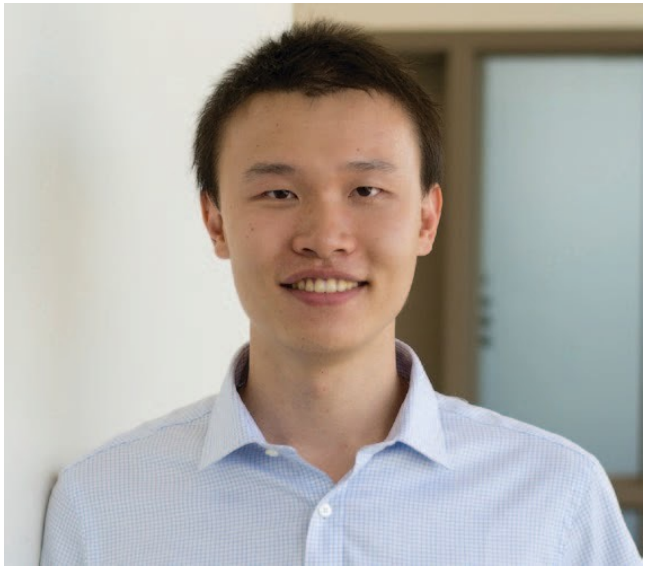
欢迎大家学习
《人工智能导论》

艾清遥

清华大学计算机系

AI

Qingyao AI 艾清遥



- 单位： 清华大学计算机系
- 智能技术与系统
- 国家重点实验室
- 办公室：FIT楼1-506
- 开放交流时间：每周四16:00-17:00
- E-mail：
- aiqy@tsinghua.edu.cn
- 网络学堂：
  learn.tsinghua.edu.cn
- 主页：
  http://www.thuir.cn/group/~aiqy/

清华大学信息检索实验室THUIR

信息检索是人工智能的一个分支，研究信息的结构、分析、组织、存储、搜寻与检索技术

- **博士生导师**4人
- 每年招收博士生3-5人
- 在读博士22人，硕士6人
- 国际计算机学科排行榜
  CSRankings 网络与信息检索方向
  全球Top-1 (2014-2024)

THUIR公众号

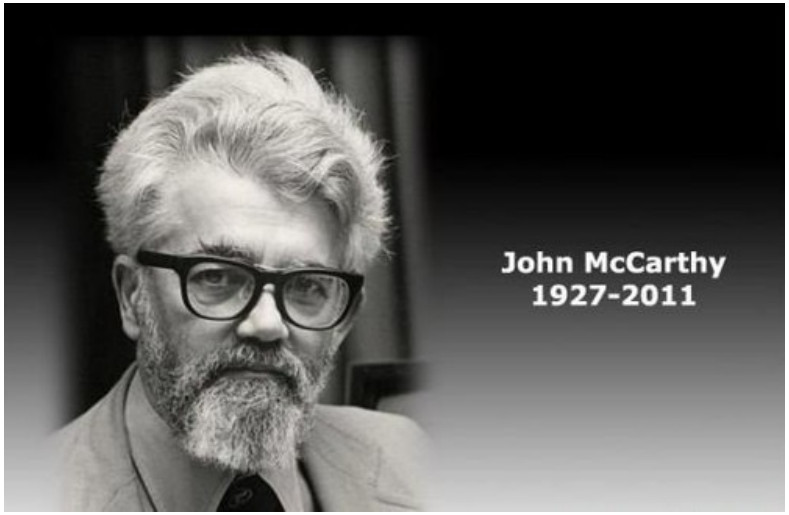
THUIR主页

www.thuir.cn

绪论：人工智能进展

人工智能的“诞生”

- 1956年达特茅斯夏季讨论会上，首次提出人工智能（会议召集人麦卡锡（John McCarthy））
  - 远古时期人类对人造智能的幻想
  - 图灵：1950年发表论文《计算机与智能》
  - 电子计算机的出现……



