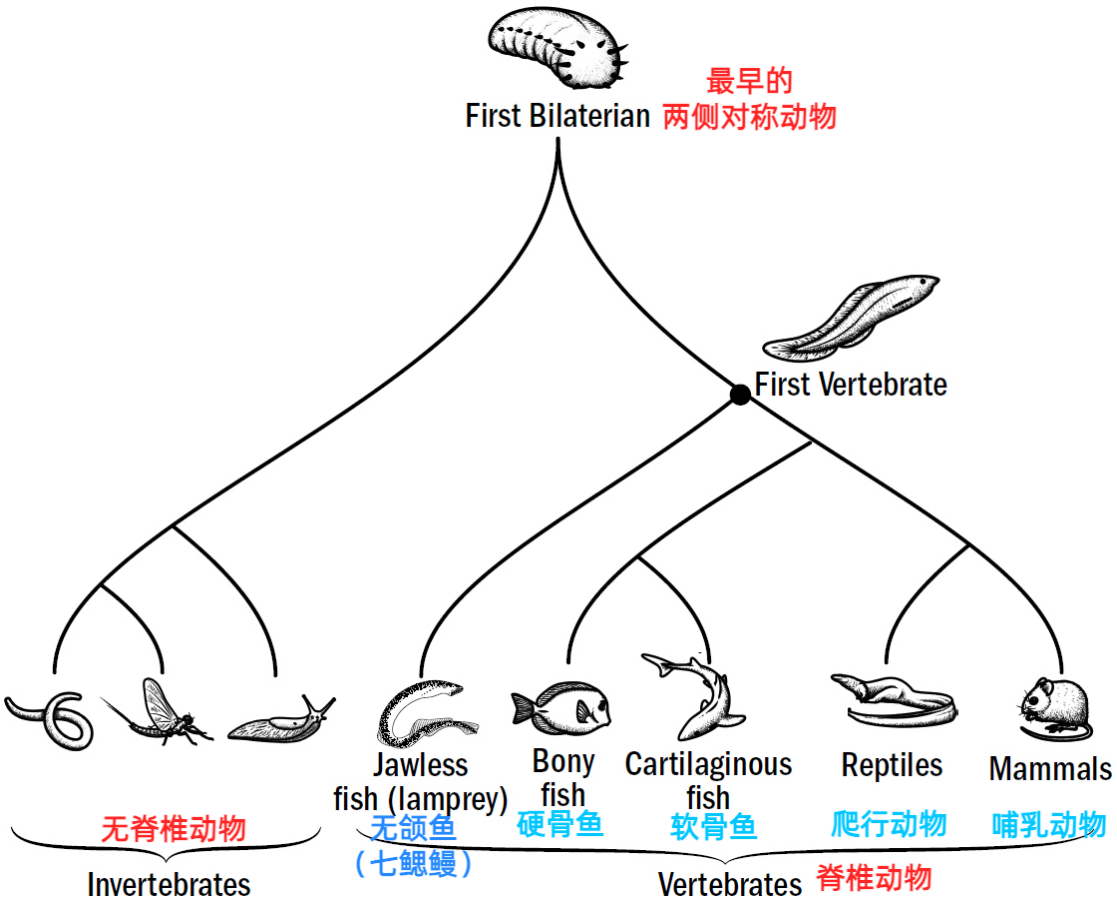
《智能简史》3：学习的革命

现在搞 AI 有个说法，模型的架构其实不重要，所有架构都很简单，只要你算力足够，参数多语量大，模型就能学会任何东西。但是我们考察生物智能的演化史，那可真不是这样。不是大脑算得快点还是慢点的问题，而是你没有那个结构就没有那个智能。进化史上往往是一次偶然的分岔，就让选错了路的这一支从此错过高级智能……

这一讲咱们说麦克斯·班尼特叙事中的**第二次智能突破，脊椎动物。**

自从地球上有了两侧对称动物，生物界就进入了一场军备竞赛。各个物种花样百出，各种捕猎与反捕猎，这就导致了所谓的「寒武纪大爆发」，一下子多出了很多物种。在大约 5 亿年前，上一讲说的线虫演化出了一个分岔，一边是脊椎动物，一边是无脊椎动物。



