# 绪论

## 一、什么是人工智能

从**类人**与**理性**以及**思考**与**行为**这两个维度来看，有4种可能的组合：

| 类人思考 | 理性思考 |
| --- | --- |
| 类人行为 | 理性行为 |

### 类人行为：图灵测试方法

- 由艾伦·图灵（Alan Turing）提出
- 这个思想实验被提出的目的是**判断机器能否像人一样思考**

计算机需要具备下列能力：

- **自然语言处理**（natural language processing），以使用人类语言成功地交流
- **知识表示**（knowledge representation），以存储它所知道或听到的内容
- **自动推理**（automated reasoning），以回答问题并得出新的结论
- **机器学习**（machine learning），以适应新的环境，并检测和推断模式

**完全图灵测试**（total Turing test），该测试需要与真实世界中的对象和人进行交互

- **计算机视觉**（computer vision）和语音识别功能，以感知世界
- **机器人学**（robotics），以操纵对象并行动

### 类人思考：认知建模方法

- **内省**（introspection）——试图在自己进行思维活动时捕获思维
- **心理实验**（psychological experiment）——观察一个人的行为
- **大脑成像**（brain imaging）——观察大脑的活动

### 理性思考："思维法则"方法

希腊哲学家亚里士多德的**三段论**（syllogism）为论证结构提供了模式，当给出正确的前提时，总能得出正确的结论

- 这些思维法则被认为支配着思想的运作，他们的研究开创了一个称为逻辑（logic）的领域
- 19世纪的逻辑学家建立了一套精确的符号系统，用于描述世界上物体及其之间的关系。人工智能中所谓的逻辑主义（logicism）流派希望在此基础上创建智能系统
- 概率（probability）论允许我们在掌握不确定信息的情况下进行严格的推理

### 理性行为：理性智能体方法

**智能体**（agent）就是某种能够采取行动的东西

**理性智能体**（rational agent）需要为取得最佳结果或在存在不确定性时取得最佳期望结果而采取行动

**理性行为**：做正确的事情

抽象地讲，智能体是从感知历史到行动的函数：$f : P^* \rightarrow A$

**人工智能专注于研究和构建做正确的事情的智能体**，其中正确的事情是我们提供给智能体的目标定义

**益机**（Beneficial Machine）：对人类有益的智能体

## 益机

益机：对人类**可证益**的（provably beneficial）智能体

在我们的真实需求和施加给机器的目标之间达成一致的问题称为**价值对齐问题**（value alignment problem），即施加给机器的价值或目标必须与人类的一致。

---

# 二、人工智能的基础

| 哲学 | 数学 | 心理学 | 经济学 | 语言学 | 神经科学 | 计算机工程 | 控制理论 |
|---|---|---|---|---|---|---|---|
| 逻辑，推理方法 | 形式化表示和证明 | 感知和行为 | 理性决策的形式化理论 | 知识表示 | 大脑运作的机制 | 算力 | 稳态系统，稳定性 |
| 将思想视为物理系统 | 可计算性，易处理性 | 行为主义 | | 语法 | | 编程语言，开发框架 | |
| 学习、语言、理性的基石 | 概率，微积分 | | | | | | |

## 人工智能的历史

**人工智能的诞生（1943—1956）**

- 人工神经元模型
- 图灵测试
- 达特茅斯会议：“人工智能”术语被第一次正式使用

**早期的人工智能（1952—1969）**

- 通用问题求解器
- 物理符号系统假说
- 西洋跳棋的研究
- 神经网络

**低谷时期（1966—1973）**

- 计算复杂性
- 基础结构存在限制

**专家系统（1969—1986）**

- 符号主义
- 领域特定知识
- 知识密集型系统
- 第五代计算机

**神经网络的回归（1986—现在）**

- 反向传播学习算法
- 联结主义

**概率推理和机器学习（1987—现在）**

- 隐马尔可夫模型

- 贝叶斯网络

**大数据（2001—现在）**

- 数据为王

**深度学习（2011—现在）**

- 卷积神经网络
- Transformer
- 扩散模型

## 人工智能的发展现状

## 人工智能的风险与收益

**收益**

- 减少重复性工作
- 增加商品和服务的生产
- 加速科学研究 (疾病治疗, 气候变化和资源短缺解决方案)

**风险**

- 致命性自主武器
- 监控与隐私
- 有偏决策
- 就业影响
- 安全关键型应用