人工智能发展的五个阶段（？）

- 初期阶段
- 知识时代
- 特征时代
- 数据时代
- 大模型时代

初期阶段（逻辑/符号时代）

- 定理证明
- 通用问题求解
- 逻辑推理
- 机器翻译
- 博弈
- 游戏
- 。 。 。

赫伯特·西蒙

Herbert A. Simon

艾伦•纽厄尔

Allen Newell

- 一个笑话:（英俄翻译）

The spirit is willing but the flesh is weak.
（心有余而力不足）

The vodka is strong but meat is rotten.
（伏特加酒虽然很浓，但肉是腐烂的）

知识时代

- 专家系统
- 知识工程
- 知识表示
- 不确定性推理
- 人为知识表示
- 。 。 。

费根鲍姆

Edward Albert Feigenbaum

- 知识获取的瓶颈问题
- 比如：如何骑自行车？

特征时代

- 统计学习方法
- 优化技术
- 特征映射（浅层）
- 人为特征定义
- 。 。 。

莱斯利
Leslie Valiant

朱迪亚
Judea Pearl

- 特征定义的困难
- 比如：语音识别

数据时代

- 深度学习
- 表示学习
- 自动特征抽取
- 不同层次的抽象特征
- 特征映射（深层）
- 。 。 。

杨立昆 Yann LeCun　辛顿 Geoffrey Hinton　本吉奥 Yoshua Bengio

- 数据时代的困难？

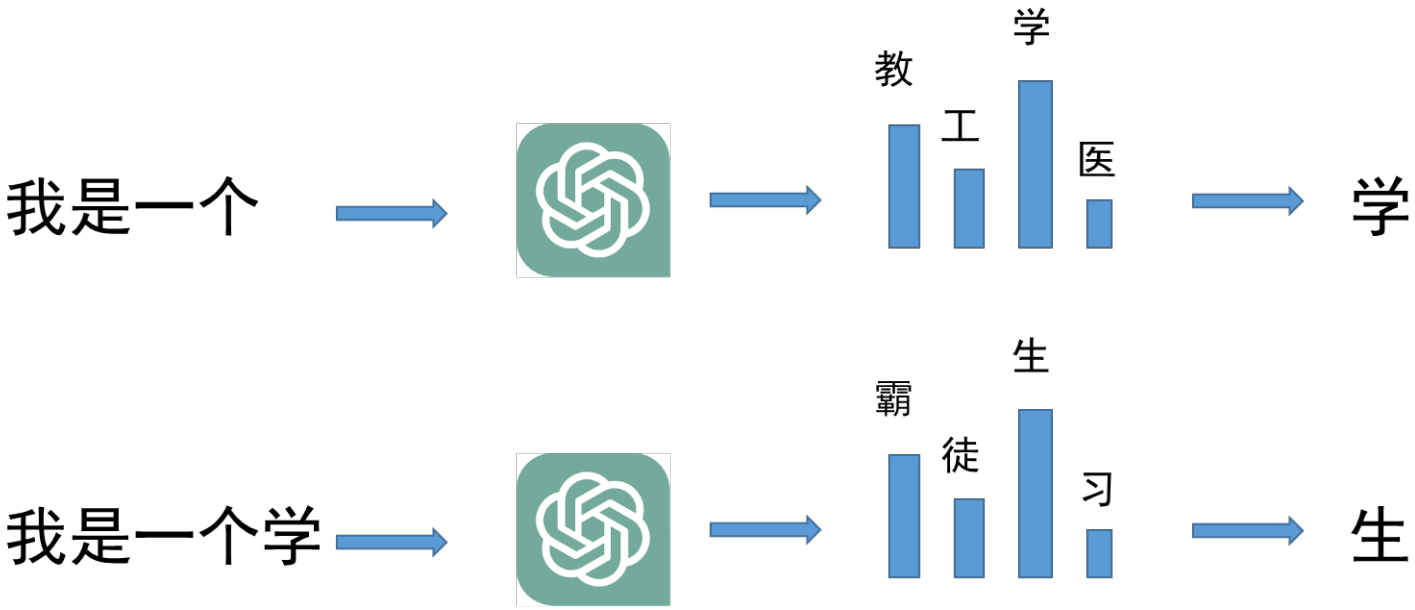
进入大模型时代?

- ChatGPT
  - 1750亿个参数
  - 45TB训练数据
  - 28.5万个CPU
  - 1万个高端GPU
  - 训练成本1200万美元
  - 语言理解能力
  - 语言生成能力
  - 多轮对话管理能力
  - 正确性还有待提高

什么是ChatGPT?

- GPT： **G**enerative **P**re-Trained **T**ransformer
- 生成式预训练变换模型

生成式模型：文字接龙



生成式模型：回答问题

白日依山进的下一句是 →  → 黄河入海流

预训练模型：自监督学习

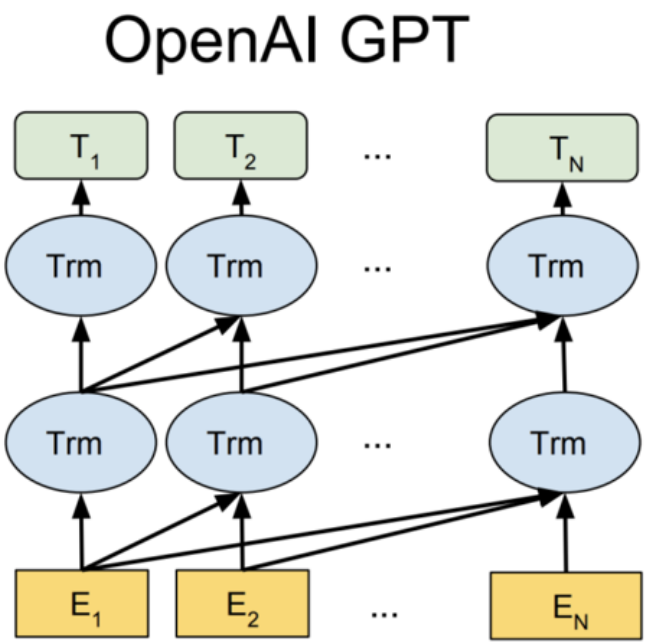
- n元语言模型

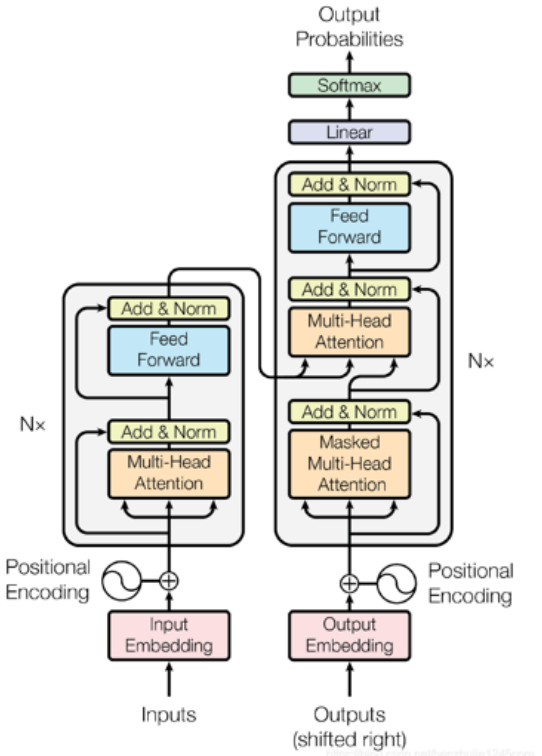
$$P(w_n|w_1 w_2 \dots w_{n-1})$$

- 自监督学习

$$w_1 w_2 \dots \underbrace{w_{i-(n-1)} \dots w_{i-2} w_{i-1}}_{\text{前n}-1\text{个字}} w_i \dots w_m$$