无脊椎动物这一支里出现了「节肢动物」，它在寒武纪是世界的统治者。海洋里的节肢动物可以长到一两米那么长，非常厉害。但是很不幸，他们这一支，包括所有的无脊椎动物，发展下去都没有多大前途 [*]。一直到今天，最聪明的节肢动物也就是螃蟹、蚂蚁、蜜蜂这些东西。

而幸运的我们属于另一支，这就是脊椎动物。最早的脊椎动物大约如下图所示，是个类似于现在的鱼一样的东西，只有几英寸长，在强手如云的寒武纪中很不起眼。

但是它大脑的主要结构，跟现在的鱼类，跟我们人类的大脑是几乎一样的：有皮层、基底神经结、丘脑、下丘脑、中脑和后脑。我们的大脑跟它唯一的区别就是在皮层之外又演化出了一个新皮层，那是以后的突破。

这么厉害的大脑，给脊椎动物带来了什么新智能呢？那就是学习。

✳

脊椎动物面临的是一个高度竞争的世界，局面极为复杂，靠线虫那点条件反射学习远远不够，你得学点高级本领。

19 世纪末，有个年轻的心理学家叫爱德华·桑代克（Edward Thorndike，1874—1949），想知道动物是怎么学习的。

他设计了一系列的笼子、迷宫之类的装置，用小鸡、小猫小狗甚至鱼类做实验，看它们怎么才能学会从中逃脱出来获得食物。有的是你得知道走哪条路，有的得会开锁，有的是按一个按钮，又或者只是做个特殊的动作 —— 比如小猫舔一下自己 —— 桑代克就会替它开门。

桑代克本来设想，动物应该是模仿学习：看别人怎么做的，记住，就像上课一样。但他发现不是这样。看是看不会的。

那些小动物都是自己先摸索，在里面到处走，这儿动一下，那儿动一下，偶尔一次成功走出来，得到奖励。然后它们下次面对同样的谜题就会解决得快一点，然后再快一点。

这也就是「试错」。连鱼都有这个能力。很多人说鱼只有三秒钟的记忆力，其实根本不是。鱼不但可以用试错的方式学会怎样从一个复杂的鱼缸迷宫中找到出口，而且哪怕过上几年之后，你再把它放回那个鱼缸，它还记得当初的路线！

你看这有啥难的？如果连鱼都能做到，AI 也应该能做到。

1950 年代迎来了第一波 AI 热潮，就有人借鉴桑代克的发现，设计神经网络用试错的方法学习走迷宫和下棋。你先随便试，具体怎么走我不管，反正赢了就给奖励输了就给惩罚，看看你能学会啥。

