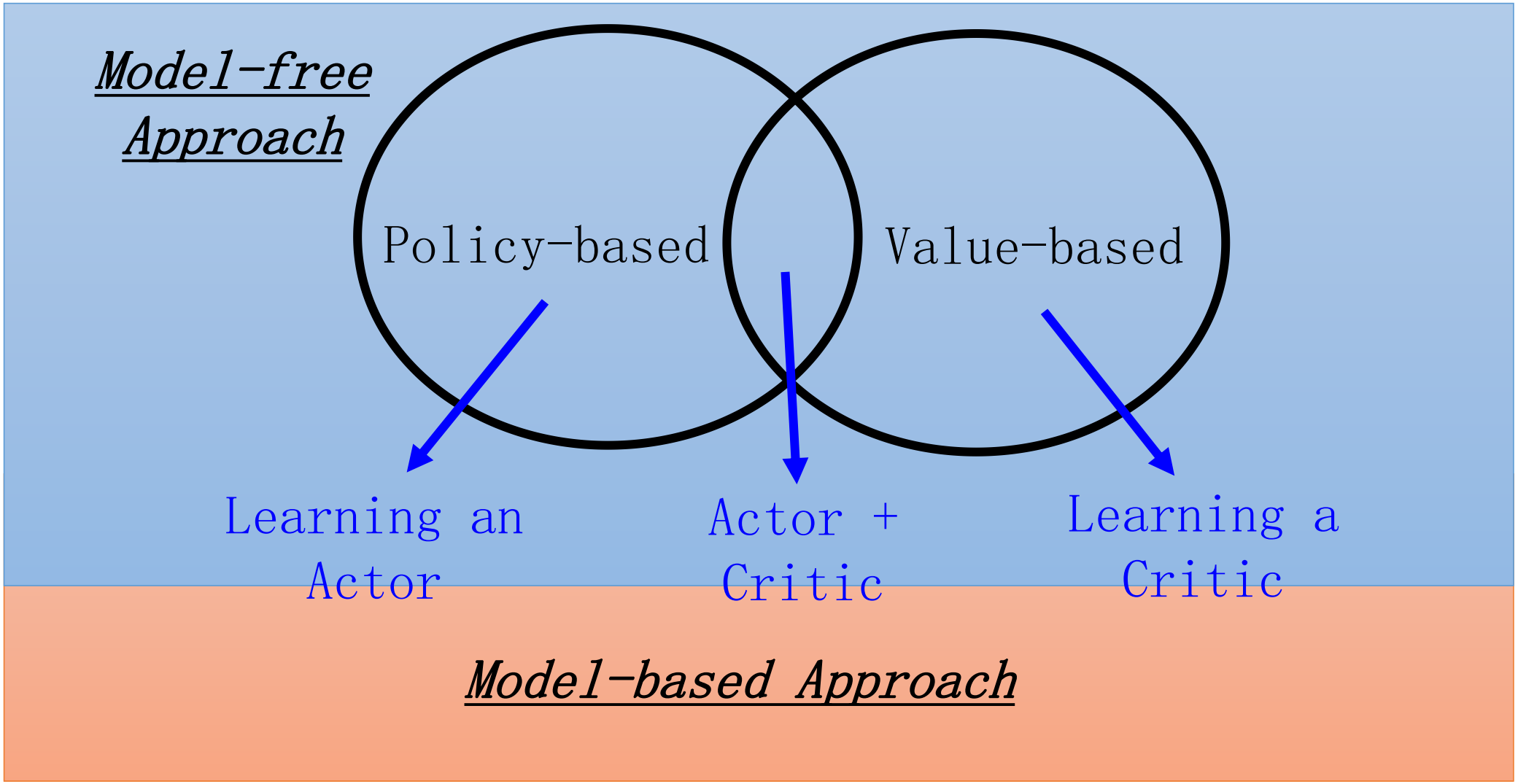
强化学习的特性

- **奖励延迟**
  - 你的机器人或许会知道开火跟得分有关系，但不能直接了解得分跟往左右移动有什么关系，这样机器最后只会不断地开火。
    - 再举个围棋的例子，在与环境对弈的过程，并不是每步都有明显的回馈说这步下得很好，有时早期的牺牲些区块，诱敌等战术都能让你在后面获得更好的期望利益
    - 学习的对象是一连串的行为（轨迹），机器才能了解，有些没有及时奖励值也是很重要，目标是最大化整个过程的奖励。
    - 另一个特性是，机器的行动会影响它接收的随后的数据。机器不是一开始便拥有标注好的资料，机器要跟环境持续做互动，改变环境获得反馈，玩许多次才会更新算法，过程整个这样持续。



Outline

Alpha Go: policy-based + value-based + model-based

