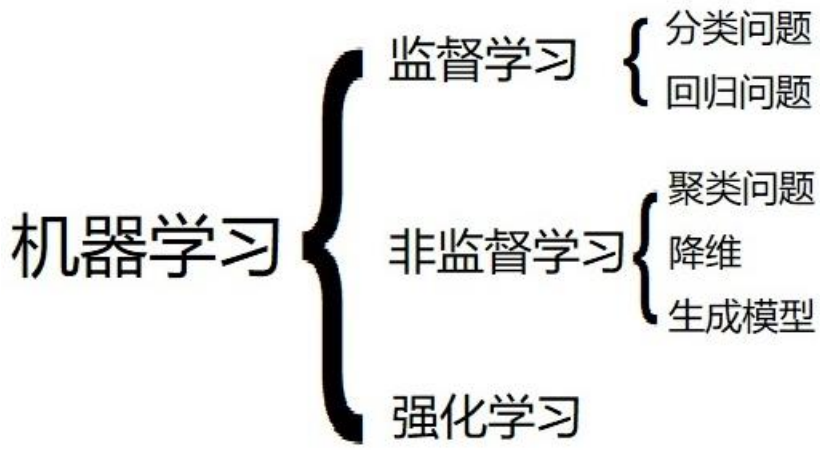
第一章 绪论

1.1 强化学习是什么

目前，在机器学习领域中，强化学习是一个非常热门的领域。在本书中，我们将介绍强化学习的方方面面。那么，首先要弄清楚的问题是“强化学习”这个词指的是什么。大家想必都听说过“深度学习”、“监督学习”、“非监督学习”、“半监督学习”这些概念。其中，“监督学习”与“非监督学习”指的是一类问题，而“深度学习”指的是一类方法。那么，“强化学习”究竟指的是一类问题呢？还是一类方法。这些同样含有“学习”二字的词语，与“强化学习”是什么关系呢？

有些前几年出版的机器学习材料中，会将机器学习的问题分为两大领域——监督学习与非监督学习（也称作无监督学习）。其中，监督学习意味着训练集同时存在着 x（feature）与 y（target），我们要学习 x 与 y 之间的映射关系。根据 y 是分类变量与还是连续变量，监督学习又可以细分为分类问题与回归问题；而非监督学习则意味着训练集只有 x，没有 y，它主要目的是研究变量 x 的一些内在结构，包括聚类问题、降维、特征提取、生成模型等具体问题。有的材料还会提及介于二者之间的半监督学习问题。但总的来说，过去很多材料会认为机器学习主要分为监督学习与非监督学习这两大类。

在最近几年新出版的机器学习材料中，一般会将“监督学习”、“非监督学习”与“强化学习”列为机器学习问题的三大领域。由此我们可以看出，“强化学习”与“监督学习”、“非监督学习”一样，指的是一类特定的问题。



那么，强化学习作为一类全新的问题，它是如何定义的呢？它和我们熟悉的分类、回归、聚类、降维有什么区别呢？为什么很多人认为，强化学习更加贴近人类的智能、是未来人工智能发展的方向呢？

如果要严谨地定义强化学习问题，就必须先定义 MDP，先后定义状态、动作、马尔可夫性、环境、奖励等，比较繁琐。我们将这部分留到第二章中。本章作为本书的绪论，我们将先通俗地讲解强化学习大致要做什么、其解题的基本思想是什么，它为什么重要，并说明本书的逻辑顺序。

我们先来看一个简单的例子——如果有一只“聪明狗”，我们如何让它学会叼住飞盘这一技能？

面对这个问题，人们通常会设计一个奖励机制，首先准备一个飞盘和一些聪明狗喜欢吃的肉，接着重复多次将飞盘扔出去。如果在这个过程中，聪明狗叼住了飞盘，那我们就给予一块它喜欢吃的肉作为奖励；如果它没有叼住，则我们不给予奖励，或者抽它一鞭子（给予负的奖励）。在一个明确的惩罚和奖励机制中，只要让聪明狗不断重复训练，就可以学会叼住飞盘。

如果从聪明狗的角度来看这个过程——叼飞盘的场地、训练它的人（包括“叼住飞盘可以获得肉”这条规则）可以视为外在环境。它目标是在给定的环境中吃到更多的肉（获得更大的奖励）。那么，通过在给定环境中不断地尝试，追求获得更大的奖励、更大的效用，它久而久之便会学会这项技能。

