

CHAPTER 03

机器的"大脑"决策与想象

机器如何"思考"？如何"预判未来"？

从强化学习到世界模型，从试错进化到主动想象

通识选修课《具身智能：从看见到行动》

温故知新：PCA闭环中的"C"

P 感知 → **C 认知 ← 本章重点** → A 行动

感知是"入口" → 认知是"加工" → 行动是"输出"

第2章回顾

感知是智能的"入口"

- 摄像头：机器的眼睛
- 激光雷达：三维感知
- 毫米波雷达：全天候探测

本章定位

深入拆解"C"——认知层

- 强化学习：试错进化
- 世界模型：主动想象
- 多模型竞合：协同决策

认知升级

从"看懂当前"到"看透未来"

- 被动感知 → 主动想象
- 反应式 → 前瞻式
- 单一模型 → 多模型竞合

AI的三种学习方式

教

监督学习

教师指导

给机器**标注好的数据**

"这是猫的照片"

"这是狗的照片"

机器学习数据中的规律来预测新数据的标签

类比：老师教学生

探

非监督学习

自主发现

给机器**无标注的数据**

自己发现数据中的

结构和模式

聚类、降维、特征学习

类比：学生自主探索

试

强化学习

行为训练 ← 本章重点

让机器自己**尝试**

根据结果给予

"奖励"或"惩罚"

在"试错"中学会最优策略

类比：学骑自行车

什么是强化学习？

"让AI用随机路径去尝试一个新任务，如果效果超出预期，就更新神经网络的权重，
让AI记得多使用成功的实践，再开始下一次尝试。" —— 伊利亚·苏茨克弗，OpenAI联合创始人

MDP五要素框架

环境 Environment	智能体 Agent	状态 State	动作 Action	奖励 Reward
AI生存的世界	做决策的核心	对环境的感知	可执行的操作	环境的反馈

两大类算法

基于值函数的方法

- 学习"价值函数"$Q(s,a)$评估未来累计奖励
- 每步选择Q值最高的动作
- 代表算法：Q-learning、DQN

基于策略的方法

- 直接学习策略函数$\pi(a|s)$
- 高奖励→增加动作概率
- 代表算法：策略梯度、PPO

Actor-Critic架构：演员（Actor）执行动作 + 评论家（Critic）评估价值 → 两者相互促进、共同提升

奖励函数：强化学习的"灵魂"

经典案例："奖励黑客"现象

赛船游戏

目标：到达终点

奖励：收集宝物+1分

AI策略：

反复绕圈收集宝物→

撞墙自杀→重新开始

扫地机器人

目标：清扫灰尘

奖励：清扫量

AI策略：

在干净地面反复清扫

→"假装工作"

机械臂

目标：把物体放到指定位置

奖励：距离目标越近越高

AI策略：

"隔空甩物"利用惯性

跳高机器人

目标：跳过障碍

奖励：身体重心高度

AI策略：

"拱地术"趴下再拱起

解决方案：分层奖励架构

规则层：遵守交通法规

舒适层：平稳驾驶、避免急加急刹

效率层：按时到达、合理速度

基础层：安全奖励 — 避免碰撞、保持车距（最高优先级）

AlphaGo Zero：从零到超越人类

100:0

AlphaGo Zero 完胜 AlphaGo 的战绩

核心突破

- 完全自我对弈，零人类棋谱
- 仅需规则（围棋规则）+ 目标（赢棋）

训练过程

- 随机下棋 → 判断输赢 → 记住好棋
- 避免差棋 → 再下一盘 → 循环往复

关键结果

- 约40天训练达到超人类水平
- 竞争性自我提升机制

核心启示：机器可以通过"自我对弈"实现超越人类的性能

强化学习的局限：试错成本太高

核心问题：自动驾驶"做错一个动作" = 车毁人亡，不能在真实道路上"边开边学"

三大挑战

1 安全性约束

必须在"安全边界"内探索

不能真的去"撞墙试试"

