

每次只改一小步

本课讲 2017 年 OpenAI 的一篇文章。John Schulman 和同事发明了 PPO：让 AI 边试边学时，每次只把习惯改一小步。这样学得又稳又快。

适合 10 岁左右

约 40 分钟

需要：铅笔、几个纸团、垃圾桶、2-5 位同伴

1 热身：零食和怪兽

先填空。想一想小狗和游戏里的你。

训练员说“坐”，小狗坐下了。训练员给它一块_____。

下次听到“坐”，小狗会_____坐下。（更愿意 / 更不愿意）

碰到怪兽掉了一条命，下一局你会_____怪兽。

没有人给小狗讲道理，也没有人给你说明书。可是你们都学会了。办法就是**试一试，看结果，再调整**。今天的主角 PPO，就是教 AI 这样学习的一个办法。

2 边试边学：强化学习

让电脑边试边学，有个大名字，叫**强化学习**。“强化”的意思是：做得好的事，以后多做。

学习的那一方叫**智能体**。它可以是游戏里的小人，也可以是机器人。它所在的世界叫**环境**。它每做一个**动作**，环境就给它一个**奖励分**。奖励分可以是正的，也可以是负的。



🔄 再看下一个情况，重复几百万次

和第 1 课比一比：第 1 课电脑猜字，每次都**知道正确答案**。强化学习没有正确答案，智能体只知道**自己得了几分**。

3 策略：它的做事习惯

智能体心里有一张**习惯表**。表上写着：在什么情况下，做什么动作，做的可能性有多大。这张表的大名叫**策略**。

比如一个闯关游戏里的小跳人。它的习惯表可能是这样（示意）：

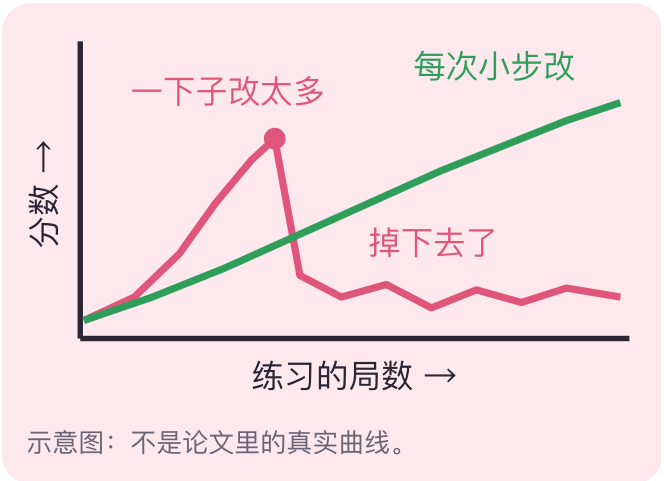
看到什么	跳起来	往前走	蹲下
前面有个坑	10 次里 5 次	10 次里 4 次	10 次里 1 次
头顶飞来一只鸟	10 次里 2 次	10 次里 3 次	10 次里 5 次

遇到坑时，它跳过去了，得了 10 分。那就把“跳起来”调高一点。它往前走掉进坑里，扣了 50 分。那就把“往前走”调低一点。**学习，就是一点一点改这张习惯表**。每次该改多少呢？

4 麻烦：一下子改太多，会崩

想想学骑自行车。你刚能骑一小段，有点歪。如果一次把坐姿、手和蹬车的力气**全改了**，很可能摔得更狠。原来会的那一点也没了。

电脑也会这样。它根据刚玩的几局改习惯，改得太猛，习惯就可能变坏。更麻烦的是：接下来的每一局，都是坏习惯玩的。小跳人一直掉坑，就学不到怎么跳过去。分数掉下去，很难再爬起来。



论文里真的试过：什么限制都不加，让它想改多少就改多少。结果在机器人走路的测试里，它的平均成绩比**随便乱动**还差。

5 PPO 的办法：每次只许小步改

PPO 定了一条规矩：**新习惯和旧习惯比，差别不能超过大约两成**。这一招叫**裁剪**，就像用剪刀把多出来的部分剪掉。

还看小跳人。遇到坑时，它原来 10 次里跳 5 次。这一轮学完，最多改到 10 次里跳 6 次，最少改到 4 次。下图横线上的数字就是“10 次里跳几次”（示意）。



PPO 给每一次改动打分。在绿色范围里改得好，就算好处。跳出绿色范围的那部分，**不再算好处**。就像集点卡每次最多盖 2 个章，再多也不盖。改过头得不到好处，所以电脑没理由改过头。

还有一点很聪明：**上限只管好处，不管坏处**。如果改动让习惯变差了，照样扣分。电脑不会因为有限，就躲过惩罚。

好处 1：学得稳。每一步都不大，原来会的本领不会一下子丢光。

好处 2：一批经验能多学几遍。论文的机器人实验里，每玩一批就复习 10 遍，不用每改一次就重新去玩。

好处 3：简单。2015 年，Schulman 发明过一个类似的办法，叫 TRPO。它要用很复杂的数学。PPO 只改了打分规则，程序只比普通办法多几行代码。

6 结果：机器人走得稳，小游戏打得好

机器人模拟 (MuJoCo)：在电脑里的虚拟世界控制机器人走路、跳跃、保持平衡，共 7 个任务。PPO 几乎每个都超过了老办法，包括 TRPO。它还教会 3D 小人奔跑、转弯，摔倒后爬起来。