我们可以把聪明狗抽象成智能体，将叼飞盘视为某项技能，则上述过程可以抽象为智能体完成某个学习的过程。

下面，我们给出维基百科对于强化学习的描述：

强化学习是机器学习中的一个领域，强调如何基于环境而行动，以取得最大化的预期利益。其灵感来源于心理学中的行为主义理论，即有机体如何在环境给予的奖励或惩罚的刺激下，逐步形成对刺激的预期，产生能获得最大利益的习惯性行为。

在上面的描述中，我们可以看到强化学习的广泛性与重要性——强化学习是一种智能体在与环境进行交互的过程中进行学习的方法 ，主要研究作为主体的智能体与作为客体的环境交互的序列决策过程，以及主体在环境中逐渐学习到能产生最大的利益的习惯性行为的过程。

一般而言，我们会为强化学习问题定义如下几个元素：即智能体（agent）、环境（environment）、状态（state）、动作（action）和奖励（reward）。在某一个时刻，环境处于某一状态。智能体针对当前状态采取一个动作后，环境的状态发生改变，同时向智能体反馈奖励信息。策略是指面对状态应该如何采取动作。强化学习的目标是，通过与环境的交互，找到最佳策略，以获得最多的奖励。在宠物狗学习叼飞盘的例子中，可以将草地视为宠物狗学习叼飞盘的环境；将宠物狗与飞盘的距离视为状态；将宠物狗的跑、跳和叼视为动作；将肉视为奖励。宠物狗在草地上通过判断与飞盘的距离，选择不同的动作，不断试错努力尝试完成叼住飞盘的任务，从而获取更多的肉作为奖励。在第二章中，我们会将强化学习的问题通过更严谨的数学语言规范化为马尔可夫决策过程（MDP）的形式。

现实中有许多场景能够符合主体与客体交互的定义，例如机器人控制、无人驾驶汽车等等，它们都可以按照上述的方式定义为强化学习问题。我们以一个经典游戏——黄金矿工为例，它的目标是在规定时间内用钩子钩取黄金获取更多的分数以超过目标分数从而过关。我们将游戏本身视为环境，将玩家视为智能体，状态是当前屏幕上呈现出来的游戏情况，动作是所能采取的动作，包括“下钩”、“放炸弹”、“等待”三种。如果我们在恰当的状态采取了恰当的动作，能够获得的分数便是奖励。我们的目标是通过不断地进行游戏与环境交互，提高自己玩游戏的技巧，使得我们能够针对当前的状态更好地选择动作，使得自己能够获得更多的奖励。

下象棋的过程也同样可以定义为一个强化学习问题。要注意的是，下象棋涉及到二人的对弈，可以将对手及其走棋策略视为环境，而将自己视为需要训练提高的智能体，状态是当前棋盘的情况，动作是所采取的下法。当我们针对当前棋盘局势选择了一步走棋操作后，对方会根据我们走出的结果走棋改变场上的局势，然后又轮到我们针对新的局势选择走棋，这相当于环境给我们的反馈。我们的目标是通过与对手的对弈，观察场上的局势变化以更好地了解环境，从而争取战胜对手。

1.2 强化学习的基本思想

强化学习是机器学习中除有监督学习和无监督学习以外的第三大类方法。在机器学习中，有监督学习和无监督学习的特点是基于已有的数据，去学习数据的分布或蕴含的其它重要信息。强化学习与上述这二者最显著的不同在于，首先它不是基于已有的数据进行学习，而是针对一个环境进行学习。其次，它的目标不是学习数据中蕴含的信息，而是寻找能够在环境中取得更多奖励的方法。

概括地说，强化学习主要涉及两个部分：一是通过与环境交互产生大量的数据，二是利用这些数据去求解最佳策略。在给定数据集的问题中，我们往往只用考虑算法的计算量。而在强化学习中，我们不但要考虑算法的计算量，也要考虑产生数据所消耗的成本，即数据效率（data efficiency）。如何能高效地与环境交互产生数据（提升数据效率），并高效地求解最佳策略（提升训练效率），这也正是强化学习的困难所在。

下面，我们分别介绍这两个部分所用到的主要思想，以及其意义。

1.2.1 从环境中产生数据

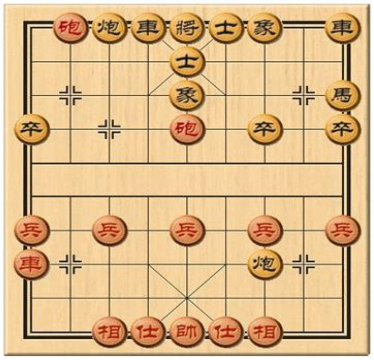
在有监督学习中，我们假定自然中有一个我们不了解的分布$P(Y|X=x)$，而我们有许多服从于P分布的数据。我们的目标是通过数据学习出P使得在测试集上的误差最小。而强化学习问题中，我们假定有一个能够自由与其交互产生数据的环境，