变换模型：Transformer

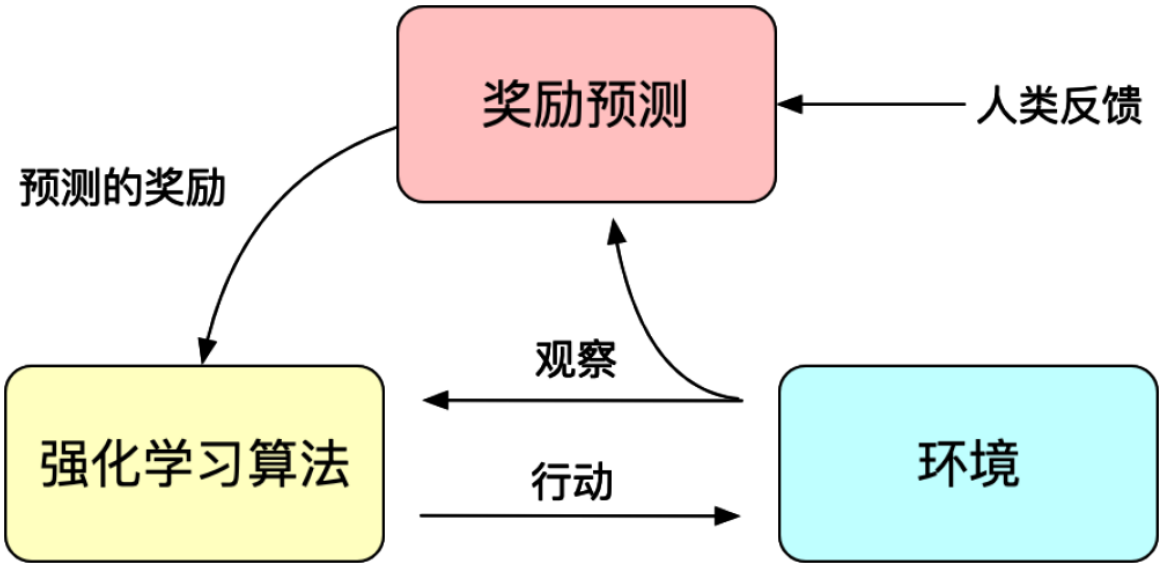




模型	发布时间	层数	头数	词向量长度	参数量	预训练数据量
GPT-1	2018 年 6 月	12	12	768	1.17 亿	约 5GB
GPT-2	2019 年 2 月	48	-	1600	15 亿	40GB
GPT-3	2020 年 5 月	96	96	12888	1,750 亿	45TB

ChatGPT

- 以GPT3/3.5为基础
- 采用人工反馈的强化学习（RLHB）



共同特点：如何定义问题

人工智能 = “定义” + 算法

定义：描述问题

算法：把智能问题转化为计算问题

举例：什么是猫?

- 百科上“猫”的定义：

"头圆、颜面部短，前肢五指，后肢四趾，趾端具锐利而弯曲的爪，爪能伸缩。夜行性。以伏击的方式猎捕其他动物，大多能攀缘上树。"