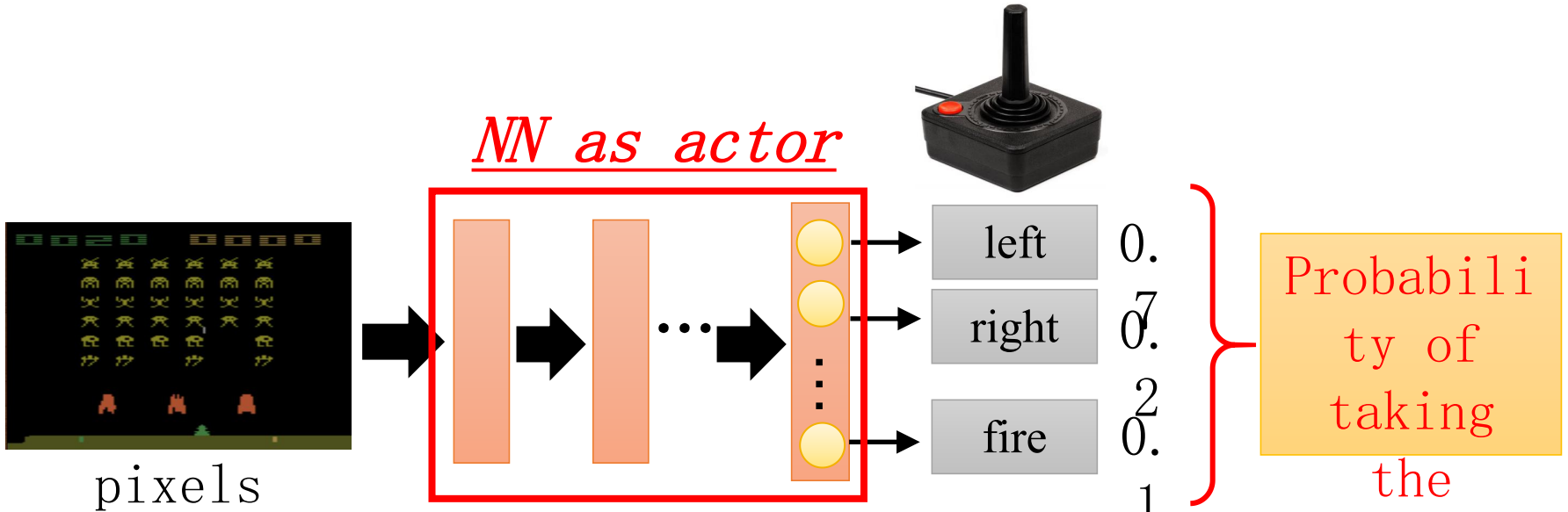


Policy-based Approach

Learning an Actor

神经网络作为 Actor （ ）

- 神经网络的输入：机器的观察表示为向量或矩阵
- 神经网络的输出：输出层，每个行动相关的神经元



使用网络代替查表的好处？

泛化generalization

Value-based Approach

Learning a Critic

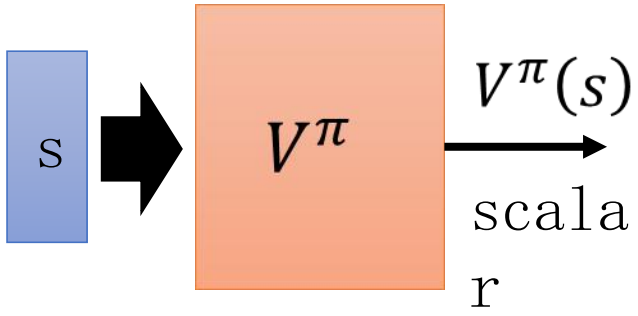
Critic

- Critic是什么呢？Critic并不会跟你的训练过程有直接关系，它要做的是评估一个Actor的好坏，好的Actor会由Critic挑出，Q-学习就是这样的方法。

