

左邊好還是右邊好

本課講 2017 年的一篇文章。OpenAI 和 DeepMind 兩個研究團隊的 Christiano、Leike、Brown 和同事們，不寫打分規則，只讓人挑「哪段錄像更好」，就教會了電腦做事。

適合 10 歲左右

約 40 分鐘

需要：鉛筆、白紙、2—3 位同伴

1 熱身：給「有禮貌」寫規則

同學問你：「你的橡皮能借我用一下嗎？」什麼樣的回答算**有禮貌**？請寫 3 條打分規則。

規則 1：_____ 規則 2：_____ 規則 3：_____

寫好了嗎？再看看下面兩個回答。不用規則，直接選：哪個更有禮貌？在你選的那個後面畫 ✓。

左邊	右邊
「請用吧，用完放我桌上就行。」 _____	「謝謝請您好不客氣對不起，不借。」 _____

如果你的規則是「說了‘請’和‘謝謝’就加分」，右邊那句反而得分更高！**寫規則很難寫全，可是二選一，一眼就能選出來**。這就是今天這篇文章的秘密。

2 難題：獎勵分從哪裡來

第 29 課講過**強化學習**：電腦一遍一遍地試，做得好就拿**獎勵分**，再往分高的方向改。遊戲屏幕上本來就有分數，很方便。

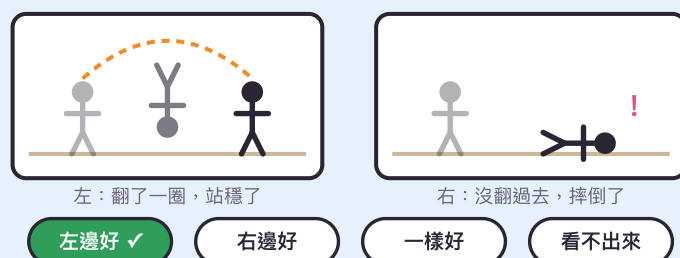
可是很多事情**沒有現成的分數**。比如，讓模擬的小機器人做一個**漂亮的後空翻**，或者讓聊天 AI 寫一段**有禮貌**的回答。

「漂亮」和「有禮貌」，怎麼算成分數？以前有兩條路，都不好走：

- **寫規則**：規則寫不全，電腦會**鑽空子**，就是分很高，事情卻沒做好。比如規則是「轉一圈加 10 分」，機器人可能躺在地上一打滾（我們編的例子）。
- **人一步一步打分**：電腦要練幾百萬步，人根本看不過來。

3 新辦法：只問「左邊好還是右邊好」

這篇文章走了第三條路：**不寫規則，讓人來比較**。電腦把機器人練習時的樣子錄成小錄像，每段只有 1 到 2 秒。人每次看兩段，只要回答一個問題：



示意圖：真的錄像裡是電腦模擬的機器人，不是小人。選「看不出來」的那一對，電腦不拿來學。

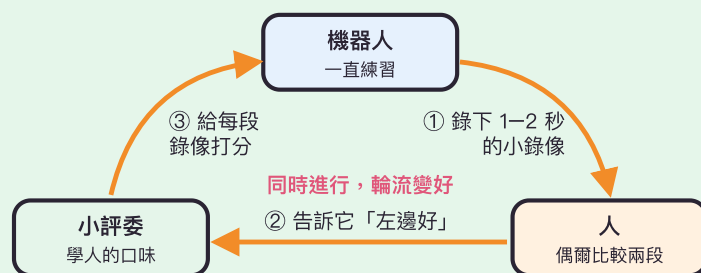
就像嘗果汁：給一杯打「87 分」很難，兩杯裡挑更好喝的，馬上就能說。**人不用會寫規則，只要會挑。**

4 小評委：替人猜喜好

人不能一直守在旁邊。所以電腦又練出一個**小評委**，它的工作是**猜人會喜歡哪一段**，給每段錄像打分。科學家管它叫**獎勵模型**，意思是「負責給獎勵分的模型」。

小評委怎麼學？和第 1 課猜字很像：先猜「人會選左邊」，再看人真正選了哪邊。猜錯了，就悄悄調整自己。

然後，機器人就用第 29 課的強化學習去練，獎勵分由小評委給。也就是說，**機器人在努力討好小評委**。下面這 3 件事同時進行，輪流變好：



5 結果：看得很少，學得很好

這個辦法**非常省人力**：機器人每練 100 步，人看過的還不到 1 步。

任務	人做了多少次比較	結果
8 個模擬機器人任務（走路、跳躍這類）	大約 700 次	接近直接看真實分數學出來的水平
7 個 Atari 老遊戲	大約 5500 次	乒乓球等遊戲接近看分數的水平。賽車遊戲還 超過了 看分數學的
一條腿的小機器人 後空翻	大約 900 次，人花了不到 1 小時	學會了連續後空翻

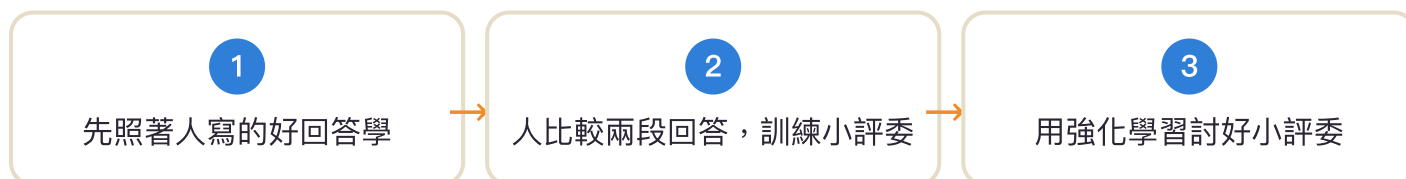
玩遊戲時，電腦**完全不看遊戲分數**，只聽小評委的。後空翻更是**沒有現成的打分規則**。當然，不是每個遊戲都這麼好。

