



南京大學  
NANJING UNIVERSITY

# 人工智能导论

## Introduction

郭兰哲

南京大学 智能科学与技术学院

<https://www.lamda.nju.edu.cn/guolz>

Email: [guolz@nju.edu.cn](mailto:guolz@nju.edu.cn)

# Outline

---

- 课程信息
- 什么是人工智能：两种视角
  - AI Tools
  - AI Agent
- 人工智能简史：三种主义
  - 符号主义、联结主义、行为主义
- 挑战

# 课程定位

---

人工智能方向第一门专业课程

掌握人工智能的主要流派与代表性思想

具备动手实践能力

# 课程信息

## □ 上课时间:

✓ 周一2-4节 (10:10-12:00 AM)

✓ 鼓楼校区 馆3-103

## □ 课程主页:

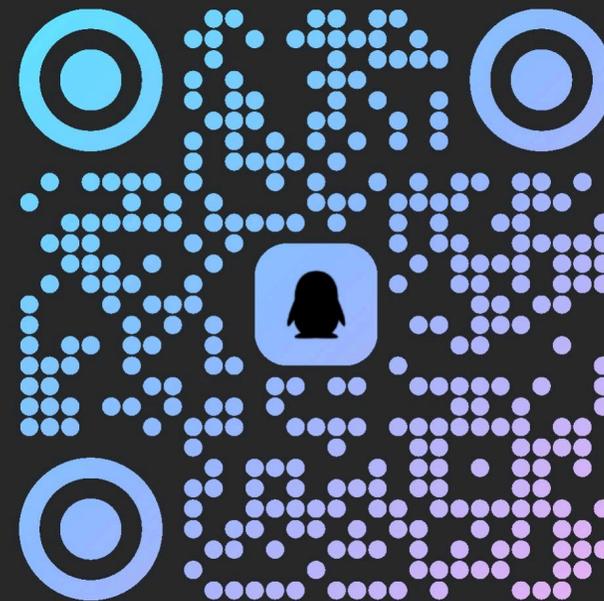
✓ <https://www.lamda.nju.edu.cn/guolz/IntroAI/sp2026/index.html>

✓ 课程讨论QQ群: 1087827042



2026春人工智能导论...

群号: 1087827042



# 课程信息

## □ 主讲教师:

- ✓ 郭兰哲, 准聘助理教授(特聘研究员), 博士生导师
- ✓ 智能科学与技术学院
- ✓ 机器学习与数据挖掘研究所 (LAMDA)
- ✓ 研究方向: **神经符号学习、智能推理与规划、Agent**
- ✓ 邮箱: [guolz@nju.edu.cn](mailto:guolz@nju.edu.cn)
- ✓ 个人主页: <https://www.lamda.nju.edu.cn/guolz/>
- ✓ 办公室: 南雍楼-东523; **有问题需要交流可以约线上会议**

# 课程信息

## 课程助教



陈旻，博士研究生

[chenyang@lamda.nju.edu.cn](mailto:chenyang@lamda.nju.edu.cn)