挪威森林猫
孟买猫
英国短毛猫
缅因猫
异国短毛猫

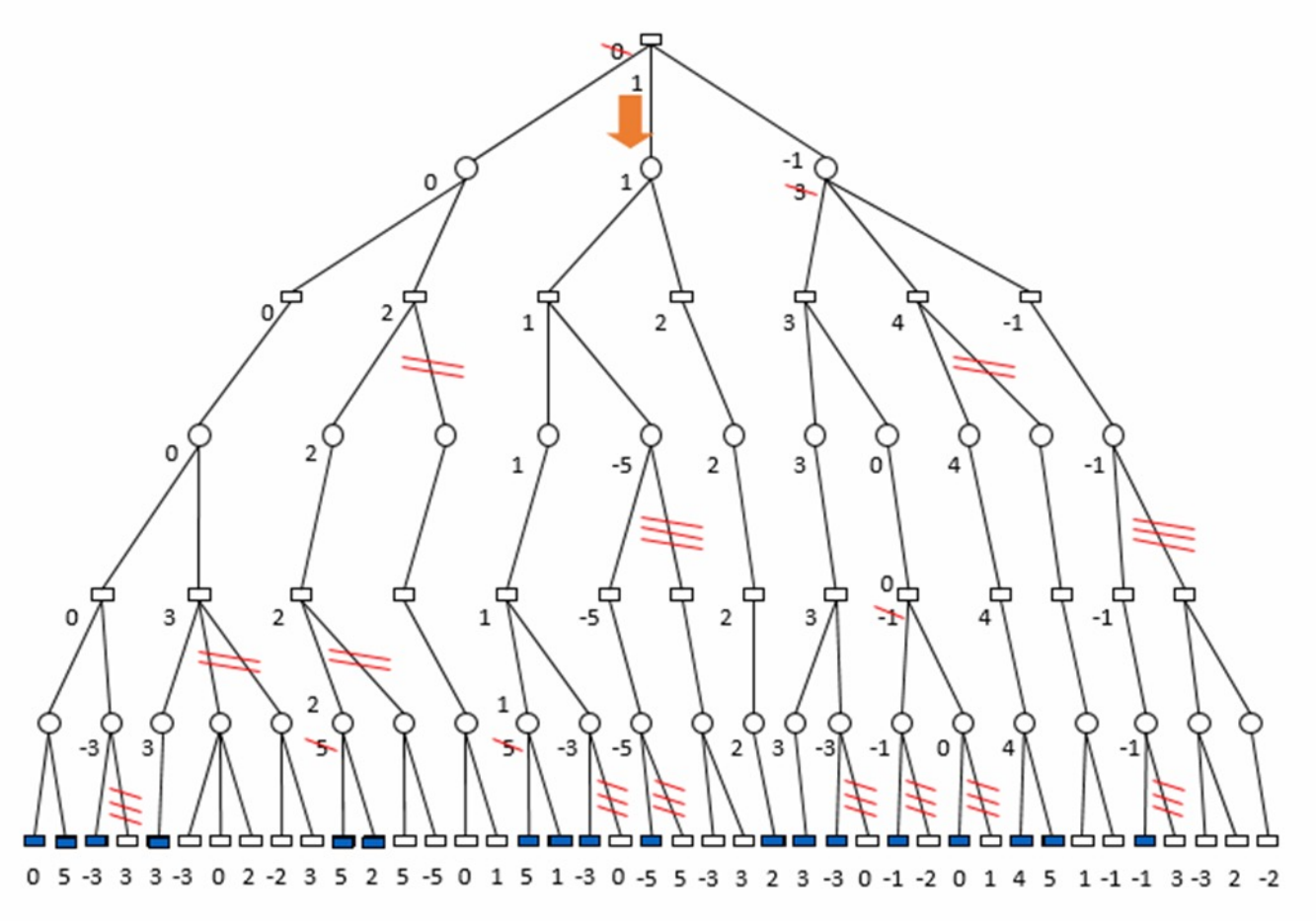

土耳其安哥拉猫
褴褛猫
东奇尼猫
马恩岛猫
沙特尔猫


伯曼猫
斯芬克斯猫
缅甸猫
新加坡猫
土耳其梵猫


波米拉猫
塞尔凯克卷毛猫
拉邦猫
东方猫
美国卷毛猫

深蓝 （1997）

深蓝的算法框架：α-β剪枝



深蓝何以取胜?

- 专家的知识 + 搜索（α-β剪枝）
- 依赖于国际象棋大师的参与

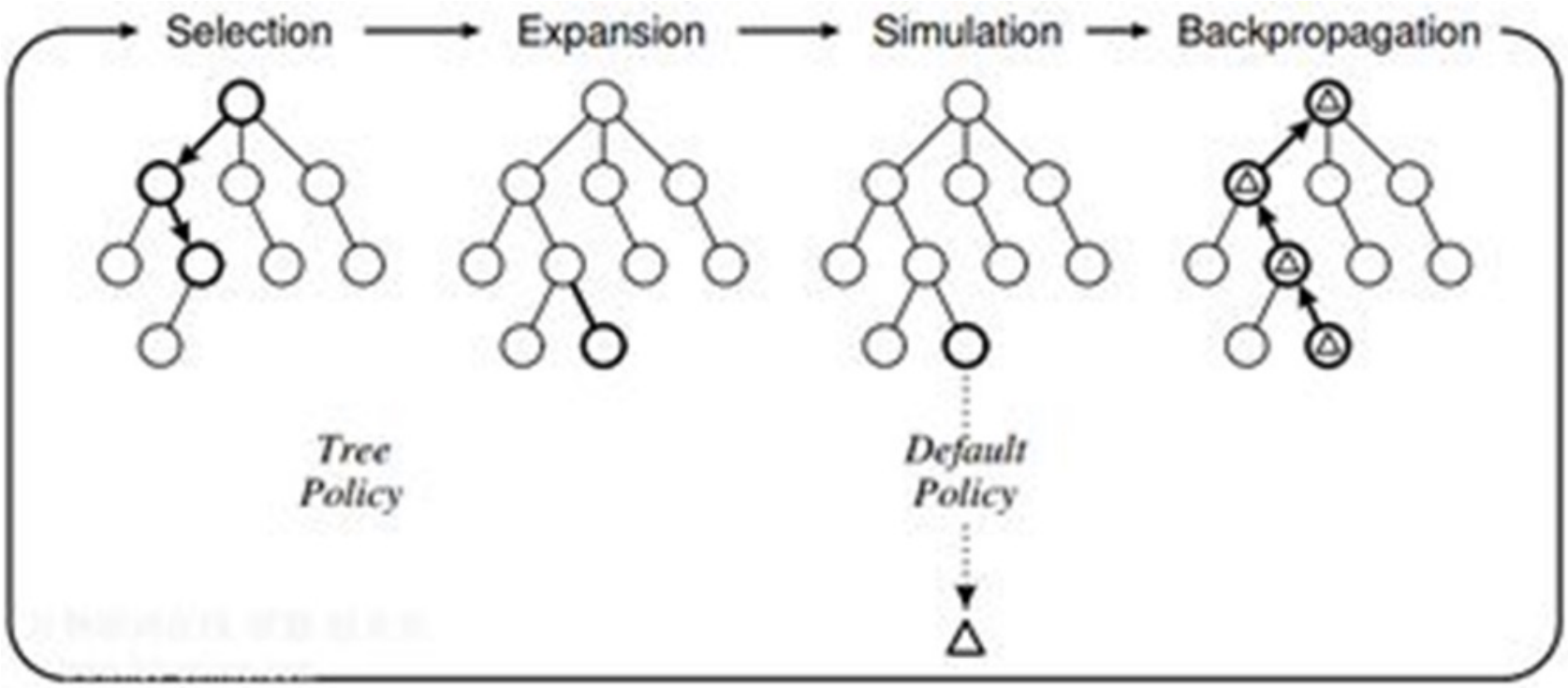
中国象棋浪潮杯比赛

- 2006年为纪念人工智能诞生50周年，中国人工智能学会主办了浪潮杯中国象棋人机大战，先期举行的机器博弈锦标赛获得前5名的中国象棋软件，分别与汪洋、柳大华、卜凤波、张强、徐天红5位中国象棋大师对弈，**人机分别先行共战两轮10局比赛**，双方互有胜负，最终**机器以11:9的总成绩战胜人类**大师队。