一個鑽空子的真實例子：乒乓球

科學家試過只在訓練**剛開始**時請人比較。結果在乒乓球遊戲裡，電腦只把球擋回去，不去得分。小評委沒見過後來的情況，就猜不準了。所以**必須一邊練一邊問人**。

6 這和今天有什麼關係

這個辦法後來有了名字，叫 **RLHF**，意思是「根據人的反饋做強化學習」。**反饋**就是人給的意見。2022 年，研究者用它的升級版訓練會聽指令的 AI，大致分 3 步：



結果，一個小了 100 多倍的模型，回答反而更受人喜歡。同一年，ChatGPT 發佈了。

也要誠實地說：今天的做法有了變化。有時評委換成了 AI，比如 Anthropic（做 Claude 的公司）2022 年讓 AI 按寫好的原則來比較。有時省掉小評委。數學題和程序，可以直接對答案給分。問題也沒全解決：模型越拼命討好小評委，真實的好壞**先變好，後變差**。

30 天回顧：一條路走到這裡

第 1 課猜下一個字 → 第 13、15 課注意力和 Transformer → 第 22、25 課**越大越好** → 第 30 課聽懂人的喜好。最後一步，是學會做人真正想要的事。

7 動手：二選一評委

玩法（3 個人，約 15 分鐘）

同學問：「今天數學作業是哪幾頁？」每組有兩個回答。

1. 評委：讀第 1—3 組，每組只說「左邊好」或「右邊好」。記錄員把結果填進表裡。
2. 小評委：只看表，不問評委，寫出評委的 2 條喜好規律。
3. 小評委先猜第 4、5 組評委會選哪邊，再請評委選。猜中了嗎？

組	左邊	右邊	評委選	小評委猜
1	第 12 到 13 頁。	不知道，別問我。		—
2	你自己看黑板啊。	第 12 到 13 頁，最後一題不用做。		—
3	可能 12 頁，也可能 20 頁吧。	第 12 到 13 頁。		—
4	作業？我才不管呢。	第 12 到 13 頁，不會可以問我。		
5	反正在數學書上，你自己找。	第 12 到 13 頁，明天交。		

評委的喜好規律：_____

你會發現：評委從沒說出規則，小評委只看「選了哪邊」，也能總結出規律，還能猜中新的一組。電腦裡的小評委，做的就是這件事。

8 動手：小評委鑽空子

玩法（2—4 個人，約 10 分鐘）

假設有一個偷懶的小評委，它只學到了一條規律：**字越多越好**。它打分的辦法是數字數：一個字或一個數字算 1 分，標點不算。

每人回答「1 加 1 等於幾？」，想辦法從它那裡拿**最高分**。參考答案「等於 2。」只得 3 分。

把回答寫在白紙上。得分：_____ 真的好嗎？_____

你會發現：為了拿高分，大家會寫得又長又囉嗦。分數最高的，往往不是最好的。AI 只討好這樣的小評委，就會養成「說廢話」的壞習慣。

討論：怎樣讓小評委變聰明？提示：多給它看「又長又差」和「又短又好」的比較，一邊練一邊問人。

9 詞語小卡片

二選一比較

人不打分，只說兩段裡哪段更好。比寫規則容易得多。

小評委（獎勵模型）

從人的比較裡學會猜人的喜好，再給電腦的表現打獎勵分。

鑽空子

分數拿得很高，人真正想要的事卻沒做好。

RLHF

根據人的反饋做強化學習。今天的聊天 AI 靠它學會聽懂人的喜好。

10 想一想

1. 為什麼「給後空翻打 87 分」很難，「兩段裡挑一段」卻很容易？

2. 如果只在訓練剛開始時問人，小評委會出什麼問題？想想乒乓球的例子。

3. 你想讓 AI 學會「講笑話講得好笑」。你會怎樣請人幫忙比較？小評委可能被怎樣鑽空子？

一句話記住：寫不出打分規則，就讓人比較「哪個更好」，電腦從中學會人的喜好。

給家長和老師：本課講的是 Paul Christiano、Jan Leike、Tom B. Brown 等人（OpenAI 與 DeepMind 合作）2017 年的論文 [Deep Reinforcement Learning from Human Preferences](#)（NeurIPS 2017）：人比較 1—2 秒的片段，獎勵模型從比較中學習，策略用 A2C/TRPO 優化它，三者異步進行。人類反饋不到交互量的 1%；MuJoCo 8 個任務約 700 次比較，Atari 7 個遊戲 5,500 次比較（Enduro 超過真實獎勵訓練的 A3C）；Hopper 900 次比較、不到 1 小時學會後空翻；只在開頭標註時，Pong 智能體學會不丟分卻不得分。

熱身回答、「打滾轉圈」、果汁比喻、小人圖和活動問答都是示意。第 6 節依據 InstructGPT（2022，13 億參數模型的人類評估勝過 1750 億的 GPT-3）、Constitutional AI、DPO 和 Gao 等人（2022）的獎勵過度優化研究。想深入可看本站的 [第 30 天中文導讀](#)，代碼見開源項目 [30u30 第 30 天目錄](#)（英文）。