强化学习：智能体学习在环境中采取一定动作以最大化奖励的策略。

顺序决策（Sequential Decision Making）：选择一系列顺序的动作以最大化总奖赏

历史 (History): 是观察observations, 动作actions, 奖赏rewards组成的序列

状态 (State): 通过历史来确定接下来会发生什么的信息(动作、观察、奖赏)

状态是历史的函数

策略 (Policy)：从状态到动作的映射

值函数（Value function）：表征在某个策略𝜋下获得的累计总奖赏

状态值（State Value）：评价某个状态有多好

动作值（Action Value）：评价在某个状态下，采取某个动作有多好

模型 (Model)：预测接下来的状态或奖赏

有模型：奖赏和策略（状态转移矩阵）等已知，在已有的模型上优化，无需与环境交互。

无模型：奖赏和策略（状态转移矩阵）等未知，通过与环境交互学习。

* 基于价值的学习 (value-based)：从值函数中学习到策略
* 基于策略的学习 (policy-based)：有策略，无值函数
* Actor-Critic：有策略、有值函数

权衡

利用 (exploitation)：基于已有的经验，选择最大化奖赏的动作

探索 (exploration)：尝试新的动作，该动作可能会带来更大的奖赏