http://combiboilersleeds.com/picaso/critics/critics-4.html

Critic

- 价值函数V是怎么评估一个值的呢?
- V评估的方法是输入进你的前状态，然后给出后面会累积奖励的值。如果是游戏还没开始多久，画面上可得分的目标还挺多，V产出的值便会很大。如果目标已经被击落的差不多了V值便会比较小。但这前提是你的Actor够强，如果Actor在前面阶段便被射中，当然V值也会较小。
- State value function $V^{\pi}(s)$



$V^{\pi}(s)$ is large

$V^{\pi}(s)$ is smaller

Critic

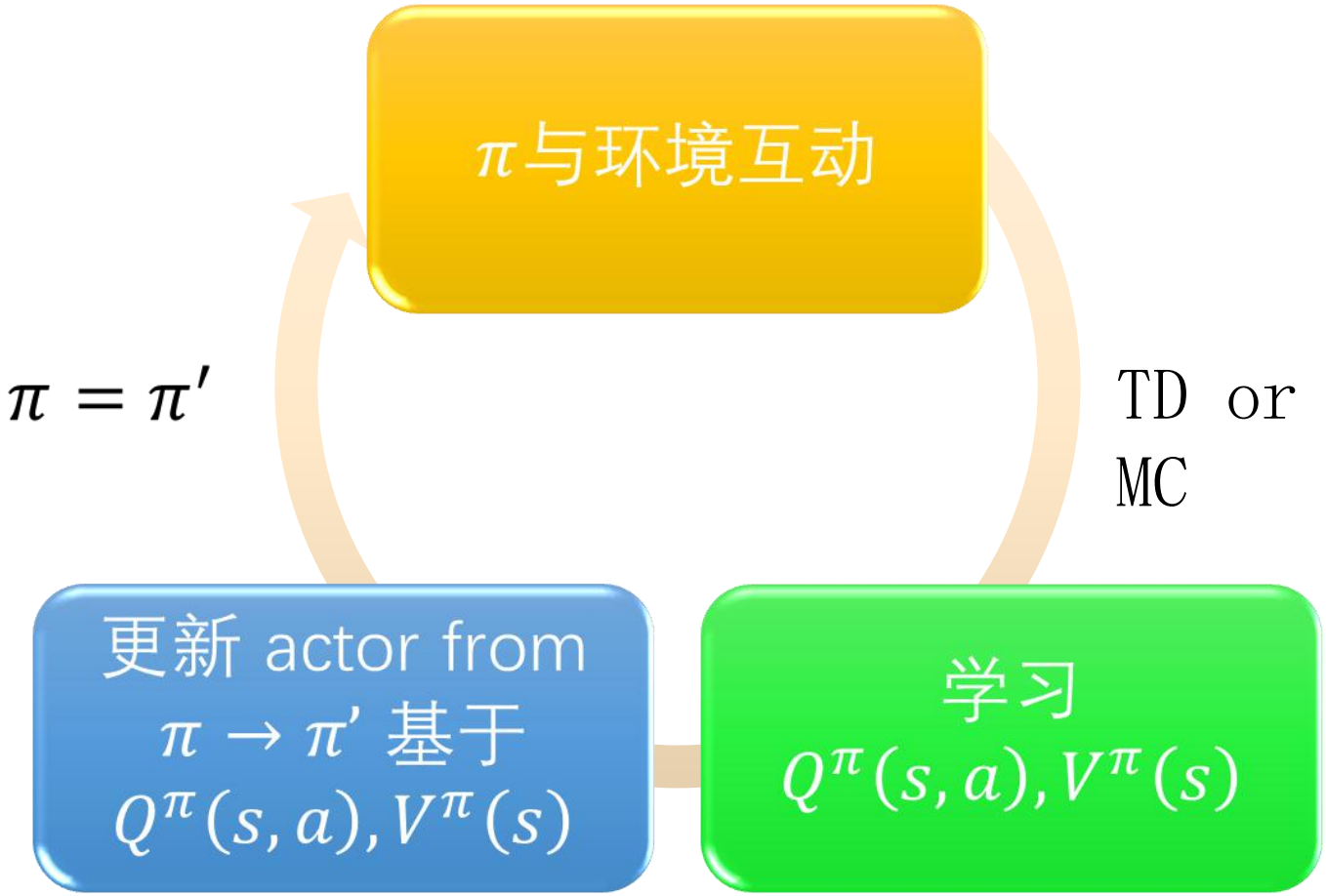
v以前的阿光(大馬步飛) = bad

v變強的阿光(大馬步飛) = good

Deep Reinforcement Learning

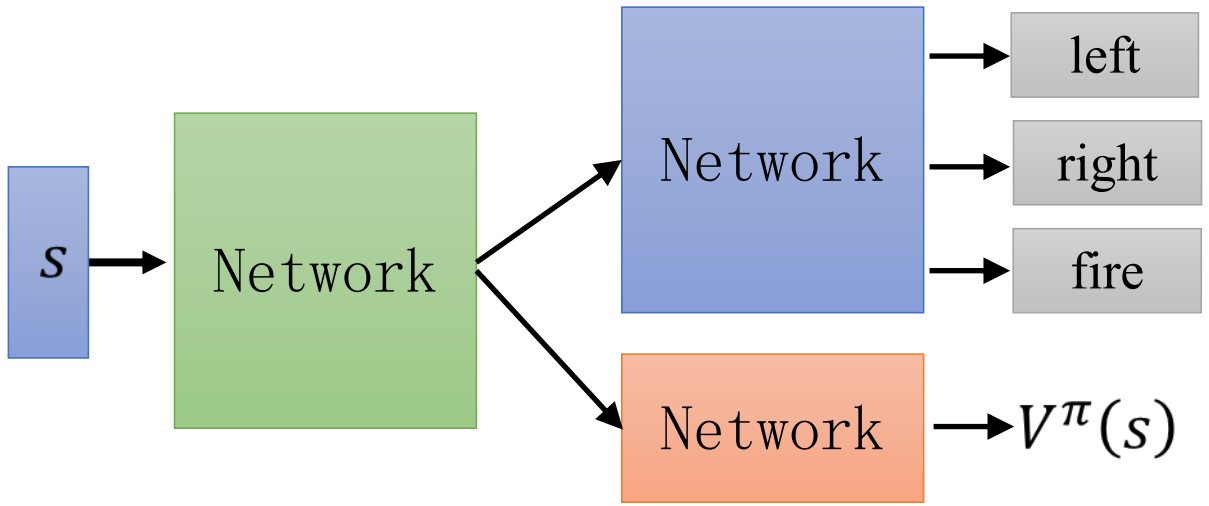
Actor-Critic

Actor-Critic



Actor-Critic

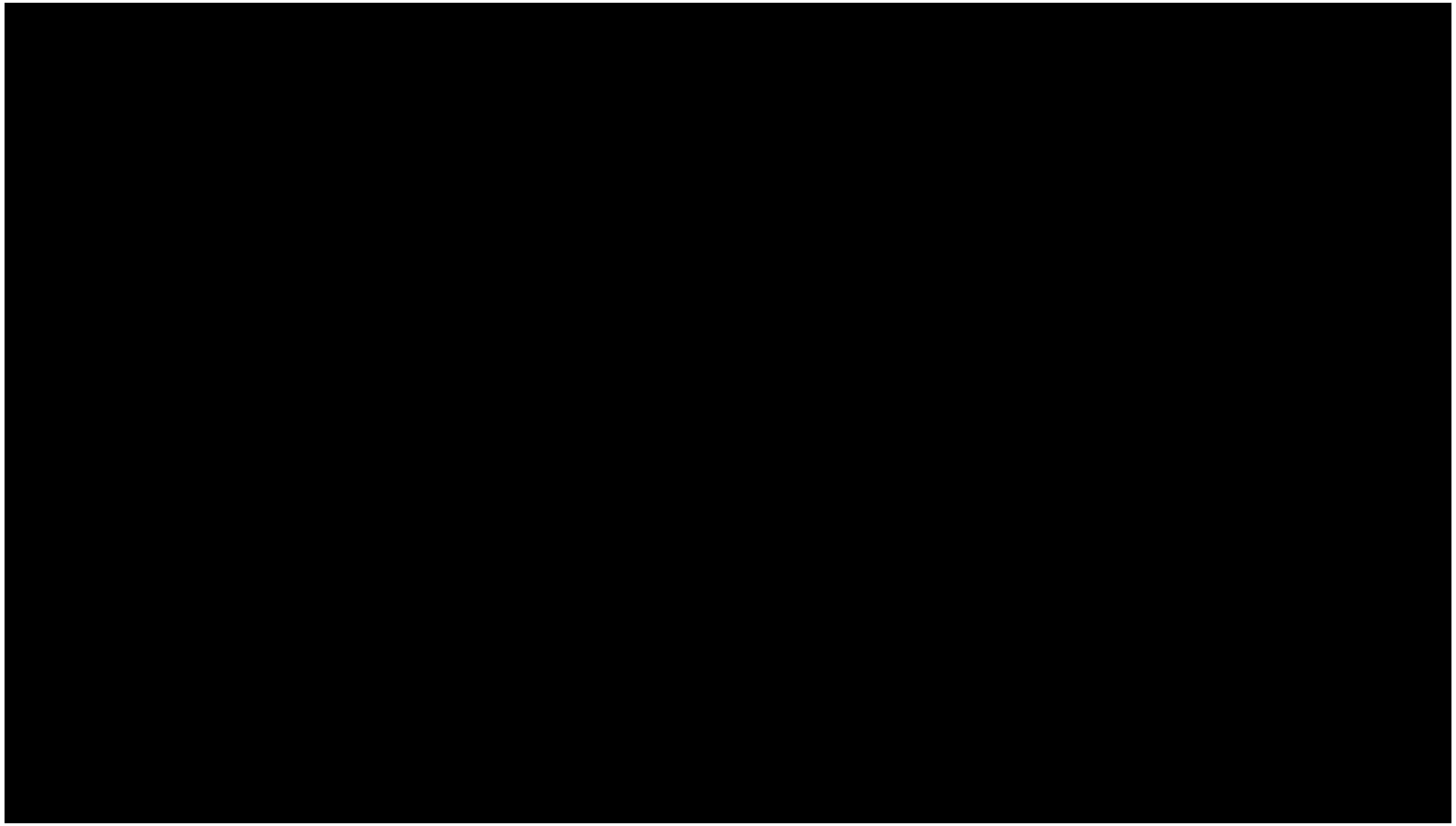
- Tips
  - actor $\pi(s)$ and critic $V^{\pi}(s)$ 的参数能被共享



Demo of A3C

https://www.youtube.com/watch?v=0xo1Ldx3L5Q

- Racing Car (DeepMind)



A3C

- A3C解决了Actor-Critic难以收敛的问题，同时更重要的是，提供了一种通用的异步的并发的强化学习框架，也就是说，这个并发框架不光可以用于A3C，还可以用于其他的强化学习算法。这是A3C最大的贡献。目前，已经有基于GPU的A3C框架，这样A3C的框架训练速度就更快了。