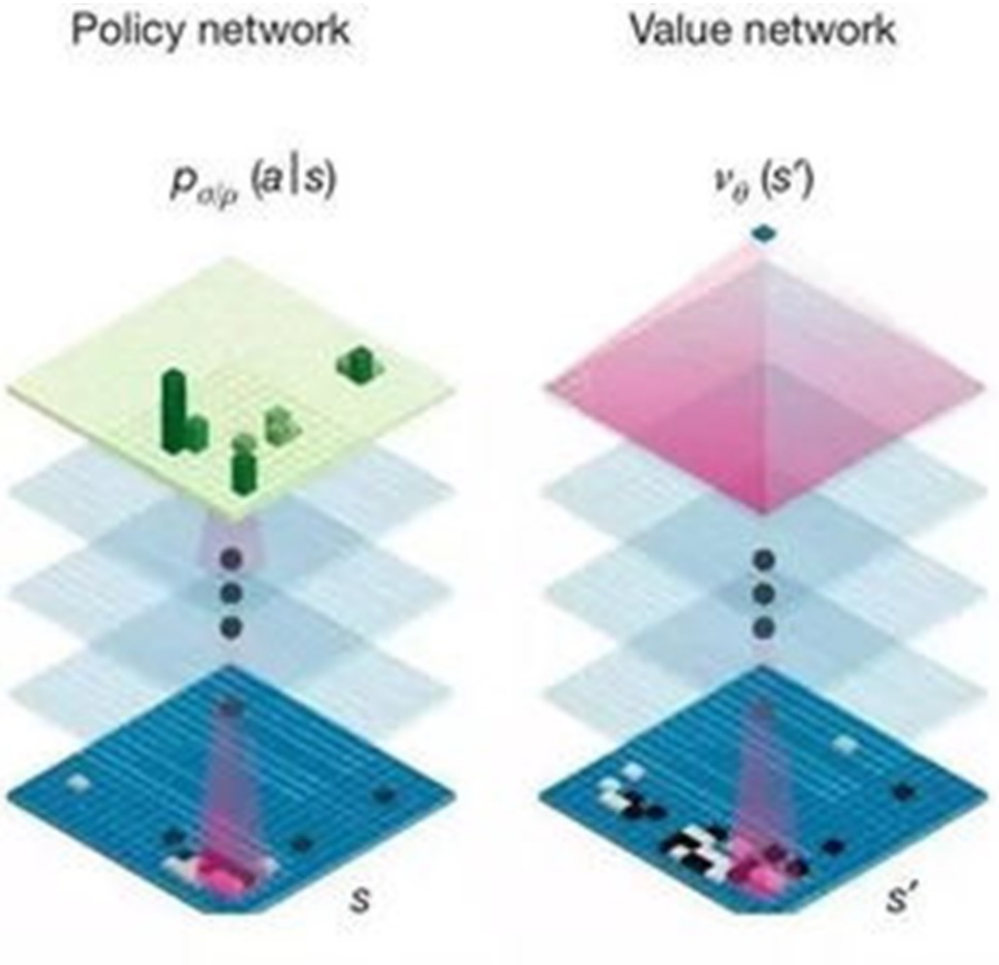
AlphaGo （2016）

AlphaGo

AlphaGo的算法框架：MCTS



策略网络与估值网络



AlphaGo何以取胜?

- 人类数据 + （蒙特卡洛树搜索+深度学习）
- 不需要围棋大师的直接参与

AlphaGo Zero：从零学习?

AlphaGo Zero：从零学习?

- 完全摆脱人类知识，从零学习
  - 不依靠人类棋谱
  - 不再使用人工特征作为输入
  - 。 。 。
- 3天，战胜AlphaGo Lee
- 40天，战胜AlphaGo Master