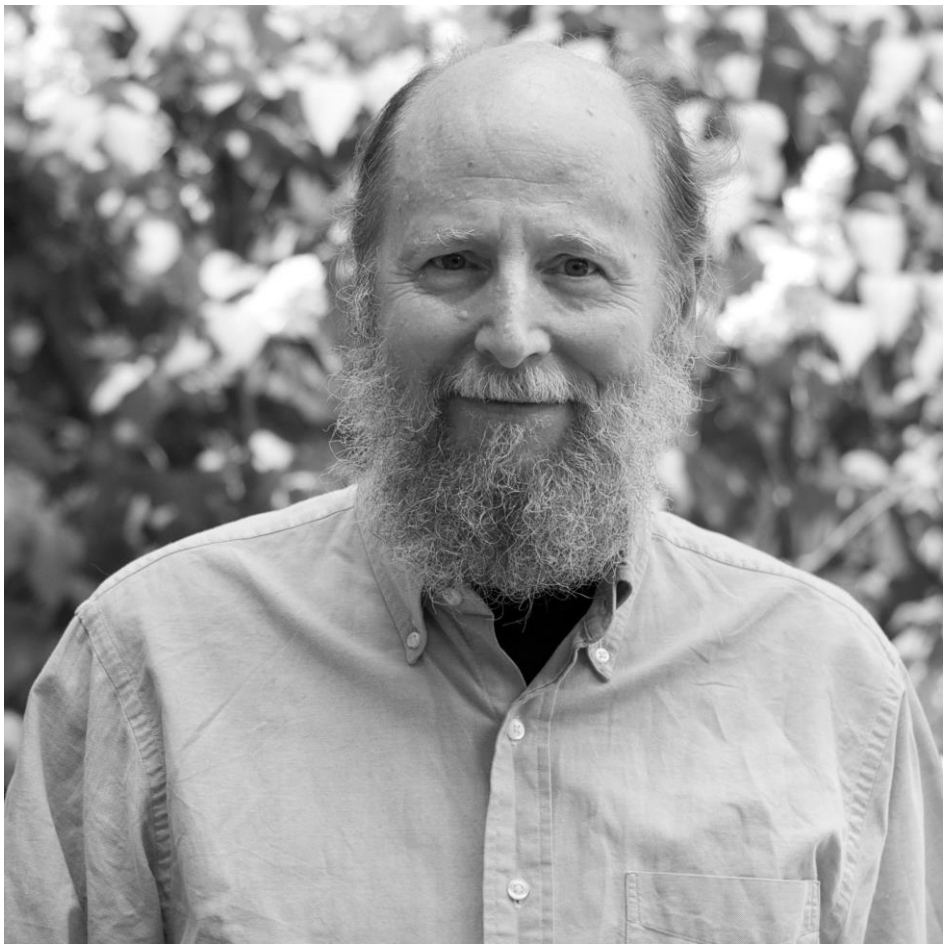
如你所能想见，这就是「强化学习」。

然而科学家很快发现，这种方法只对简单的问题有效。稍微复杂一点的东西，比如下棋，一局要走好几百步，走到最后才知道是赢是输，中间那么多步骤怎么学习啊？

此后三四十年间，没有人能解决这个问题。说着简单，可实际上人们不知道鱼到底是怎么试错的。

✳

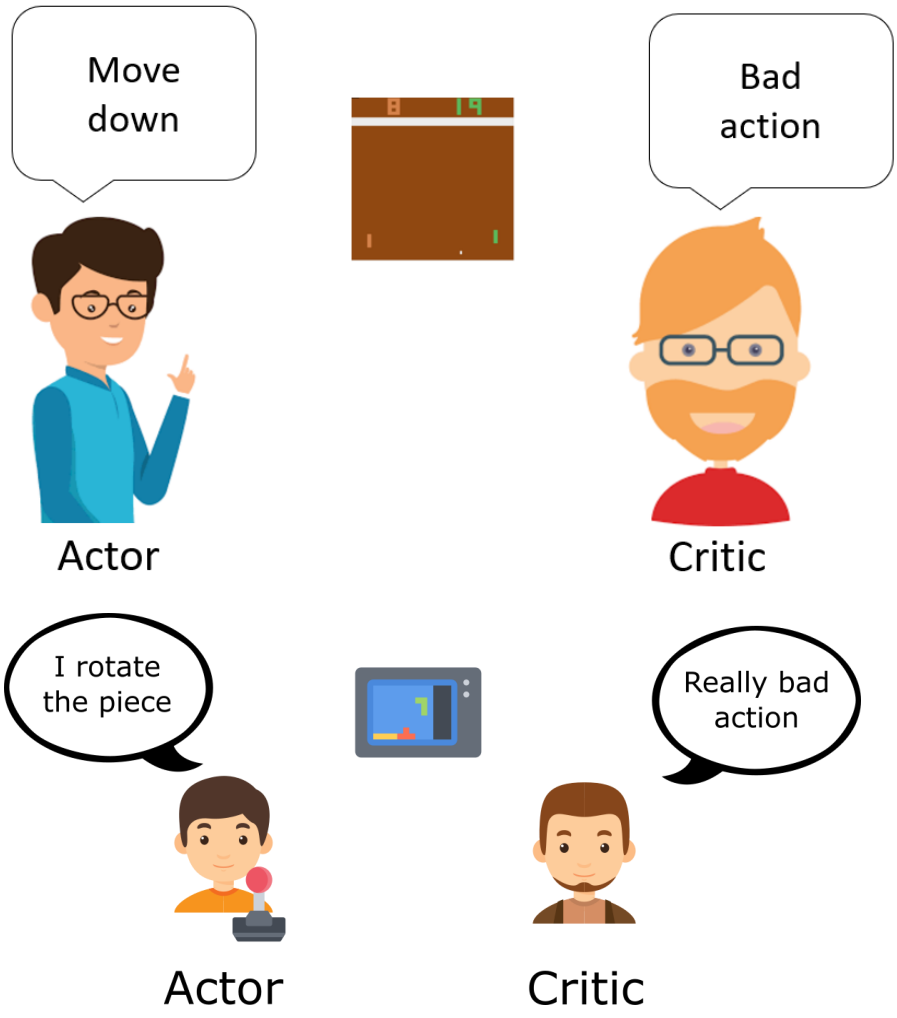
一直到 1984 年，才由一位大神级人物，理查德·萨顿（Richard S. Sutton），解决了这个问题。