我们可以不断从环境中获取数据，以训练智能体在环境中的行为方式，使其获得更多的奖励。有监督学习中我们拥有的是数据，而强化学习中我们拥有的是环境。

有人可能不解——拥有数据与拥有环境有什么区别呢？即使强化学习中我们拥有环境，但我们仍然要通过与环境交互产生数据，最后通过数据来学习。如此说来，拥有环境和拥有数据岂不是一样的吗？

这里我们要特别强调的是，拥有数据意味着我们拥有环境中随机产生的数据。拥有环境意味着我们可以自主地选择与环境交互的方式，从环境中产生我们需要的数据。简单地说，有监督学习中的训练数据并不包含人为设计的成分；而强化学习的训练数据则包含了主观设计的成分，它无疑比随机产生的数据包含更多的价值。正因如此，我们可以专门选择环境中我们感兴趣的或对于目标有帮助的部分进行探索、根据需要来获得数据。这也正是强化学习中的探索-利用取舍（exploration-exploitation dilemma）发挥作用的地方。

比如，我们的目标是训练一个下象棋的智能体，使得它能够尽可能地在标准的象棋对局中战胜对手。下图中的s^1是一个正常对弈中很可能出现的局势。s^1中红方已经处于绝对优势，如果走法恰当，只需要两步就可以取胜了。即使走法不那么恰当，也几乎不可能被对手逆转。与之相比，s^2是一个非常罕见的局势，只有在专门设计的残局挑战中才会出现。在s^2中，红方处于极其危险的境地，只有步步紧逼的走法才有可能反败为胜。只要有一步的疏忽，就立即会被对手打败。如果单从技术难度的角度来说，局势s^2下选择走法无疑是更难的。但问题是，我们的目标是让智能体在标准的象棋对局中取胜，这两种局势对于我们的目标是否同样重要？



常见而重要的局势s^1

罕见而不重要的局势s^2

如果我们站在游戏规则本身来看，这两种局势都是游戏中可能出现的。假设对弈双方在符合游戏规则的条件下随机选择走法，则出现这两种局势的可能性或许不相上下。在这个角度下，s^1与s^2应该是同样重要的。但是如果两个有一定象棋基础的玩家对弈，则对局中几乎不可能出现s^2的情况。即使s^2局势下的走法非常精妙，

但是它对于我们的目标——在标准对局中尽可能地战胜对手——并没有那么重要。由于我们每次产生模拟对局的数据都是有成本的，所以我们应该产生更多有关s^1局势的数据，以提升算法的数据效率。

那么我们如何知道s^1与s^2相比更加重要？我们如何知道s^1是有可能出现的局势，而s^2则是完全不可能出现的局势？在训练刚开始的时候，智能体的参数是随机初始化而得到的。如果让它与人对弈，相当于是在随机地走棋，它也没有能力分辨出s^1与s^2这两个局势哪一个是更重要的。

为此，我们只有对智能体进行初步训练，当它有了初步的分辨能力之后，就能够判断并产生相对更加重要的数据。然后，智能体又会用这些数据优化自身的策略，并能够更准确地判断哪些数据更加重要。随后，它又能够更有针对性地从环境中产生更加有价值的数据，并进一步优化自身的策略。将这个过程迭代下去，便是一个越学越强的过程。这也正是强化学习这个名称的由来。

总的来说，“拥有环境”意味着我们可以源源不断地、有针对性地从环境中产生我们所需要的数据。换一个角度说，强化学习的目标是寻找能够在环境中取得较多奖励的策略。如果仅仅依靠给定的数据是难以求解出很好的策略的。只有拥有可交互的环境，才能充分验证策略的有效性。

有一类方法称作模仿学习（imitation learning）。它的思想简而言之就是将强化学习问题转换为有监督学习问题。以自动驾驶为例，我们需要从状态（包括车辆雷达检测的路况以及车载摄像头拍摄的周围环境图像）选择动作（包括转弯、加油、刹车）。一个很自然的想法是收集大量人类司机驾驶的数据，从数据中学习状态到动作的映射关系，这是一个典型的有监督学习问题。