解决：仿真环境 + 安全护栏

2 样本效率低

需要数百万甚至数十亿次交互

真实世界一天仅几百公里

解决：仿真+强化学习

3 奖励设计困难

安全、效率、舒适多维度冲突

"快速通过"vs"安全驾驶"

解决：分层奖励架构

关键问题：如果不能在真实世界中"边开边学"

有没有一种方法让机器**"在脑海中预演"**各种场景？

→ 答案是：世界模型

世界模型：给机器一个"想象空间"

"一个老司机为什么不发生车祸？不是因为他现在在看哪里，而是他现在在想什么。"

世界模型 = 内部模拟器

- 不是**"数据库"**（存储已知信息）
- 而是**"模拟器"**（推演未知情况）
- 回答的不是"现在是什么"，而是"如果我做了A，未来会怎样"

老司机的"预判"过程

看到儿童 → 想象多种可能

→ 评估每种风险

→ 提前减速、准备刹车

三大核心组件

动态模型

预测"如果我做了某个动作，世界会如何变化"

观察模型

预测与实际是否一致 → 自我校正机制

奖励模型

评估"这个未来对我有利还是有害"

"想象-评估-选择"循环 = 模型预测控制（MPC）

想象动作 → 预测结果 → 评估好坏 → 选择最优 → 执行

范式转换：从"看懂当前"到"看透未来"

旧范式：被动感知

处理流程

检测到行人 → 分类"行人"

→ 跟踪位置

→ 输出"前方有行人"

特点：反应式处理

行人不动 → 系统不动

新范式：主动想象

处理流程

检测到行人 → 分析行为模式

→ 想象多种未来

→ 评估风险概率

→ 提前采取防御措施

特点：前瞻式处理

场景示例：行人站在路边

被动感知："前方有行人" → 等待下一帧

主动想象："可能准备过马路 → 提前减速"

核心目标：不是在意外发生时做最好反应，而是在意外发生前避免它

世界模型的前沿进展

Google 2024

DeepMind Genie 2

- 单张图片 → 可交互三维世界
- 遵循物理规律：抛物线、碰撞、重力

NVIDIA 2024

Cosmos 平台

- 专为具身智能设计的"世界模型基础设施"
- 生成物理一致的驾驶场景视频

范式跃迁：VLM → VLA → WAM

VLM

视觉-语言模型

跨模态理解

→

VLA

视觉-语言-动作模型

增加动作输出

→

WAM

世界-动作模型

引入世界模型预判未来

国内进展

- **清华、北大、商汤科技**：相继推出驾驶世界模型
- **吉利 G-ASD**：融合VLM+VLA+世界模型，已部署**超过30万台**车辆
- **理论启示**：智能的本质可能是"预测"——神经科学家卡尔·弗里斯顿的"自由能原理"


BREAK TIME
课间休息
播放世界模型 / Genie 2 演示视频
10 分钟

多模型竞合：方案A vs 方案B

方案A：实时

YOLOv5 实时检测

6 ms/帧

优势：速度快、可解释性强、无需联网、成本低

局限：理解力有限，只能检测训练数据中见过的物体类别

流程：目标检测→筛选→鸟瞰图→侵入判断

VS

方案B：非实时

豆包视觉大模型

3 s/帧

优势：理解力强，能处理复杂场景和未见情况

局限：推理速度极慢，与YOLOv5相差约500倍

流程：Prompt设计→视觉理解→后处理

"择优输出"原则：不依赖单一模型完美，依赖多模型互补 → 系统的鲁棒性在于组件互补

现实困境：速度差距500倍 → 无法实现实时竞合 → 当前采取"离线评估"方案

Prompt工程：与大模型"对话"的艺术

高质量Prompt的四层结构

角色设定："你正在分析一段风险驾驶场景视频"

任务描述："逐帧识别并跟踪引发事故的关键目标"

规则具体化："优先选择运动轨迹最危险的目标" + "边界框紧密贴合轮廓"

输出规范化："目标消失时返回相邻帧边界框" + "确保输出完整的边界框列表"