我们专栏以前说过萨顿 [1]。他本科学的是心理学，从中获得了灵感。萨顿的洞见是，你不应该把最终结果作为奖励，而应该把中间每一步对结果的「预期」作为奖励。

简单说，萨顿把强化学习分解成两个独立的部分，各自训练：「行动者（actor）」和「批评者（critic）」。行动者每走一步，批评者都要预测这一步之后的全局取胜概率是多少。萨顿说我们应该强化的不是最终的胜负奖励，而是走完这一步后胜率的变化。

比如赢棋的概率本来是 51%，你走了这一步，批评者说现在赢棋的概率变成了 61%，那我们就说你这一步是个好棋，应该强化。



这样一来，我们学习参考的不是棋局最终的输赢，而是每一步是好棋还是坏棋，我们就等于是每一步都在学习！哪怕最后这局棋输了，你还是从中学到了很多步好棋。

这叫「时序差分学习（temporal difference learning）」，是现在强化学习的基本原理，包括 AlphaGo 也是这么做的。其实早在 1990 年代初，就有人用萨顿这个方法做了一个能下西洋双陆棋的 AI，达到了极高的水平。

那么，动物也是这样学习的吗？

✳

没错，动物的试错学习也是用的时序差分学习法。这是 AI 反哺脑科学最漂亮的一个例子，正是因为 AI 的研究在前，脑科学家才真正理解了多巴胺。

原来多巴胺是强化学习的关键。

起初人们以为多巴胺是一种奖励物质。在实验中，给猴子喂点糖水，猴子大脑立即产生多巴胺，似乎多巴胺代表「喜欢」。但这样进行几次之后，研究者发现猴子大脑不再是得到糖水*之后*释放多巴胺，而是在之前，在它*预期*会得到糖水的时候，大量释放多巴胺。

多巴胺是对好东西的预期，而不是好东西的奖赏。

而且这个预期可以量化。预期的好东西距离现在越近，预期获得好东西的概率越高，多巴胺释放得就越多。这恰恰就是萨顿的时序差分学习算法中对强化训练 AI 的信号的处理方法！

1997 年，有人结合 AI 的原理，用一篇论文彻底讲清楚了多巴胺的工作机制。多巴胺是一个强化信号，而不是奖励信号。多巴胺的作用是让我

们「想要」，告诉我们好东西就在附近，你现在的做法是对的，继续前进！

我们前面讲了，哪怕是最早的两侧对称动物，线虫，也有多巴胺。但线虫的多巴胺比较粗糙，只能告诉你附近有好东西。而脊椎动物的多巴胺则是一种量化信号：多巴胺越多，你就知道好事儿发生的可能性越高，时间越近，你的动力就越大。