对于一些现实中的复杂问题，模仿学习的实际应用效果很差，这是因为有监督学习与强化学习的性质不同。在有监督学习问题中，我们的目标是要优化真实值y与预测值$f(x)$之间的差距。模型对所有x给出的预测$f(x)$都难免会和实际上的y有误差。而在整个测试集上，总的误差由所有$f(x)$与y的误差线性叠加而成；而在强化学习中，假设我们训练出来的模型有误差，这会导致我们采取的动作A_1'与司机给出的动作A_1有一定的误差，继而导致进入的下一个状态S_2'也和S_2有一定的误差。然后由于我们采取的动作A_2'与司机给出的动作A_2有一定的误差，这会导致状态S_3'在S_2'的基础上有更大的误差。最后，我们得到的S_t'可能已经与人类司机所面对的S_t有了巨大的误差，甚至出现已有数据中完全没有的情况。既然在已有数据中完全没有出现过这种情况，那根据监督学习训练的智能体当然无法给出很好的应对方式。

为了在模仿学习的框架下解决这种问题，我们就不能仅仅用已经产生的数据进行模仿，而要找一个专家，让它对我们新产生的数据不断进行标注。当智能体遇到

已有数据中完全没有出现过的情况时，专家会为它进行标注、告诉他如果专家遇到了这种极端应该会采取何种措施。这就相当于，智能体训练用到的数据集在不断进行有针对性的扩展。

这种训练技巧被称为 DAgger（Dataset Aggregation），基本思路如下所示：

算法 1.1 DAgger

1: **repeat**
2: 通过数据集 $D = \{(o_1, a_1), (o_2, a_2), \dots\}$ 训练出策略 $\pi_\theta(a_t|o_t)$;
3: 执行 $\pi_\theta(a_t|o_t)$ 得到一个新的数据集 $D_\pi = \{o_1, o_2, o_3, \dots\}$;
4: 人工为 D_π 中的状态标上动作 a_t;
5: 进行聚合组成新的数据集：$D \leftarrow D \cup D_\pi$;

当采取了 DAgger 的技巧之后，模仿学习就不再只是经典的有监督学习问题，而是具有了一定强化学习的性质。这是因为，我们不再只是拥有给定的数据，而是拥有一个可以按需要不断产生数据的“环境”（能够为数据进行标注的专家）。

当然，模仿学习与强化学习性质不同，这里的“环境”和强化学习所说的环境还是有区别的——在强化学习中，面对着从未出现过的状态，智能体要去真实环境中亲身试错、获得正或负的真实奖励，以此来修正自己的行为；而在模仿学习中，智能体只需询问专家、再对着模仿就可以了。我们举这个例字的主要目的是为了说明，拥有固定的数据，或是可以不断交互产生新的数据，这二者是极其不同的。至于如何利用这些数据，那又是另一个问题了。

虽然我们最终要讨论的强化学习问题是持续多步的，但是直接对于这样的问题讨论如何产生数据，无疑是非常困难的。为了能够讲清楚如何产生数据这件事，在第三章中，我们会讨论多臂老虎机（Multi－Armed Bandit，MAB）问题。MAB 本身是一个简单而高度退化的强化学习问题。它有助于我们剥离强化学习其他方面的困难，专注于弄清楚如何面对未知环境产生训练数据的基本思想，这就是探索-利用困境（exploration-exploitation dilemma）

这里还要补充说明的是，强化学习中有一个分支领域叫做离线强化学习（Batch Reinforcement Learn）。在近年，它也成为了研究的热点。它的最大特点就是规定我们只有离线的数据，而没有可以自由交互、产生新数据的环境。在这类问题中，人们往往会尝试用已有的离线数据“模拟”出一个可以交互的环境。当出现离线数据中不存在的全新状态S_t时，人们会尝试用已有的离线数据“模拟”出与环境交互的结果。比方说，在 KADP 算法（kernel-based approximate dynamic programming）中，人们会利用离线数据与核函数来“模拟”出全新的数据。因此，

即使 Batch RL 中我们仅仅拥有离线数据，但它也和传统的监督学习有很大不同，具有部分强化学习的性质。

当然，由于 Batch RL 在 RL 中并不是最典型的一类问题，许多方法也比较前沿，所以本书中不会涉及到它。在后续说到的所有强化学习问题中，我们默认我们拥有一个可以自由交互、产生新数据的环境。面对未知的环境，如何有针对性地产生更加符合我们需要的数据，将始终是贯穿全书最重要的问题之一。