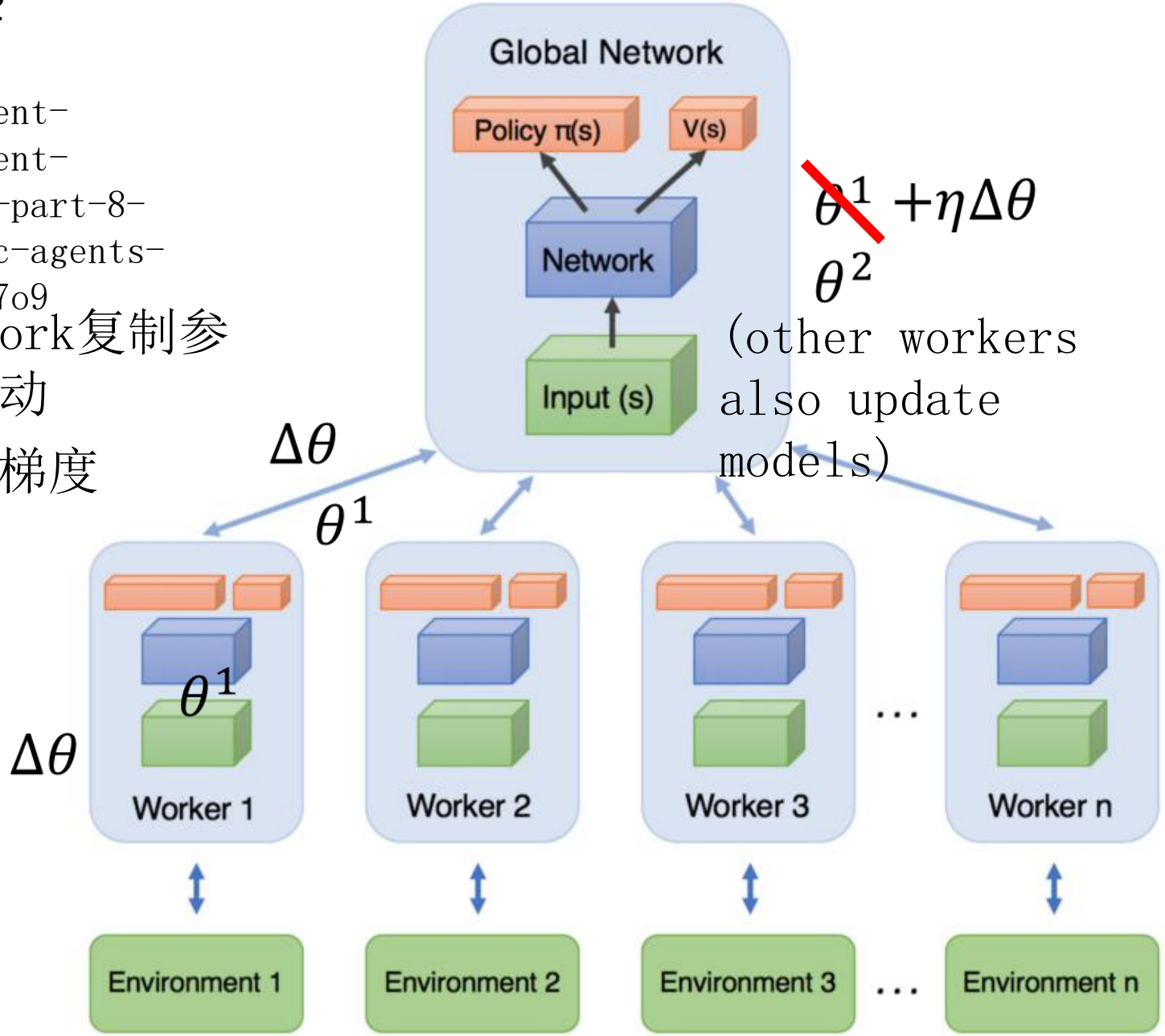
Asynchronous

Source of image: https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2#.68x6na7o9

1. 每个worker从global network复制参数;不同的worker与环境去做互动

3. 不同的worker计算出各自的梯度

4. 不同的worker把各自的梯度传回给global network，global network接收到梯度后进行参数更新。




Thanks!
Obrigado!
Gracias!
Dziękuję!
谢谢！
Danke!
Grazie!
شكرا لك !
Merci!
감사합니다!
Kiitos!