这里面一个重要变量是时间感。脊椎动物有时间感，能精确感知两个事情间隔的时间长短 —— 而无脊椎动物，哪怕是其中比较高级的螃蟹、蜜蜂，都不能感知时间间隔，这就大大限制了它们的学习能力。

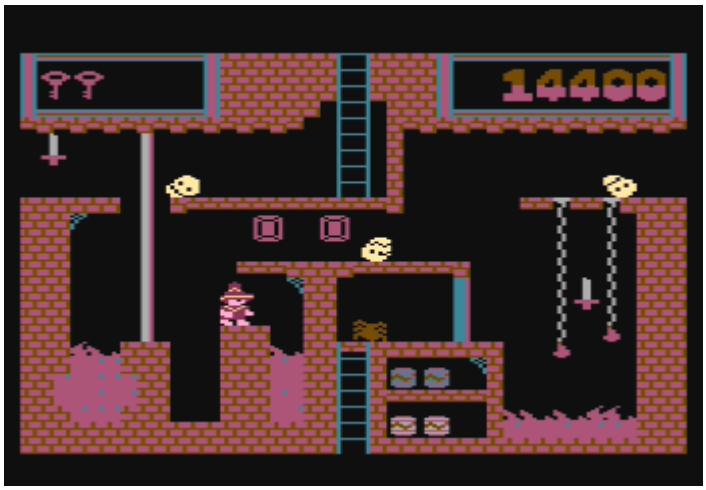
在脊椎动物的大脑中，下丘脑负责释放多巴胺。它是一个奖励系统，只看结果，认为是好东西就释放多巴胺。但大脑真正的学习机制不是释放，而是感知多巴胺，这一步由基底神经结负责。基底神经结中有两个回路，一个扮演行动者，一个扮演批评者。批评者负责感知多巴胺，它们共同学习。

猴子第一次喝到糖水的时候，下丘脑释放多巴胺，基底神经节就知道这个事件值得学习。几次之后，学习变得精确化，基底神经节的两个回路学会了判断奖励发生的概率，从而量化感知多巴胺，形成强化学习。

但这个故事还没完。

✳

强化学习算法取得了巨大成功，让 AI 在很多个电子游戏里都超过了人类水平。但是，有一个游戏，1984 年发售的《蒙特祖玛的复仇》（Montezuma’ s Revenge），AI 的水平却是怎么也上不去。



这是一个迷宫类游戏，要求玩家穿过一个充满障碍的房间，找到暗门出口。跟其他游戏相比，这里的暗门可以出现在任何地方，没有里程碑性的中间步骤，在找到暗门之前你根本不知道自己做对了还是做错了什么。

这就很不适合强化学习。强化学习为了试错，一般会留下比如说 5%的空间允许行动者随机做动作，但是在这里似乎远远不够。可你如果纯粹毫无章法地乱走，你又找得太慢。

一直到 2018 年，才由 Google 的 DeepMind 团队攻克了《蒙特祖玛的复仇》，打败了人类玩家。

他们的解决方法是引入「好奇心」。

我们专栏讲《为什么伟大不能被计划》[2] 一书的时候说过这个思想，这叫「新奇性搜索」。核心思路是如果一个动作虽然没有给你带来什么回报，但是它很新颖，让你探索了未知区域，满足了好奇心，那么这个动作也应该得到强化鼓励。AI 就是凭着好奇心，主动探索房间里没去过的地方，才找到暗门。

脊椎动物，以及一些后来演化出的高级无脊椎动物，都有好奇心。我们仅仅因为满足好奇心就能获得多巴胺。这就是为什么我们那么容易被随机的奖励所吸引，为什么我们在赌场里输着钱还那么投入。