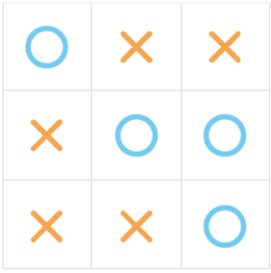
1.2.2 求解最佳策略

如何从环境中产生数据是强化学习中很重要的组成部分，这决定了学习的效率。但是，强化学习的最终目的在于求解环境中的最佳策略。面对未知的环境，我们需要通过智能体与环境的不断交互产生大量数据，并通过这些数据来学习最佳策略。如果我们既要考虑产生有价值的数据，又要考虑从数据中学习最佳策略，这个问题无疑会非常困难。为此，我们会尝试进行一些简化。

在上世纪的中叶诞生了一门叫做最优控制（Optimal Control）的学科，并被广泛视作是强化学习的前身。所谓的最优控制问题，简而言之就是要求解一个环境已知的强化学习问题的最佳策略。

要注意的是，即使环境是完全已知的——你清晰地知道自己做出的每一个选择，会有多少概率将自己引入哪一种情况——这也不意味着这个问题是平凡的。事实上，这是因为我们要面临的决策过程往往是“持续多步”的。正因为我们必须在持续多步的过程中最优化总的收益，所以问题是困难的。这一点我们在第二章定义 MDP 的时候还将仔细说明。这里仅仅举一个例子。

比方说，在第四章中，我们首先将用一个“井字过三关”的游戏举例——在规则清晰、简单的童年游戏中，我们定义了对手的走法（当你在同一直线上有两个棋子时，对手会在另一个位置上落子阻止你；否则，对手将随机选择一处落子）。在这种定义下，所有变化都已经在你的预料之中、出现每一种局势的概率都可以计算出来。那么，你能很简单地找到让你获胜概率最大的策略么？答案是否定的。如果你仅仅凭着直觉判断，很可能在第一步就做出错误的选择（读者不妨先试试想一下自己第一步应该在何处落子，等看到第四章时候再揭晓答案）。



在第四章中，我们将讲解最优控制问题。由于在这类问题中环境是已知的，所以我们不必考虑如何产生数据，可以将精力集中于如何求解最佳策略的思想与技巧上。在这一章中，我们将看到强化学习的两种主要思路——基于价值的方法与基于策略的思路，二者有什么不同。我们可以在相对纯粹的角度下，去思考基于价值与基于策略这两大类方法的本质，而不用受到其它的干扰。

总的来说，强化学习针对的是一个未知环境求解最佳策略的复杂问题。在强化学习中，我们既需要与环境交互产生数据，又需要通过这些数据去求解能够取得最大奖励的最佳策略。因此，我们必须兼顾上述两个方面，同时提升训练的数据效率与计算效率。在本书中，为了便于大家理解强化学习方法，我们会力图先将强化学习这两个方面的难点分开讲解。在第三章中，我们将先讲解 MAB 问题，介绍如何从环境中产生更有价值的数据；而在第四章中，我们将介绍最优控制问题，介绍如何在环境完全已知的情况下，求解强化学习问题的最佳策略；完成了这两个部分的讲解之后，我们再开始正式讲解强化学习有关的算法。

1.3 强化学习为什么重要？

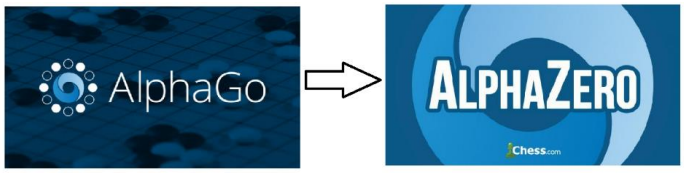
目前，强化学习是人工智能学术界受到关注最多的话题之一。那么一个很重要的问题是，强化学习为什么这么重要？

有监督学习是机器学习领域中最基础的问题。在这类问题中，我们的目标是利用给定的训练集(X,Y)去拟合函数$Y=f(X)$，使得预测误差尽量小，其中根据Y是连续变量还是分类变量可以分为回归问题与分类问题。有监督学习是最直观易懂的，所以早期机器学习研究的重点在于有监督学习。但是有监督学习要求数据中有标注，标签Y与数据X具有不同的地位，它代表着高度概括的信息，有价值的知识。然而标签的获取代价往往较为昂贵。

无监督学习主要研究数据的内在结构，它不需要标注，可以直接从数据中进行学习。具体而言，无监督学习分为目标不同的几类问题。例如生成模型寻找的是X的分布$P(X=x)$，比较有代表性的成果是生成对抗网络（GAN）。降维问题寻求用

低维的随机变量Z来表征X，代表性成果是经典的主成分分析（PCA）与自动编码器（AutoEncoder）。聚类问题寻求将数据集X分为几类的方法，代表性成果是 K-Means、层次聚类等等。由于无监督学习适用于各种数据，其成本较低，适用范围较广，所以近年来受到较多的关注。