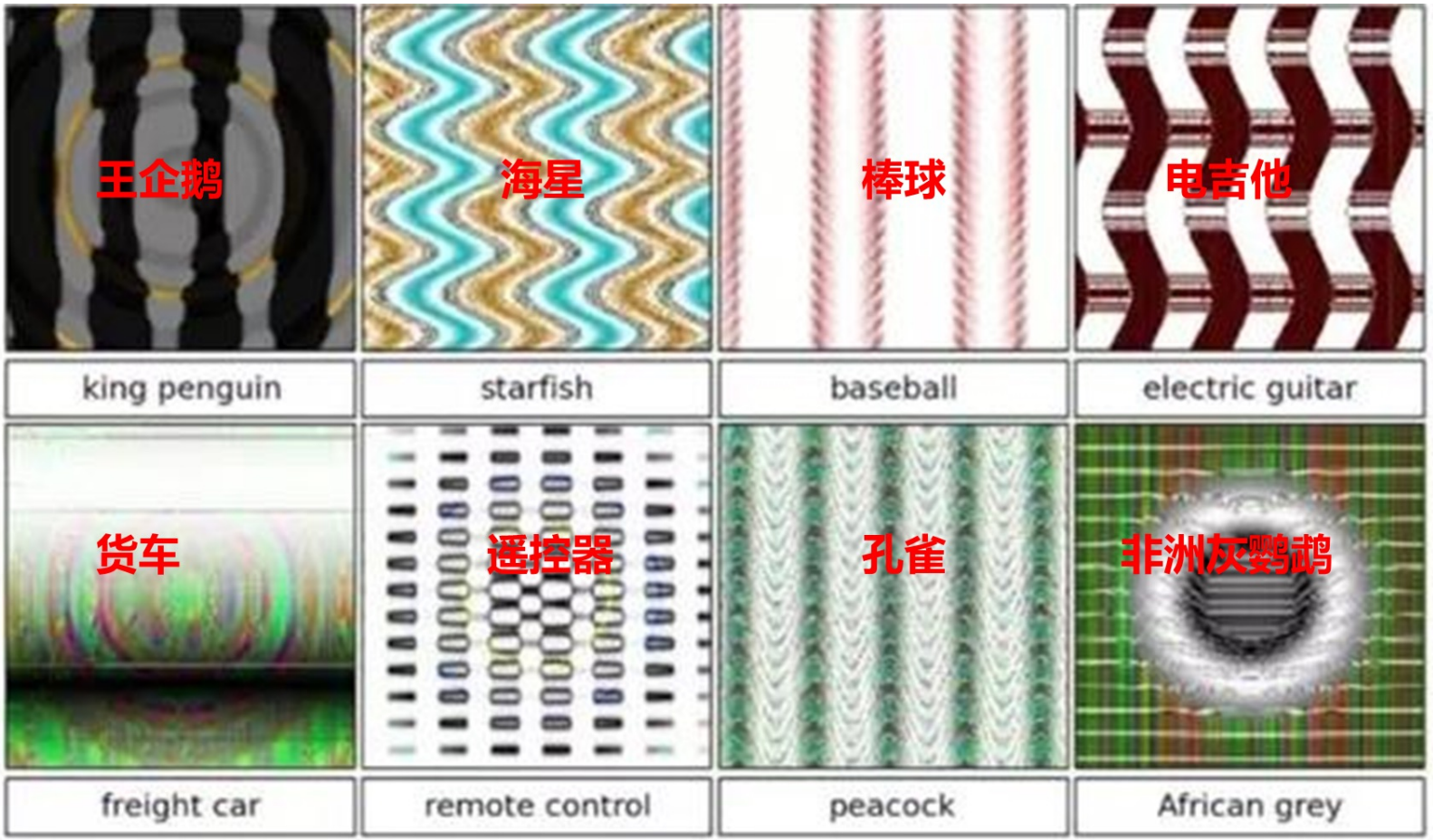
AlphaGo Zero：从零学习?

- 棋类问题的特殊性
  - 可以自己判断胜负
  - 本质上还是依靠数据
- 不具有一般推广性

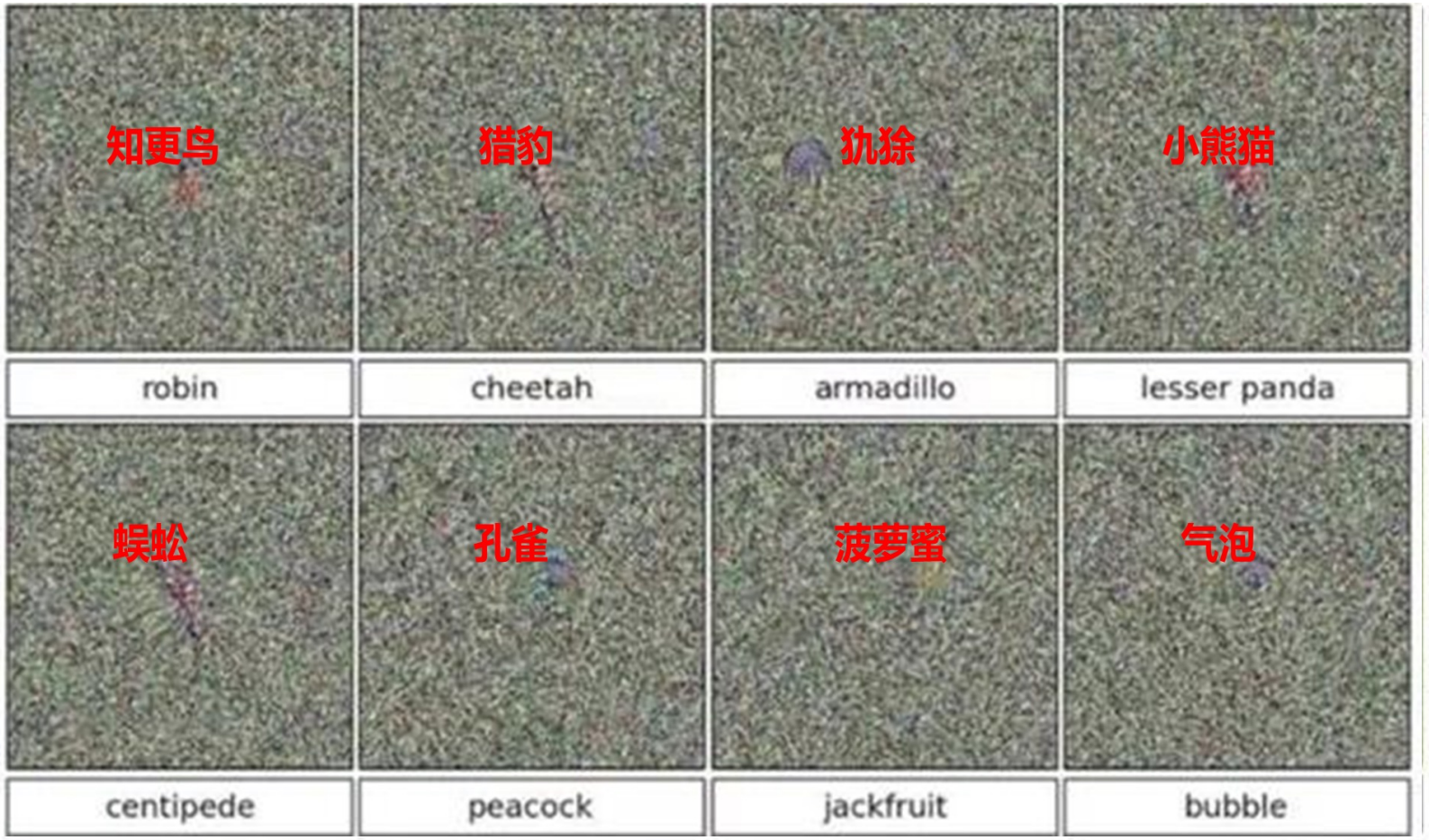
深度学习存在的问题

- 大数据 vs 小样本
- 黑箱 vs 可解释
- 一次性学习 vs 增量学习
- 固执己见 vs 知错能改
- 猜测 vs 理解

深度学习存在的问题


王企鹅
海星
棒球
电吉他
king penguin
starfish
baseball
electric guitar
货车
遥控器
孔雀
非洲灰鹦鹉
freight car
remote control
peacock
African grey

