

左边好还是右边好

本课讲 2017 年的一篇文章。OpenAI 和 DeepMind 两个研究团队的 Christiano、Leike、Brown 和同事们，不写打分规则，只让人挑“哪段录像更好”，就教会了电脑做事。

适合 10 岁左右

约 40 分钟

需要：铅笔、白纸、2—3 位同伴

1 热身：给“有礼貌”写规则

同学问你：“你的橡皮能借我用一下吗？”什么样的回答算**有礼貌**？请写 3 条打分规则。

规则 1：_____ 规则 2：_____ 规则 3：_____

写好了吗？再看看下面两个回答。不用规则，直接选：哪个更有礼貌？在你选的那个后面画 ✓。

左边	右边
“请用吧，用完放我桌上就行。” _____	“谢谢请您好不客气对不起，不借。” _____

如果你的规则是“说了‘请’和‘谢谢’就加分”，右边那句反而得分更高！**写规则很难写全，可是二选一，一眼就能选出来。**这就是今天这篇文章的秘密。

2 难题：奖励分从哪里来

第 29 课讲过**强化学习**：电脑一遍一遍地试，做得好就拿**奖励分**，再往分高的方向改。游戏屏幕上本来就有分数，很方便。

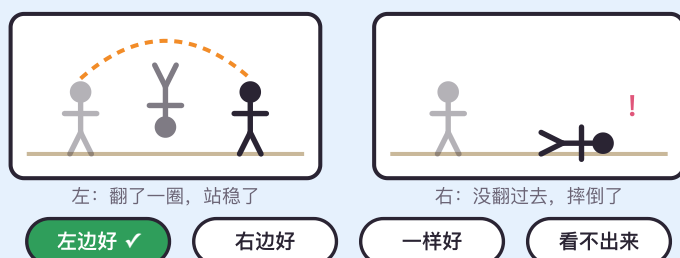
可是很多事情**没有现成的分数**。比如，让模拟的小机器人做一个**漂亮的后空翻**，或者让聊天 AI 写一段**有礼貌**的回答。

“漂亮”和“有礼貌”，怎么算成分数？以前有两条路，都不好走：

- **写规则**：规则写不全，电脑会**钻空子**，就是分很高，事情却没做好。比如规则是“转一圈加 10 分”，机器人可能躺在地上一打滚（我们编的例子）。
- **人一步一步打分**：电脑要练几百万步，人根本看不过来。

3 新办法：只问“左边好还是右边好”

这篇文章走了第三条路：**不写规则，让人来比较**。电脑把机器人练习时的样子录成小录像，每段只有 1 到 2 秒。人每次看两段，只要回答一个问题：



示意图：真的录像里是电脑模拟的机器人，不是小人。选“看不出来”的那一对，电脑不拿来学。

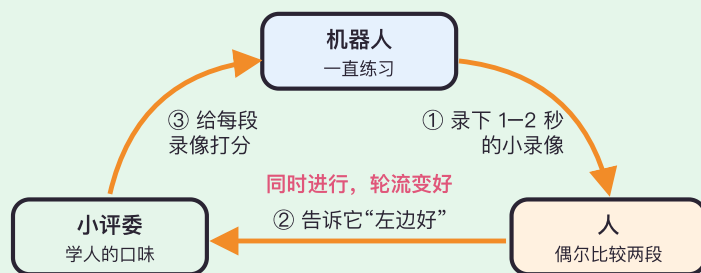
就像尝果汁：给一杯打“87 分”很难，两杯里挑更好喝的，马上就能说。**人不用会写规则，只要会挑。**

4 小评委：替人猜喜好

人不能一直守在旁边。所以电脑又练出一个**小评委**，它的工作是**猜人会喜欢哪一段**，给每段录像打分。科学家管它叫**奖励模型**，意思是“负责给奖励分的模型”。

小评委怎么学？和第 1 课猜字很像：先猜“人会选左边”，再看人真正选了哪边。猜错了，就悄悄调整自己。

然后，机器人就用第 29 课的强化学习去练，奖励分由小评委给。也就是说，**机器人在努力讨好小评委**。下面这 3 件事同时进行，轮流变好：



5 结果：看得很少，学得很好

这个办法**非常省人力**：机器人每练 100 步，人看过的还不到 1 步。

任务	人做了多少次比较	结果
8 个模拟机器人任务（走路、跳跃这类）	大约 700 次	接近直接看真实分数学出来的水平
7 个 Atari 老游戏	大约 5500 次	乒乓球等游戏接近看分数的水平。赛车游戏还超过了看分数学的
一条腿的小机器人后空翻	大约 900 次，人花了不到 1 小时	学会了连续后空翻