Atari 小游戏：49 个老式电子游戏，比如打砖块。看整个练习过程的平均分，PPO 赢了 30 个，最多。看最后的分数，另一个办法 ACER 赢得更多。

诚实地说：PPO 不是每一项都第一。它的长处是**学得快、很稳、程序简单**。所以大家都爱用它。

7 这和今天有什么关系

PPO 很快成了 OpenAI 默认的强化学习办法。2019 年，OpenAI 用大规模的 PPO 训练出一支打 Dota 2 游戏的 AI 队伍。它打败了当时的世界冠军队。

聊天 AI 也用到了它。聊天 AI 先学会猜下一个字（第 1 课）。可是会猜字，不等于会好好回答问题。所以科学家又让它边试边学：

- ① AI 写出几个回答。
- ② 人挑出更喜欢的那个。
- ③ 按人的喜好给回答打奖励分。
- ④ 用 PPO 小步调整 AI 的习惯。

2022 年，OpenAI 的 InstructGPT 就用了这个办法。后来的 ChatGPT 也沿用了同一套流程。第 30 课会细讲“怎样让 AI 懂得人的喜好”。

诚实地说：这不是唯一的办法。后来的新办法有的完全不用强化学习，有的去掉了 PPO 的一部分。不过很多新办法还保留着它最核心的一招：**每次只许小步改**。

8 动手：小步改投篮

玩法（4—6 个人分成两组，约 15 分钟）

把垃圾桶放在地上。每人揉 1 个纸团，站在离桶 2—3 米的地方。每轮每组投 5 次，数投进几个，一共投 3 轮。

小步组：每轮只改一个地方，而且只改一点点。比如往前挪半步，或者力气大一点点。**大改组：**每轮可以随便改。站位、力气、姿势，想换就全换。每轮投完，在表里写上投进几个、改了什么。

第几轮	小步组投进	小步组改了什么	大改组投进	大改组改了什么
第 1 轮		(先随便投)		(先随便投)
第 2 轮				
第 3 轮				

你会发现：小步组的成绩多半一点点往上走。大改组容易忽高忽低，原来投得准的地方也可能被改丢。PPO 选的就是小步组的走法。

9 动手：训练小狗

玩法（3 个人，约 10 分钟）

一人当小狗，一人当训练员，一人当记录员。训练员偷偷写下一个目标动作，只给记录员看，比如“走到门边摸门把手”。

小狗每次只能做一个**小动作**：走一步、转身、抬手。训练员不许解释，只能说“热”（更近了）或“冷”（离远了）。做对了，就说“加 10 分！”记录员记下用了几次提示、几秒，然后换人再玩。

第 1 轮：提示 _____ 次， _____ 秒 第 2 轮：提示 _____ 次， _____ 秒

你会发现：小狗不知道正确答案，只能靠“热”和“冷”这样的奖励分来猜。听到“热”就接着试，听到“冷”就退回来换个方向。这就是强化学习。

10 词语小卡片

强化学习

智能体边试边学，做得好的事以后多做。

奖励分

环境给动作打的分，可加可扣。越多越好。

策略

习惯表：什么情况下做什么，可能性多大。

裁剪 (PPO)

新旧习惯差别不超过约两成，改过头不加分。

11 想一想

1. 学骑车或学游泳时，为什么一次把姿势全改了，反而容易更糟？

2. 如果上限定得特别小，每次只许改一丁点，会有什么坏处？

3. 聊天 AI 的奖励分来自人的喜好。如果打分的人总给“说好听话”的回答加分，AI 会学成什么样？

一句话记住：边试边学时，每次只改一小步，学得稳，原来会的也不会丢。

给家长和老师：本课讲的是 John Schulman、Filip Wolski、Prfulla Dhariwal、Alec Radford、Oleg Klimov (OpenAI) 2017 年的论文 [Proximal Policy Optimization Algorithms](#) (arXiv)。PPO 把新旧策略的概率比裁剪到 $[1 - \epsilon, 1 + \epsilon]$ (论文取 $\epsilon = 0.2$)，再与未裁剪项取较小值，只去掉改进方向上超出范围的收益，不是硬约束。MuJoCo 7 个任务上，不裁剪不惩罚的版本归一化平均分为 -0.39 (随机策略为 0)，裁剪版本为 0.82。Atari 49 个游戏按全程平均奖励 PPO 赢 30 个，按最后 100 局 ACER 赢 28 个、PPO 19 个。

小跳人习惯表、“10 次跳 5 次改到 4—6 次” (对应概率比 0.8—1.2)、10 分和 50 分、分数曲线、集点卡比喻都是示意。聊天 AI 四步是 InstructGPT 式 RLHF 的简化，奖励来自人类比较训练的奖励模型。第 30 天的 RLHF 原论文用 A2C 和 TRPO，用到语言模型时才换成 PPO；今天另有 DPO、GRPO 等。

想深入可看本站的 [第 29 天中文导读](#)，代码见开源项目 [30u30 第 29 天目录](#) (英文)。