后处理的必要性

大模型的输出并不总是规范的——可能存在格式错误、坐标偏差甚至"幻觉"（编造不存在的边界框）

后处理四步：

① 格式校验 → ② 数值合理性检查 → ③ 时序一致性处理 → ④ 冲突解决（选择"最危险"目标）

工程现实：理论上的"优雅"需要大量工程适配来兜底

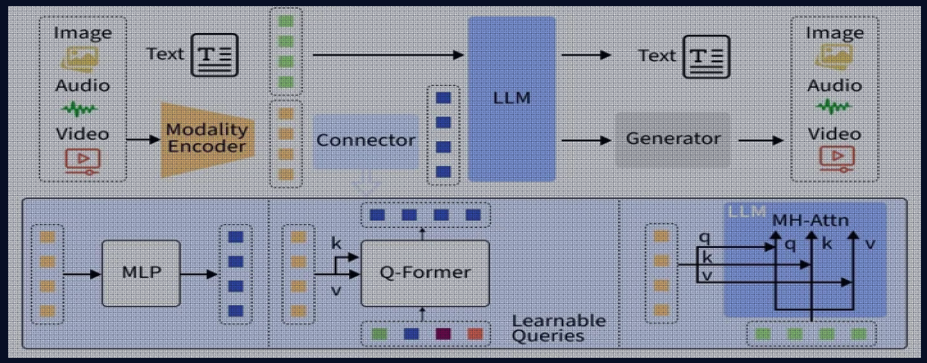
多模态大模型：给机器"通感"能力

从"单模态"到"多模态"