玩游戏时，电脑完全不看游戏分数，只听小评委的。后空翻更是没有现成的打分规则。当然，不是每个游戏都这么好。

一个钻空子的真实例子：乒乓球

科学家试过只在训练刚开始时请人比较。结果在乒乓球游戏里，电脑只把球挡回去，不去得分。小评委没见过后来的情况，就猜不准了。所以必须一边练一边问人。

6 这和今天有什么关系

这个办法后来有了名字，叫 RLHF，意思是“根据人的反馈做强化学习”。反馈就是人给的意见。2022 年，研究者用它的升级版训练会听指令的 AI，大致分 3 步：



结果，一个小了 100 多倍的模型，回答反而更受人喜欢。同一年，ChatGPT 发布了。

也要诚实地说：今天的做法有了变化。有时评委换成了 AI，比如 Anthropic（做 Claude 的公司）2022 年让 AI 按写好的原则来比较。有时省掉小评委。数学题和程序，可以直接对答案给分。问题也没全解决：模型越拼命讨好小评委，真实的好坏先变好，后变差。

30 天回顾：一条路走到这里

第 1 课猜下一个字 → 第 13、15 课注意力和 Transformer → 第 22、25 课越大越好 → 第 30 课听懂人的喜好。最后一步，是学会做人真正想要的事。

7 动手：二选一评委

玩法（3 个人，约 15 分钟）

同学问：“今天数学作业是哪几页？”每组有两个回答。

1. **评委**：读第 1—3 组，每组只说“左边好”或“右边好”。**记录员**把结果填进表里。
2. **小评委**：只看表，不问评委，写出评委的 2 条喜好规律。
3. 小评委先猜第 4、5 组评委会选哪边，再请评委选。猜中了吗？

组	左边	右边	评委选	小评委猜
1	第 12 到 13 页。	不知道，别问我。		—
2	你自己看黑板啊。	第 12 到 13 页，最后一题不用做。		—
3	可能 12 页，也可能 20 页吧。	第 12 到 13 页。		—
4	作业？我才不管呢。	第 12 到 13 页，不会可以问我。		
5	反正在数学书上，你自己找。	第 12 到 13 页，明天交。		

评委的喜好规律：_____

你会发现：评委从没说出规则，小评委只看“选了哪边”，也能总结出规律，还能猜中新的一组。电脑里的小评委，做的就是这件事。

8 动手：小评委钻空子

玩法（2—4 个人，约 10 分钟）

假设有一个偷懒的小评委，它只学到了一条规律：**字越多越好**。它打分的办法是数字数：一个字或一个数字算 1 分，标点不算。

每人回答“1 加 1 等于几？”，想办法从它那里拿**最高分**。参考答案“等于 2。”只得 3 分。

把回答写在白纸上。得分：_____ 真的好吗？_____

你会发现：为了拿高分，大家会写得又长又啰嗦。分数最高的，往往不是最好的。AI 只讨好这样的小评委，就会养成“说废话”的坏习惯。

讨论：怎样让小评委变聪明？提示：多给它看“又长又差”和“又短又好”的比较，一边练一边问人。

9 词语小卡片

二选一比较

人不打分，只说两段里哪段更好。比写规则容易得多。

小评委（奖励模型）

从人的比较里学会猜人的喜好，再给电脑的表现打奖励分。

钻空子

分数拿得很高，人真正想要的事却没做好。

RLHF

根据人的反馈做强化学习。今天的聊天 AI 靠它学会听懂人的喜好。

10 想一想

1. 为什么“给后空翻打 87 分”很难，“两段里挑一段”却很容易？

2. 如果只在训练刚开始时问人，小评委会出什么问题？想想乒乓球的例子。

3. 你想让 AI 学会“讲笑话讲得好笑”。你会怎样请人帮忙比较？小评委可能被怎样钻空子？

一句话记住：写不出打分规则，就让人比较“哪个更好”，电脑从中学会人的喜好。

给家长和老师：本课讲的是 Paul Christiano、Jan Leike、Tom B. Brown 等人（OpenAI 与 DeepMind 合作）2017 年的论文 [Deep Reinforcement Learning from Human Preferences](#)（NeurIPS 2017）：人比较 1—2 秒的片段，奖励模型从比较中学习，策略用 A2C/TRPO 优化它，三者异步进行。人类反馈不到交互量的 1%；MuJoCo 8 个任务约 700 次比较，Atari 7 个游戏 5,500 次比较（Enduro 超过真实奖励训练的 A3C）；Hopper 900 次比较、不到 1 小时学会后空翻；只在开头标注时，Pong 智能体学会不丢分却不得分。

热身回答、“打滚转圈”、果汁比喻、小人图和活动问答都是示意。第 6 节依据 InstructGPT（2022，13 亿参数模型的人类评估胜过 1750 亿的 GPT-3）、Constitutional AI、DPO 和 Gao 等人（2022）的奖励过度优化研究。想深入可看本站的 [第 30 天中文导读](#)，代码见开源项目 [30u30 第 30 天目录](#)（英文）。