与无监督学习一样，强化学习得到关注的重要原因是在生产生活中有许多任务可以被视为强化学习问题。例如，如果要训练一个下围棋的智能体，围棋的规则、棋盘局势、落子方式等相互关系很自然地可以被定义为一个强化学习问题。2016 年初，第一次战胜人类围棋冠军的 AlphaGo 就利用了有监督学习与强化学习相结合的算法，它使用了大量的顶级高手对局的数据（包含了棋盘局势与顶级高手走法的对应关系），让智能体去学习。而 2017 年底，研究者们推出了用纯粹强化学习训练的 AlphaZero。它只利用强化学习了解围棋规则和下围棋的方法，并没有使用任何顶级高手对局的数据。最终，AlphaZero 利用比 AlphaGo 更少的数据与运算代价，在对局中取得了压倒性胜利。



监督学习+强化学习
计算量大、战斗力低

强化学习
计算量小、战斗力强

开始人们的想法是让智能体学习人类顶级高手的下法，这就相当于将这个问题强行转化为有监督学习问题。但是归根结底，围棋取胜的标准不是下法是否和顶级高手接近，而是按照规则是否能赢。所以，智能体学会下围棋的过程应该是强化学习问题，而不是有监督学习问题。强行将其转化为有监督学习问题，不但需要大量有标注的数据集，而且效果也不好。要注意的是，这种所谓的效果不好并不只是我们在讲解模仿学习时候所说的，由于模仿的误差带来的效果不好。即使我们真的能让智能体模仿得与人类围棋大师完全一样，也无法超越它们。

在强化学习中，我们拥有的不是数据，而是环境。我们可以从环境中产生数据，但我们的最终目标不是学习数据背后的规律，而是要让智能体能够在环境中获得更多的奖励。从这个角度上说，训练围棋智能体、机器人控制、自动驾驶等问题都应该用强化学习的方法解决。

但是，如果仅仅只有上述的原因，强化学习恐怕很难得到如此广泛的关注。按照我们前面所讲的逻辑推断，人们应该对特定的问题采用特定的方法。但是有许多学者将有监督学习问题或无监督学习问题也采取强化学习方法来解决，这又是为什么呢？

人们关注强化学习还有一个重要原因——因为强化学习所采用的思维方式是与有监督学习、无监督学习完全不一样的，它比这两者更加接近现实中生命体的学习方式，更加智能。有的研究者甚至认为，强化学习很可能是通向强人工智能的重要路径。我们所说的强人工智能到底是什么呢？维基百科的解释如下：

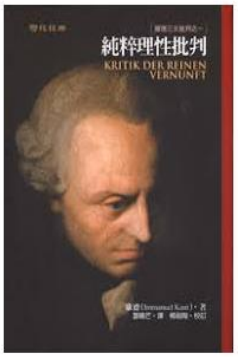
强人工智能是人工智能研究的主要目标之一，同时也是科幻小说和未来哲学家所讨论的主要议题。相对地，弱人工智能只处理特定的问题，不需要具有人类完整的认知能力，甚至是完全不具有人类所拥有的感官认知能力，只要设计得看起来像有智慧就可以了。而强人工智能也指通用人工智能，或具备执行一般智能行为的能力。强人工智能通常把人工智慧和意识、感性、知识和自觉等人类的特征互相连接。

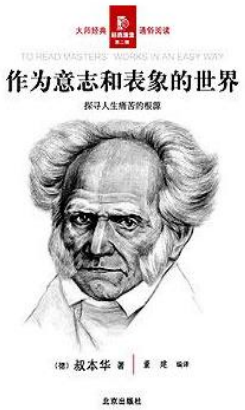
弱人工智能只能实现特定的任务，像是一个辅佐人类的工具。采用有监督学习与无监督学习的算法或许可以得到一台高级的机器，但是它本质上还是一台机器、只能用于特定的任务。而强人工智能则追求真正建立一个和人类一样能够自己感受、思考、行动的智能体，真正用机器造出一个具有智能的人。目前，强化学习虽然远远没有实现强人工智能，但它的思维方式和有监督学习及无监督学习相比完全不同，无疑更加接近强人工智能。

为什么说强化学习的思维方式更加接近生命体呢？下面，我们粗略地讲一些哲学上与之相关的概念，给大家一个启发。

在古典时代，哲学家们更加关注的是世界的本质。无论是毕达哥拉斯的“万物皆数”，德谟克利特的“原子论”还是柏拉图的“理念论”，都是对于世界的不同认识方式。他们都在追求能够更加正确地认识世界。这些以寻求世界本质为目的的理论均被称作“本体论”。我们可以想象，有监督学习或无监督学习追求探索数据背后隐藏的规律，二者和“本体论”具有相似甚至相同的思维方式。