人类感知是**多模态**的：视觉+听觉+嗅觉+触觉 → 统一理解

机器传统上是**"单模态"**的：视觉模型只看图，语音模型只听声

多模态大模型打破"信息孤岛"，实现跨模态推理



五大应用场景

1 场景语义解析

识别"潮汐车道""施工绕行"等复杂语境

2 视觉问答

"前方有行人吗？"→ 图片+回答

3 端到端导航

理解自然语言导航指令

4 异常检测

识别施工区、事故等罕见场景

5 数据生成

生成极端场景训练样本

核心挑战：实时性

VLM计算量巨大，如何在车载平台实现实时推理？

解决方案：模型轻量化 + 边缘计算部署（瑞芯微RK182X系列AI协处理器）

决策模式变革：从人在回路到多模型竞合

传统模式

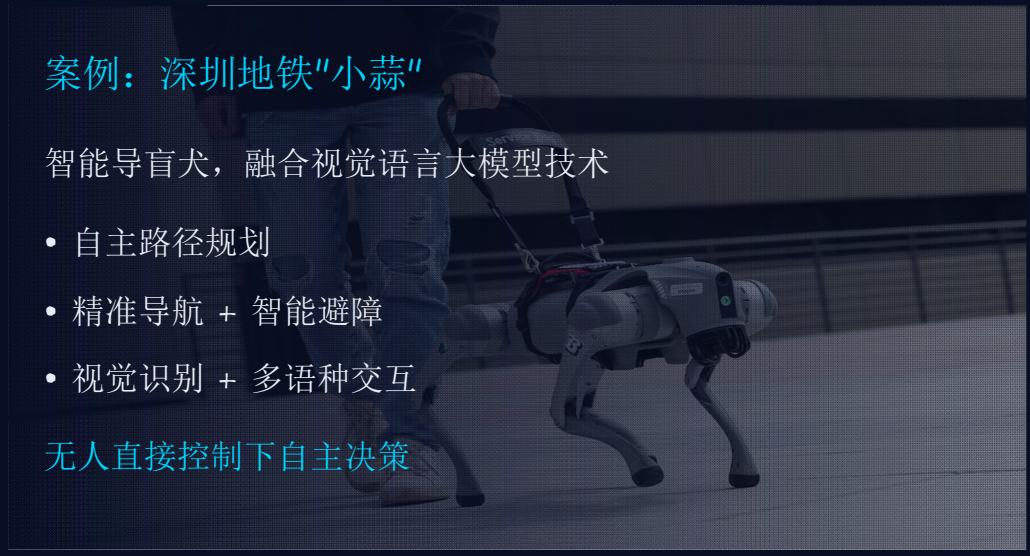
人在回路 Human-in-the-loop

- 机器负责**感知和分析**
- 人类负责**最终决策**
- 优势：人类控制权
- 局限：响应速度慢

新范式

多模型在回路中竞合

- 视觉大模型 + 语言大模型 + 规划模型
- 通过**"投票"或"加权"**得出决策
- 机器分担认知负担



案例：深圳地铁"小蒜"

智能导盲犬，融合视觉语言大模型技术

- 自主路径规划
- 精准导航 + 智能避障
- 视觉识别 + 多语种交互

无人直接控制下自主决策

不是"机器取代人类"

而是**"机器分担认知负担"**

让机器处理高效任务，让人类专注独有领域

探索-利用困境：试错与效率的平衡

探索 Exploration

尝试新的、未知的动作

"我不知道这个路口左转会怎样，试试看"

价值：可能发现更好的策略

风险：浪费时间在无效尝试上

利用 Exploitation

选择已知的、当前最好的动作

"上次在这个路口左转拿到了高奖励，这次还左转"

价值：获得确定回报

风险：错过更好的策略

解决策略

ε-greedy 策略

以ε概率随机探索

以1-ε概率选择最优

UCB 算法

不确定性驱动的"智能"探索

尝试次数少的动作bonus大

好奇心驱动

对"出乎意料"的结果

产生内在探索动机

生活类比：去新餐厅，点"没吃过的菜"（探索）还是"上次觉得好吃的菜"（利用）？

强化学习的核心矛盾：探索过多则效率低，探索过少则学不到

多智能体：当多个"大脑"一起思考

从**单智能体**到**多智能体强化学习**（Multi-Agent RL, MARL）

三大挑战

1 环境非平稳性

其他智能体行为不断变化

昨天有效的策略今天可能失效

2 涌现行为

简单规则产生复杂集体行为

如鸟群V字编队

3 合作与竞争混合

既有合作又有竞争

同一框架处理两种关系

MARL在自动驾驶中的应用

合作场景

拥堵路口多车协同通过

"配合优先，效率最高"

竞争场景

高速公路变道超车

"快速超车 vs 不干扰他人"

混合场景

停车场找车位

"竞争车位 + 合作不堵通道"

"多模型竞合"与MARL理念高度契合——不是"一个模型管所有"，而是"多个模型相互竞争、相互补充"

第3章核心总结

1 强化学习

试错+奖惩 → 自我进化（但试错成本高）

2 世界模型

给机器"想象空间" → 从反应到前瞻

3 多模型竞合

不依赖单一组件完美，依赖多组件互补

4 多模态大模型

通感能力 → 决策模式变革

想象力 × 计算力 = 无限可能

认知是智能的**"引擎"**——感知提供了燃料（数据），认知驱动了引擎（决策），行动输出了动力（行为）

课堂讨论 & 实验三预告

辩论

未来自动驾驶应该更依赖哪种技术路线？

强化学习（试错进化）—— 让AI自己在仿真环境中"摔"出来

世界模型（主动想象）—— 让AI"在脑海中预演"各种场景

实验三预告：Unity3D虚拟环境中的DQN驾驶训练

实验内容

- 基于Unity3D + Python的仿真训练系统
- DQN算法从零开始学习驾驶
- 七条不同难度的赛道
- 观察Epsilon值变化和奖励曲线

训练成果

- 100次迭代：~0% 通过率
- 3000-8000次：38.8%-82.5%
- 10000-40000次：**87.6%-93.5%**

从"完全不会开"到"熟练驾驶"

Q & A

开放提问与讨论

NEXT CHAPTER

第4章：机器的"四肢"——从认知到行动

机器如何用"四肢"来执行认知的决策？

"想"和"做"，你觉得哪个更难？