深度学习存在的问题


知更鸟
猎豹
犰狳
小熊猫
robin
cheetah
armadillo
lesser panda
蜈蚣
孔雀
菠萝蜜
气泡
centipede
peacock
jackfruit
bubble

深度学习存在的问题

加上0.7%的噪声后，99.3%的信任度识别成长臂猿

图灵测试与中文屋子问题

- 图灵
  - 1912年出生于英国伦敦，1954年去世
  - 1936年发表论文“论可计算数及其在判定问题中的应用”，提出图灵机理论
  - 1966年为纪念图灵的杰出贡献，ACM设立图灵奖

图灵测试

- 如何知道一个系统是否具有智能呢？
- 1950年，图灵发表论文《计算机与智能》，提出了著名的“图灵测试”（模仿游戏）。

图 1　图灵测试示意图

希尔勒的中文屋子

- 罗杰•施安克的故事理解程序

故事理解程序举例

- “一个人进入餐馆并订了一份汉堡包。当汉堡包端来时发现被烘脆了，此人暴怒地离开餐馆，没有付帐或留下小费。”
- “一个人进入餐馆并订了一份汉堡包。当汉堡包端来后他非常喜欢它，而且在离开餐馆付帐之前，给了女服务员很多小费。”
- 作为对“理解”故事的检验，可以向计算机询问，在每一种情况下，此人是否吃了汉堡包。

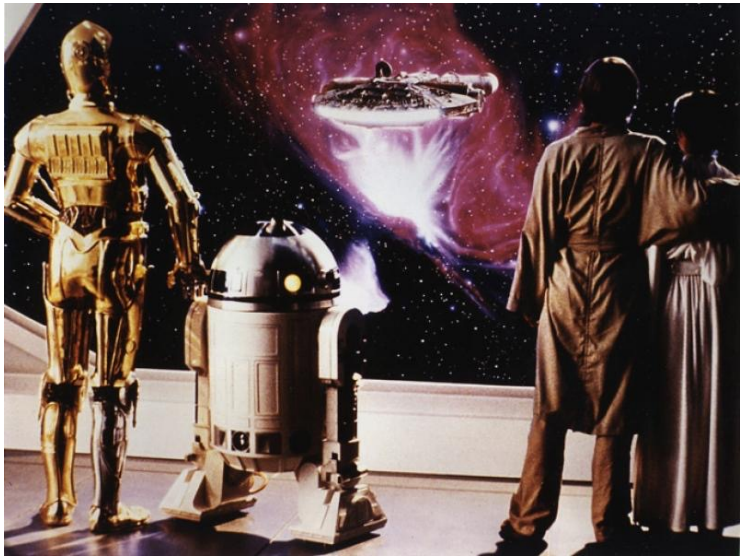
返回

希尔勒的中文屋子

- 罗杰•施安克的故事理解程序
- 机器是否真的理解了呢？
- 希尔勒的中文屋子
- 问题：通过了图灵测试就具有了智能吗？
- 思考题：如何理解希尔勒的中文屋子？