而在近代，哲学经历了重要的“从本体论向认识论”的转向。其代表是康德对于“本体”与“现象”的划分。如果用通俗的话来说，“真实世界”和“人眼中的世界”是两个不同的东西。在此基础上，康德认为“真实世界”是不重要的，“人眼中的世界”才是值得我们关注的重点。这种让“真实世界”适应我们认识的想法，是近代哲学中一个极其重要的转向。





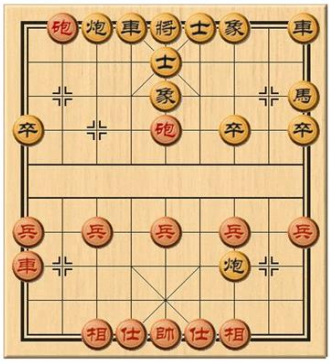
德国哲学家叔本华继承以及进一步发展了康德的理论。他将“人眼中的世界”称之为“表象”，而将人的本能称作“意志”，并且认为“表象”是“意志”外化出来的。简而言之，他认为人对于世界的认识是被人的目的所支配的，或者说，“人如何认识世界”是受到“目标是什么”所支配的。叔本华认为，人们经历的世界，不是真实世界，而是人的“意志”在“真实世界”中产生的“表象”，我们是时刻被“意志”和“目的”所支配的。“意志”是非理性的生物本能。“意志”决定了我们想要什么，而“理性”只是帮助我们设法得到它。

我们尝试用叔本华的这套理论去解释人类认识世界的过程：在原始社会，人们为了能够活下来就要去打野觅食。受这个目的支配，人们在一次次实践的过程中逐渐认识到哪些是不好对付的动物，需要进行团队协作，哪些是容易对付的动物，自己一个人也可以应对……后来，人们发现种田能够养活更多人，并且进行试验研究出哪些谷物更加适合播种……再后来，随着农业社会的发展与剥削阶级的产生，酋长与祭司脱离体力劳动，开始观测天空中的星座，思考世界的本质……如此说来，人类拥有打猎、种田和观测星座的能力是存在先后次序的，而这是由人的目的决定的。人类先有了目的，再在它的支配下学习到了各种技能。

如果我们把这种人的“目的”、“目标”或“追求”，也就是叔本华所说的“意志”定义为“最大化累积奖励”，那么人应该是在时刻追求“最大化累积奖励”的动机支配下的，所有的能力都是为了满足这个动机的手段。“正确认识世界”其实也是一种能力，本质上也是追求“最大化累积奖励”在某些具体的场景中的应用。如果追求不同，认识到的世界也会不同。

在有监督学习或无监督学习中，我们假设现实世界有一个未知的分布，代表着大自然的规律。训练数据是由这个未知分布产生的，我们的目标就是通过这些数据去学习出隐藏在数据背后的未知分布，这就像是通过知识去探求大自然中隐藏的真理。在这个过程中，我们对于所有的知识一视同仁，目标是平等地认识大自然中存在的所有客观规律，我们选用的不同模型就相当于认识世界的不同方式。

在强化学习中，我们的目标不是学习出环境而是最大化累积奖励。在这个过程中，我们会按照需要去产生数据。利用在上一节中举的一个例子：在求解下象棋的策略时，s^1是一个常见而重要的状态，而s^2是一个罕见且不重要的状态。它们都是世界的组成部分，地位应该是平等的。但是，由于我们的目标是在标准对局中以最大可能取胜，所以我们会更加重视s^1。大千世界中有无数的信息，而我们只关注感兴趣的或者和生活息息相关的信息。利用这种强化学习的思维模式，我们可以更加高效地完成真正需要完成的任务。通过这个简单的例子，希望大家能够理解强化学习思维方式之中的独特之处，理解强化学习与监督学习相比，是更加接近现实中智能体的学习方式的。



