

每次只改一小步

本課講 2017 年 OpenAI 的一篇文章。John Schulman 和同事發明瞭 PPO：讓 AI 邊試邊學時，每次只把習慣改一小步。這樣學得又穩又快。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、幾個紙團、垃圾桶、2-5 位同伴

1 熱身：零食和怪獸

先填空。想一想小狗和遊戲裡的你。

訓練員說「坐」，小狗坐下了。訓練員給它一塊 _____。

下次聽到「坐」，小狗會 _____坐下。(更願意 / 更不願意)

碰到怪獸掉了一條命，下一局你會 _____怪獸。

沒有人給小狗講道理，也沒有人給你說明書。可是你們都學會了。辦法就是**試一試，看結果，再調整**。今天的主角 PPO，就是教 AI 這樣學習的一個辦法。

2 邊試邊學：強化學習

讓電腦邊試邊學，有個大名字，叫**強化學習**。「強化」的意思是：做得好的事，以後多做。

學習的那一方叫**智能體**。它可以是遊戲裡的小人，也可以是機器人。它所在的世界叫**環境**。它每做一個**動作**，環境就給它一個**獎勵分**。獎勵分可以是正的，也可以是負的。



🔄 再看下一個情況，重複幾百萬次

和第 1 課比一比：第 1 課的電腦猜字，每次都知道**正確答案**。強化學習沒有正確答案，智能體只知道**自己得了幾分**。

3 策略：它的做事習慣

智能體心裡有一張**習慣表**。表上寫著：在什麼情況下，做什麼動作，做的可能性有多大。這張表的大名叫**策略**。

比如一個闖關遊戲裡的小跳人。它的習慣表可能是這樣（示意）：

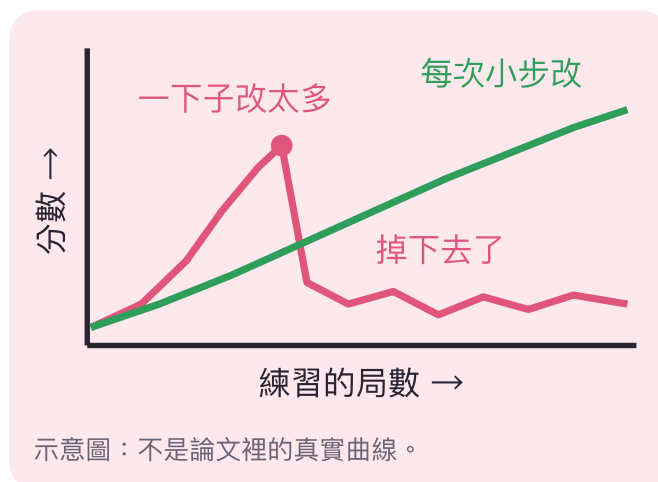
看到什麼	跳起來	往前走	蹲下
前面有個坑	10 次裡 5 次	10 次裡 4 次	10 次裡 1 次
頭頂飛來一隻鳥	10 次裡 2 次	10 次裡 3 次	10 次裡 5 次

遇到坑時，它跳過去了，得了 10 分。那就把「跳起來」調高一點。它往前走掉進坑裡，扣了 50 分。那就把「往前走」調低一點。**學習，就是一點一點改這張習慣表**。每次該改多少呢？

4 麻煩：一下子改太多，會崩

想想學騎自行車。你剛能騎一小段，有點歪。如果一次把坐姿、手和蹬車的力氣**全改了**，很可能摔得更狠。原來會的那一點也沒了。

電腦也會這樣。它根據剛玩的幾局改習慣，改得太猛，習慣就可能變壞。更麻煩的是：接下來的每一局，都是壞習慣玩的。小跳人一直掉坑，就學不到怎麼跳過去。分數掉下去，很難再爬起來。



論文裡真的試過：什麼限制都不加，讓它想改多少就改多少。結果在機器人走路的測試裡，它的平均成績比**隨便亂動**還差。

5 PPO 的辦法：每次只許小步改

PPO 定了一條規矩：**新習慣和舊習慣比，差別不能超過大約兩成**。這一招叫**裁剪**，就像用剪刀把多出來的部分剪掉。

還看小跳人。遇到坑時，它原來 10 次裡跳 5 次。這一輪學完，最多改到 10 次裡跳 6 次，最少改到 4 次。下圖橫線上的數字就是「10 次裡跳幾次」（示意）。



PPO 給每一次改動打分。在綠色範圍裡改得好，就算好處。跳出綠色範圍的那部分，**不再算好處**。就像集點卡每次最多蓋 2 個章，再多也不蓋。改過頭得不到好處，所以電腦沒理由改過頭。

還有一點很聰明：**上限只管好處，不管壞處**。如果改動讓習慣變差了，照樣扣分。電腦不會因為有上限，就躲過懲罰。

好處 1：學得穩。每一步都不大，原來會的本領不會一下子丟光。

好處 2：一批經驗能多學幾遍。論文的機器人實驗裡，每玩一批就複習 10 遍，不用每改一次就重新去玩。

好處 3：簡單。2015 年，Schulman 發明過一個類似的辦法，叫 TRPO。它要用很複雜的數學。PPO 只改了打分規則，程序只比普通辦法多幾行代碼。

6 結果：機器人走得穩，小遊戲打得好

機器人模擬 (MuJoCo)：在電腦裡的虛擬世界控制機器人走路、跳躍、保持平衡，共 7 個任務。PPO 幾乎每個都超過了老辦法，包括 TRPO。它還教會 3D 小人奔跑、轉彎，摔倒後爬起來。