强化学习有明确的目标，是一种非常功利的态度，它必须有好奇心的指引，才能走得远。你必须宁可牺牲一点回报，只为探索新的地方。

好奇心让学习本身成了一个值得追求的活动。

✳

事实证明强化学习和好奇心是最容易理解的学习方法，脊椎动物还有些特别重要的学习能力，我们至今没完全搞清楚。

这就是模式识别。昨天你在深海中遇到一条大鱼，差点被它吃掉，你记住了它。今天又遇到一个看上去很相似的动物，你怎么知道它跟昨天那条鱼一样是捕猎者，还是你潜在的交配对象呢？这就如同人类走在大街上，闻到一种混合了蛋白质和碳水化合物的味道，你怎么知道那是包子，还是某种不能吃的东西呢？

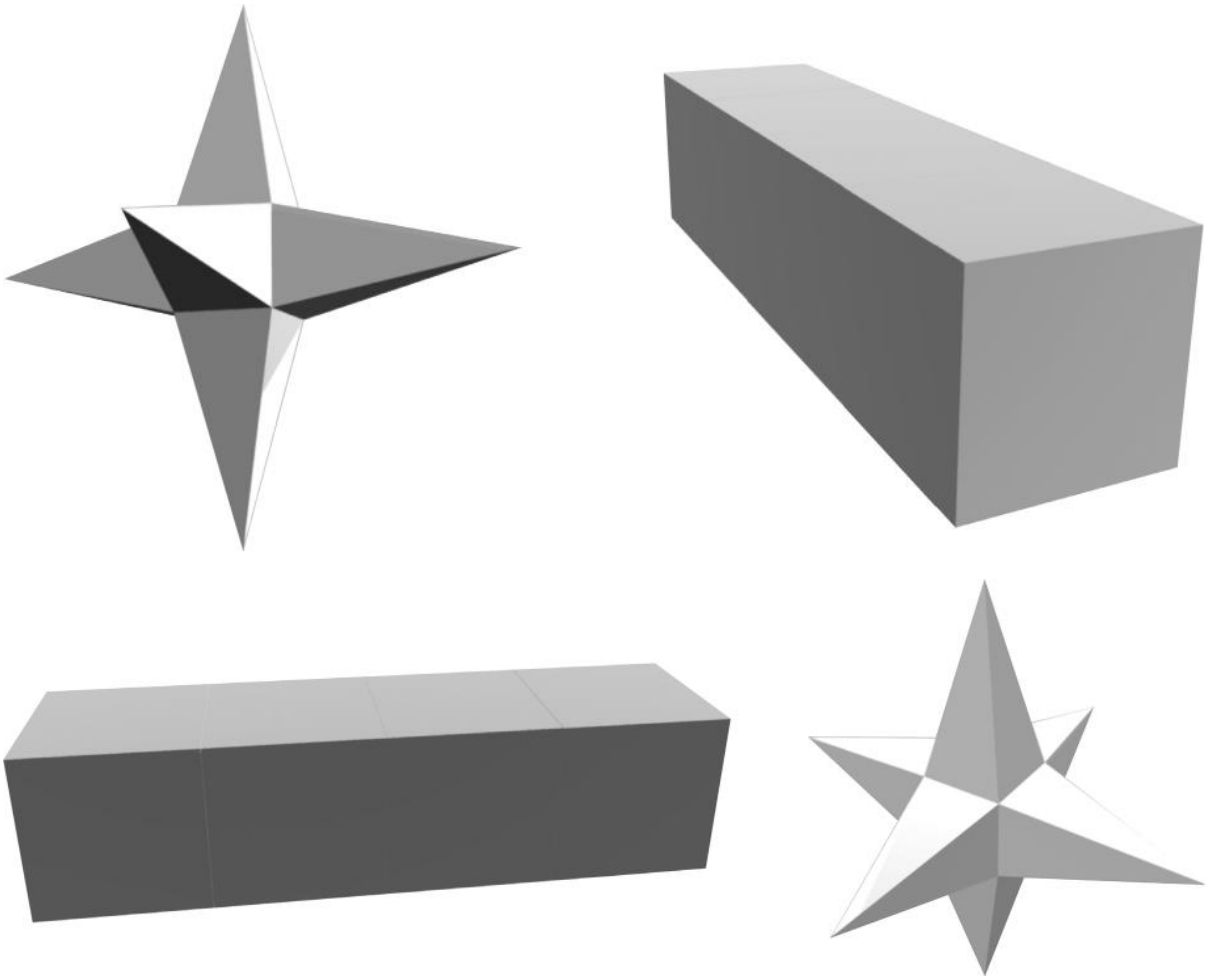
这些都是模式识别。脊椎动物并没有对每个特定物体都分配一个特定的神经元，我们收到的都是一大堆复杂的、互相有重复的信息，但是我们可以精确识别。这是怎么做到的呢？

