# 1 - 绪论

## 为什么要学习强化学习?

- 强化学习的本质就是试错学习（Trail and error learning）
- 试错学习一开始适合行为心理学等工序哦联系在一起的，主要部分包括了：
  - 尝试：为了实现目标所做出的努力
  - 错误：在尝试的过程中失败了，可能是由环境的不确定性导致的，也可能是自身的行为导致的
  - 结果：每一次尝试后的后果，不论正面或负面
  - 学习：在失败的尝试后，自身积累经验以便下一次可以避免失败
- 试错学习在生活中的例子屡见不鲜，例如经典的条件反射实验和观察学习。
- 在强化学习中，我们称好的结果为奖励，而坏的结果为惩罚。然后我们将通过一个决策过程实现目标，这个目标通常是奖励最大化，这个过程就是序列决策（Sequential decision making），是目前强化学习主要实现方法。

## 强化学习的应用

### 游戏

- 《星际争霸》的AlphaStar
- 通用游戏AI的AlphaZero
- Dota2的OpenAI Five

### 机器人

- NICO学习抓取

### 金融

- 股票交易
- 期货交易
- 外汇交易

### 其他

- 自动驾驶
- 推荐系统
- 交通派单
- 广告投放
- ChatGPT

## 强化学习方向概述

### 多智能体强化学习

- 多智能体强化学习（Multi-agent reinforment learning, MARL）是多个智能体的环境下的强化学习。

- 通常拥有对抗性实验，例如合作，竞争等等

# 从数据中学习

- 从数据学习或从演示中学习（Learn from demostration）的分类：
  - 以模仿学习（Imitation learning, IL）为代表的从专家数据中学习策略
  - 以逆强化学习（Inverse reinforment learning, IRL）为代表的从人类数据中学习奖励函数
  - 从人类反馈中学习（Reinforcement learning from human feedback, RLHF）为代表的从人类标注的数据中学习奖励模型来进行微调（Fine-tune）

## 模仿学习

- 指在奖励函数难以名确定以或者策略本身很难学习的情况下，选择通过模仿人类的行为来学习到一个较好的策略。
- 例子：行为克隆（Behavioral Cloning, BC），是每一个状态为训练样本，通过监督学习来学习策略。但是这种方法往往鲁棒性较低，无法应对未见过的状态。

## 逆强化学习

- 指通过观察人类行为来学习奖励函数，然后再通过奖励函数来学习一个策略。
- 由于需要专家数据，逆强化学习会受到噪声影响

## 挑战

- 探索策略
  - 如何在探索和利用之间做出权衡
  - 例子：
    - $\epsilon$ - greedy
    - 置信上界（Upper Confidence topologies, UCB）
    - NEAT（Neuro evolution of augmenting topologies）
    - PBT（Population based training）
- 实时环境
  - 智能体往往需要在实时或者在线环境中进行决策，这种情况下训练不仅会降低效率，而且还会带来安全隐患
  - 解决方法：
    - 离线环境进行训练
      - 可能出现分布漂移现象，即两个环境的状态分布不同，这就导致了训练好的模型在在线环境中可能会出现意外
    - 训练世界模型
      - 拥有两个部分：一个世界模型和一个控制器
      - 世界模型预测下一个状态；控制器根据当前状态决策动作
      - 世界模型的预测误差可能导致控制器的决策出错
- 多任务强化学习
  - 智能体往往需要同时解决多个任务
  - 如何在多个任务之间做出权衡
  - 解决方案：
    - 联合训练（Joint training）

    - 将多个任务的奖励进行加权求和，然后通过强化学习来学习一个策略
- 分层强化学习（Hierarchical reinforcement learning）
    - 将多个任务分为两个层次，一个是高层策略，另一个是低层策略
    - 高层策略的作用是决策当前的任务，而低层策略的作用是决策当前任务的动作
    - 高层策略的决策可能会导致低层策略的决策出错