Atari 小遊戲：49 個老式電子遊戲，比如打磚塊。看整個練習過程的平均分，PPO 贏了 30 個，最多。看最後的分數，另一個辦法 ACER 贏得更多。

誠實地說：PPO 不是每一項都第一。它的長處是**學得快、很穩、程序簡單**。所以大家都愛用它。

7 這和今天有什麼關係

PPO 很快成了 OpenAI 默認的強化學習辦法。2019 年，OpenAI 用大規模的 PPO 訓練出一支打 Dota 2 遊戲的 AI 隊伍。它打敗了當時的世界冠軍隊。

聊天 AI 也用到了它。聊天 AI 先學會猜下一個字（第 1 課）。可是會猜字，不等於會好好回答問題。所以科學家又讓它邊試邊學：

- ① AI 寫出幾個回答。
- ② 人挑出更喜歡的那個。
- ③ 按人的喜好給回答打獎勵分。
- ④ 用 PPO 小步調整 AI 的習慣。

2022 年，OpenAI 的 InstructGPT 就用了這個辦法。後來的 ChatGPT 也沿用了同一套流程。第 30 課會細講「怎樣讓 AI 懂得人的喜好」。

誠實地說：這不是唯一的辦法。後來的新辦法有的完全不用強化學習，有的去掉了 PPO 的一部分。不過很多新辦法還保留著它最核心的一招：**每次只許小步改**。

8 動手：小步改投籃

玩法（4—6 個人分成兩組，約 15 分鐘）

把垃圾桶放在地上。每人揉 1 個紙團，站在離桶 2—3 米的地方。每輪每組投 5 次，數投進幾個，一共投 3 輪。

小步組：每輪只改一個地方，而且只改一點點。比如往前挪半步，或者力氣大一點點。**大改組：**每輪可以隨便改。站位、力氣、姿勢，想換就全換。每輪投完，在表裡寫上投進幾個、改了什麼。

第幾輪	小步組投進	小步組改了什麼	大改組投進	大改組改了什麼
第 1 輪		(先隨便投)		(先隨便投)
第 2 輪				
第 3 輪				

你會發現：小步組的成績多半一點點往上走。大改組容易忽高忽低，原來投得準的地方也可能被改丟。PPO 選的就是小步組的走法。

9 動手：訓練小狗

玩法（3 個人，約 10 分鐘）

一人當**小狗**，一人當**訓練員**，一人當**記錄員**。訓練員偷偷寫下一個目標動作，只給記錄員看，比如「走到門邊摸門把手」。

小狗每次只能做一個**小動作**：走一步、轉身、抬手。訓練員不許解釋，只能說「熱」（更近了）或「冷」（離遠了）。做對了，就說「加 10 分！」記錄員記下用了幾次提示、幾秒，然後換人再玩。

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

你會發現：小狗不知道正確答案，只能靠「熱」和「冷」這樣的獎勵分來猜。聽到「熱」就接著試，聽到「冷」就退回來換個方向。這就是強化學習。

10 詞語小卡片

強化學習

智能體邊試邊學，做得好的事以後多做。

獎勵分

環境給動作打的分，可加可扣。越多越好。

策略

習慣表：什麼情況下做什麼，可能性多大。

裁剪 (PPO)

新舊習慣差別不超過約兩成，改過頭不加分。

11 想一想

1. 學騎車或學游泳時，為什麼一次把姿勢全改了，反而容易更糟？

2. 如果上限定得特別小，每次只許改一丁點，會有什麼壞處？

3. 聊天 AI 的獎勵分來自人的喜好。如果打分的人總給「說好聽話」的回答加分，AI 會學成什麼樣？

一句話記住：邊試邊學時，每次只改一小步，學得穩，原來會的也不會丟。

給家長和老師：本課講的是 John Schulman、Filip Wolski、Prfulla Dhariwal、Alec Radford、Oleg Klimov (OpenAI) 2017 年的論文 [Proximal Policy Optimization Algorithms](#) (arXiv)。PPO 把新舊策略的概率比裁剪到 $[1 - \epsilon, 1 + \epsilon]$ (論文取 $\epsilon = 0.2$)，再與未裁剪項取較小值，只去掉改進方向上超出範圍的收益，不是硬約束。MuJoCo 7 個任務上，不裁剪不懲罰的版本歸一化平均分為 -0.39 (隨機策略為 0)，裁剪版本為 0.82。Atari 49 個遊戲按全程平均獎勵 PPO 贏 30 個，按最後 100 局 ACER 贏 28 個、PPO 19 個。

小跳人習慣表、「10 次跳 5 次改到 4—6 次」(對應概率比 0.8—1.2)、10 分和 50 分、分數曲線、集點卡比喻都是示意。聊天 AI 四步是 InstructGPT 式 RLHF 的簡化，獎勵來自用人類比較訓練的獎勵模型。第 30 天的 RLHF 原論文用 A2C 和 TRPO，用到語言模型時才換成 PPO；今天另有 DPO、GRPO 等。

想深入可看本站的 [第 29 天中文導讀](#)，代碼見開源項目 [30u30 第 29 天目錄](#) (英文)。

極客教育 jikejiao.com 原創講義 · 2026-10