对 AI 来说，解决方法是「监督学习」。这就是现在经常出来讲话的大佬，也是前 OpenAI 首席科学家伊利亚的授业恩师，杰弗里·辛顿（Geoffrey Hinton）的杰作。

你弄一个很大的神经网络，事先标记好每个训练素材是什么东西，每次训练根据输出结果反向更新网络参数。监督学习算法非常成功。比如手机识别人脸，你今天的脸跟昨天稍微有点不同，它也知道是你；换个跟你长得很像的人来，它也知道那不是你……

但脊椎动物的大脑，并不是用监督学习的方法搞模式识别的。没人给它们标记训练素材，而且碳基大脑不可能每训练一次就调整那么多神经元的参数。脊椎动物模式识别靠的是自动联想学习，这里咱们不讨论细节。

一个很著名的模式识别难题是下面这样。两张图，每张图中有两个物体，一个长方体一个棱锥体。你一看这两张图，就知道那大约是从不同角度看的相同的物体，对吧？



连鱼都能看出来。但是对计算机来说，这可太难了。换个角度画面就很不一样，你怎么知道那是从不同角度看的同一个物体，还是两个完全不同的物体的呢？

最终 AI 科学家是用「卷积网络」的方法解决的这个难题，但是我们很清楚，人和鱼的大脑肯定用的不是卷积网络：我们的神经元不是分那么多层的结构。大脑用的什么方法？我们不知道。

还有，现在包括最先进的 GPT 在内，所有大语言模型训练好之后，都必须把参数锁死再推向市场，不能随便继续训练。因为模型的训练必须非常小心才行，你搞不好学到新知识就会覆盖旧的知识，把参数搞乱了，这叫「灾难性遗忘（Catastrophic Forgetting）」。

可是脊椎动物从来都没有灾难性遗忘，鱼不会因为学了新技能就忘了旧技能。我们都是艺多不压身。

碳基大脑是怎么避免灾难性遗忘的？现在没人确切知道。

但我们的确知道，要实现某些特殊功能，需要特殊的硬件结构支持。比如脊椎动物大脑中有个海马体，就对我们形成空间感知的能力非常重要。你每次出门行动不会走固定的路线，你会抄近路，是因为你在海马体中构建了一幅世界地图，你知道自己所处的位置。蜜蜂和蚂蚁没有这个能力。

✳

这次突破最大的启发可能是强化学习和好奇心。原来试错是比模仿更基础的学习能力，原来好奇心不是个可有可无的情绪，而是生存发展的关键战略。

那如此说来，一些现代教育把孩子们关在教室里伸着脖子听老师讲，记住一大堆知识点而从不上手实干，不但不鼓励试错而且生怕惹事，连课外书都不让读，这岂不是在泯灭脊椎动物的天性吗？

我是脊椎动物，我骄傲。

注释

【*】无脊椎动物中的章鱼，被认为具有比我们想象的更高的智能，而且完全不同于我们，但这不是重点。

【1】AI 专题 18：算力就是王道

【2】精英日课第五季，《为什么伟大不能被计划》序言

划重点

1.第二次智能突破最大的启发可能是强化学习和好奇心。试错是比模仿更基础的学习能力，好奇心不是个可有可无的情绪，而是生存发展的关键战略。

2.强化学习有明确的目标，是一种非常功利的态度，它必须有好奇心的指引，才能走得远。你必须宁可牺牲一点回报，只为探索新的地方。