常见而重要的局势s^1

罕见而不重要的局势s^2

强化学习的过程就像是在最大化累积奖励的目标支配下去探索环境，选择环境中对自己有用的知识加以学习。这个过程更加强调人的主观能动性在认识世界、改造世界中起到的重要作用。强化学习比起有监督学习或无监督学习更加接近一个生命体的学习过程，更加具有智能性，更加接近强人工智能。弱人工智能本质上只是一个辅佐人的机器，而不是具有生命的智能体。在人类需要的时候，它能完成具体的任务，但是在人类没有下指令的时候，它便什么也不做；而强人工智能应该是一个时时刻刻都在追求收益最大化的机器人，在这个终极追求下，它可以不依照人类的命令，而按照自己的方式去行事。在这个过程中，它或许会为了自己的目的去探索环境，去学习如何完成具体的任务。驱动它完成任务的目的不是外在定义的，而是它的内在动机决定的。总的来说，它更像是一个有生命的智能体。强化学习显然是更加符合我们对于强人工智能的期待的。

总的来说，强化学习在当下得到广泛关注的主要原因是因为它适合于很多的应用场景，不像有监督学习那样需要大量昂贵的、有标签的数据。与有监督学习相比，强化学习的思维方式能够更好地描述人类学习的过程，更加接近具有生命的主体的行为；有许多研究者们认为强化学习很可能是通向强人工智能的方式，这也增添了强化学习的魅力。正是由于上述这几个因素，强化学习才能在今天得到如此多的关注，成为如此重要的一个研究方向。理解这些特性，对于后文中理解强化学习的具体方法也是很重要的。

1.4 本书内容介绍

以上，我们尝试用一些相对通俗的语言讲述了强化学习的基本思想。在后面，我们将结合更加严谨的数学语言以及案例，将强化学习的具体方法介绍给大家。

在第二章中，我们将介绍马尔可夫决策过程（MDP）的定义，将强化学习问题形式化为数学模型，给强化学习要解决的问题一个最基本的数学定义。

在第三章中，我们将讲解多臂老虎机（MAB）问题，去理解未知环境带来的困难以及如何产生训练数据的技巧。

在第四章中，我们将讲解最优控制问题。本书中所介绍的最优控制的含义可能和传统意义上的最优控制有所区别。本书中的最优控制指的是求解环境已知的 MDP。我们将通过最优控制问题初步讲明强化学习的两大类基本的思路——基于价值与基于策略的基本思路。要注意的是，这里“基于价值”与“基于策略”指的是算法的一类基本思路，而不是具体的算法。事实上，有很多具体的算法同时融汇了这两种思想，很难清晰地进行归类。

第三章与第四章分别相当于在相对简化的条件下考虑“如何从环境中产生数据”与“如何求解最佳策略”。在讲解了这两个方面的基本思想之后，我们就可以初步地了解强化学习算法的基本框架了。下面我们就开始正式介绍强化学习问题。

在接下来的三章中，我们将讲述线下最流行的 Model-Free 方法。我们主要依据算法的形式（或者说算法的组建构成）将其分为三类——价值方法，策略方法，以及 Actor-Critic 型的方法，分别进行讲解。要注意的是，这里分章的逻辑是按照算法的形式与构成，而非算法背后的基本思想。因为倘若非要按照算法背后的思想来分类，则许多算法都很难被清晰的归类。

在第五章中，我们将介绍无模型（model-free）方法中的价值方法，包括 Q-learning、Sarsa、多步 Sarsa、DQN 和 NAF 等。如果就基本思想而言，你很难说

Sarsa 方法究竟是“基于价值”的还是“基于策略”的。但是，由于上述所有算法中我们只需要训练一个价值网络（或价值表格），所以我们将其归为一类来讲；

在第六章中，我们将介绍无模型方法中最基本的策略方法，包括无梯度方法、Vanilla Policy Gradient，为“基于策略”的思想做一个引入。这一类算法当然并不能代表“基于策略”这种思想的全部，但它们的特点是只需要训练一个策略网络，而不需要其他任何的辅助部分，所以被单独归为一章。

在第八章中，我们将介绍无模型方法中另一大类的算法——Actor-Critic。这一类方法最大的特点是人们需要同时训练策略网络与价值网络，所以被归为一类。不过，如果从基本思想上看，它们大多应该被归类为基于策略的思想。这一类算法包括 Actor-Critic、A3C、A2C、TRPO、PPO 以及 DDPG。不过，其中 DDPG 方法既可以看成是基于价值的思想，也可以看成是基于策略的思想。但是毋庸置疑的是，它需要同时训练价值网络与策略网络。因此，我们将其归入这一章。

在第九章中，我们将讲述基于模型（model-based）方法。这不是本书的重点内容，所以我们不会在这上面花费太多的篇幅。但这并不意味着它不重要。事实上，基于模型的方法才是未来强化学习发展的重点方向。我们会尝试为 Model-Based 作一个概述，并讲解几个经典的例子，以便于读者日后能更好地学习这个方面的知识。