AI的本质问题

研究如何制造出人造的智能机器或系统，来模拟人类智能活动的能力，以延伸人们智能的科学。



人工智能五要素

算据

算景

算法

算力

AI系统

算者

本课主要学习的内容

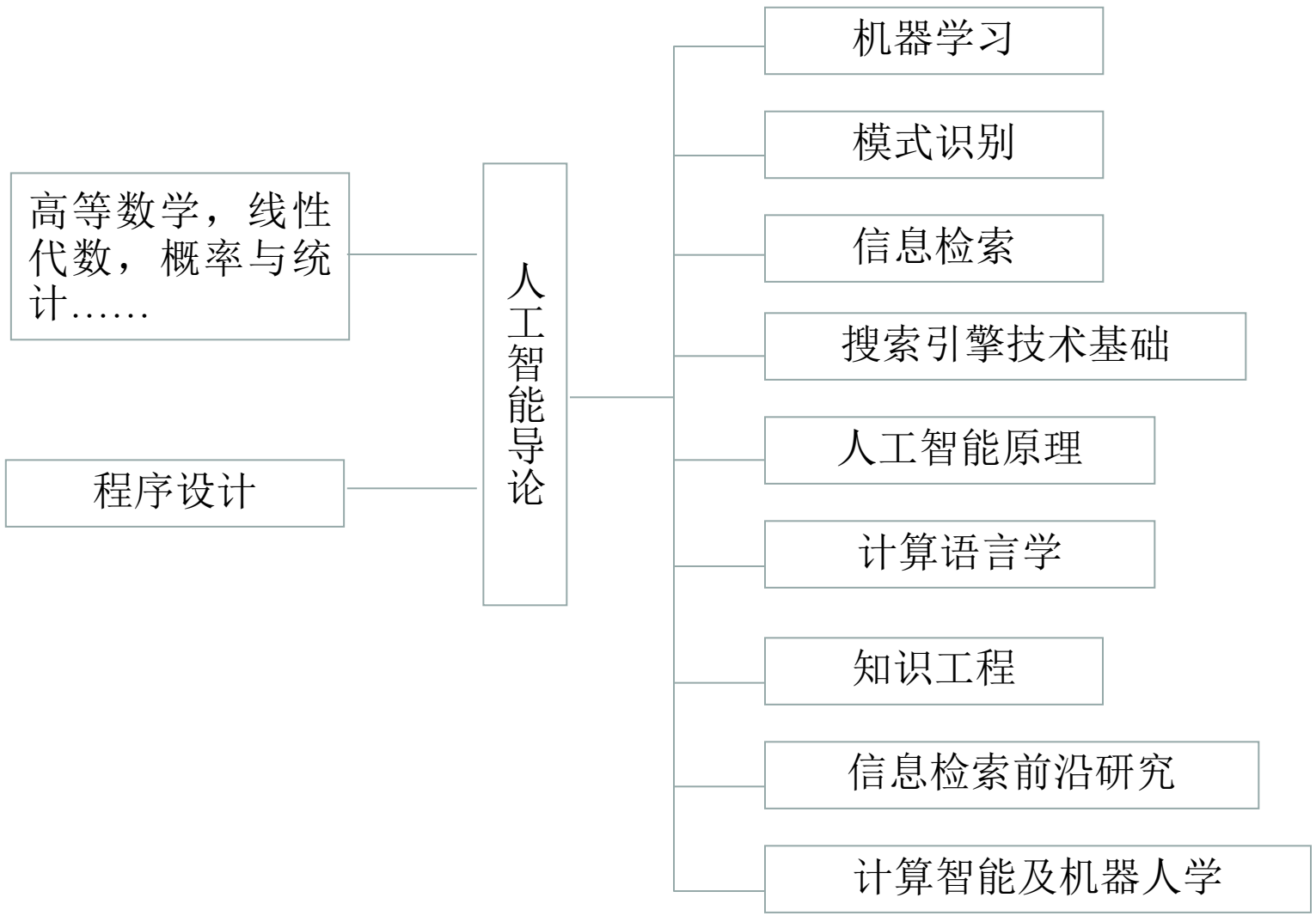
- **绪论**
  - 什么是人工智能；图灵测试；希尔勒的中文屋子；人工智能的研究目标；人工智能的发展简史；
- **第1章 搜索问题**
  - 深度优先搜索；宽度优先搜索；A*算法；改进的A*算法；
- **第2章 神经网络与深度学习**
  - 什么是神经网络，BP算法，全连接神经网络，卷积神经网络，循环神经网络
- **第3章 博弈搜索**
  - α-β剪枝，蒙特卡洛树搜索，围棋中的强化学习方法，AlphaGo实现原理
- **第4章 统计机器学习**
  - 决策树；支持向量机

学完本课后……

- 导航系统是如何实现的
- 拼音如何转化为汉字
- 计算机是如何下棋的
- 如何实现一个分类系统
- 复杂优化问题
- 深度学习是怎么回事
- 如何进行数字识别
- 如何实现文本处理
- ……

本课与其他课程的关系

——人工智能导论相关的系列课程



考核方法

- 期末考试（50%）
- 平时作业（50%）
  - 三次编程作业
  - 第一次作业不限定语言
  - 第二次作业要求python+深度学习框架
  - 第三次作业要求C++（C）
- 每次作业大概3周左右时间，但是三次作业时间上可能有覆盖
- 按时提交作业，每延迟一天扣3分（按百分制算），延迟作业提交也通过网络学堂

助教信息

- 王贝宁　Benson0704@outlook.com
  - 第一次作业
- 刘布楼　lbl20@mails.tsinghua.edu.cn
  - 第二次作业
- 詹靖涛　jingtaozhan@qq.com
  - 第三次作业
- 饶淙元　rcy22@mails.tsinghua.edu.cn
  - 第三次作业平台
- 朱书琦　zhusq22@mails.tsinghua.edu.cn
  - PyTorch教学

Honor Code

- 我们鼓励协作交流，但**对违反学术诚信的行为绝不宽容！**
- 行为认定：
  - **考试违规**：遵循清华大学《考试违规行为的处理办法》
  - **作业违规**：
    - 代码重复：两份不同同学的作业源代码（去除空格、空行与注释后）重合度超过90%
      - 有多个小题的作业中，**任何一个小题出现重复即认定重复**
      - 重合度基于计算机系TUOJ系统在线计算
      - **注意，每次提交都计入查重！**
    - 报告重复：两份不同同学的报告纯文本（去除空格、空行与符号后）重合度超过80%
      - 有多个小题的作业中，**任何一个小题出现重复即认定重复**
      - 基于转txt格式后Linux diff命令计算
      - 包含资源引用标识且用“”标注出的内容不纳入查重
  - **行为认定不考虑具体原因，不区分抄袭与被抄袭**

Honor Code

- 行为惩罚：
  - **第一次违规**：对应整个作业或考试按0分记
  - **第二次违规**：课程总分按不及格记
- 可以与同学讨论课程内容，但必须独立完成你的解答、证明和代码编写，并提交个人作业。
- 切勿抄袭他人作业或从互联网上寻找答案，不要允许他人抄袭你的作业。
- 不论是作业还是项目，都应当引用所有参考过的资料来源，无论是个人交流、书籍、论文、网站等。

参考书目

- 马少平，《艾博士：深入浅出人工智能》

参考书目

- 马少平，《艾博士：深入浅出人工智能》，清华大学出版社
  - 配套资源：公众号“跟我学AI”、B站“马少平”
- 斯图尔特·罗素，彼得·诺维格，《人工智能-现代方法（第四版）》
- 李航，《机器学习方法》，清华大学出版社
- IanGoodfellow、YoshuaBengio和AaronCourville著，《深度学习》，人民邮电出版社

有关书籍

- 尼克 人工智能简史
- 杨立昆 科学之路
- 马丁等 计算机简史