尚子乔，博士研究生

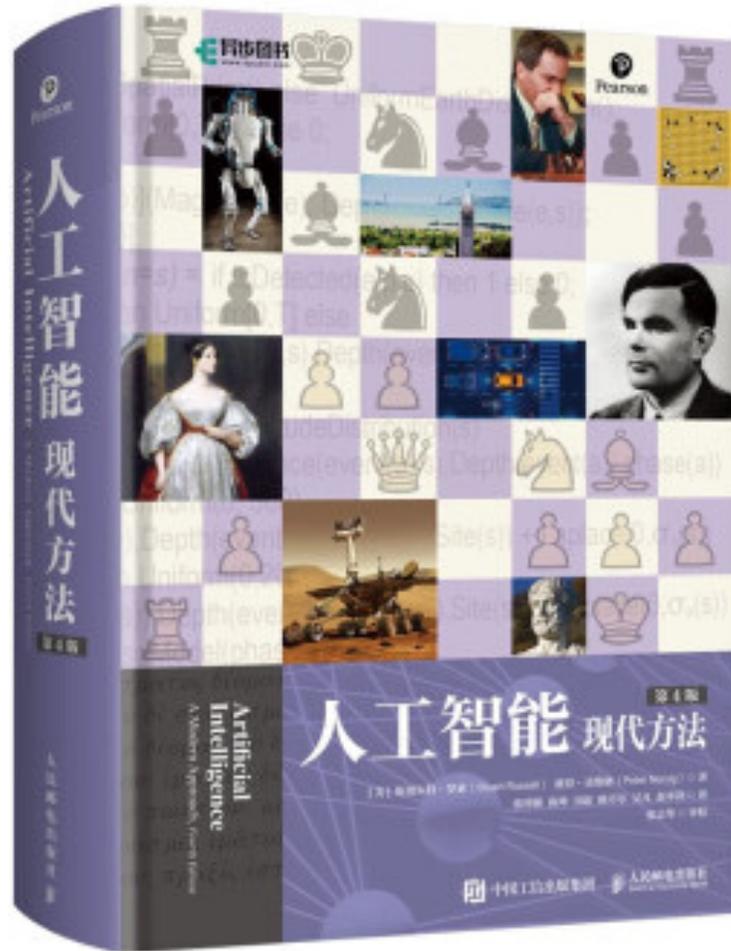
[shangzq@lamda.nju.edu.cn](mailto:shangzq@lamda.nju.edu.cn)



叶晋，本科三年级

[yej@smail.nju.edu.cn](mailto:yej@smail.nju.edu.cn)

# 参考书籍



《人工智能：现代方法》第四版  
Stuart J. Russell, Peter Norving



# 课程考核

---

## □ 成绩核算：

- ✓ 没有期中、期末考试
- ✓ 平时作业：4次理论作业+6次实践作业，60%
- ✓ 课程设计：自主选题，单人完成，40%
- ✓ bonus：课堂问答奖励3分

# 课程设计

- 为鼓励大家接触与了解人工智能科研与实践，课程设计部分提供若干个候选任务，你可以任选其一完成
- 具体来说，你需要先了解该任务的主要内容和相应的评测数据集，然后要在至少1个数据集上跑通1种算法，并得到结果；在此基础上，你还可以进一步调研更多的能够用于提升性能的方法或者自行设计算法，并进行实验分析
- 最后，你需要提交一份技术报告和对应的代码，我们会仿照学术会议的方式，构建一个线上提交平台，并邀请同学们担任审稿人，为其他同学进行评分。

# 课程设计

- 方向一：大模型数学推理增强

- Benchmark:

- [GSM8K](#)

- [Math](#)

- 参考思路：可以首先使用各种开源基座LLM得到评测结果，然后尝试调研提升基座模型数学推理能力的方法并试运行。

- 方向二：大模型逻辑推理增强

- Benchmark:

- [ProofWriter](#)

- 参考思路：可以先尝试使用各种开源基座模型得到结果，然后尝试调研更先进的方法，下面给出了一些可以参考的文献：

- [Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning](#)

- [Faithful Logical Reasoning via Symbolic Chain-of-Thought](#)

- [VeriCoT: Neuro-symbolic Chain-of-Thought Validation via Logical Consistency Checks](#)

# 课程设计

---

- 方向三：基于大模型的旅行规划
  - Benchmark:
    - [TravelPlanner](#)
    - [ChinTravel](#)
  - 参考文献
    - [Mind the Gap to Trustworthy LLM Agents: A Systematic Evaluation on Constraint Satisfaction for Real-World Travel Planning](#)
- 方向四：大模型智能体
  - Benchmark:
    - [WebShop](#)
    - [AlfWorld](#)
  - 参考文献
    - [ReAct: Synergizing Reasoning and Acting in Language Models](#)

# 课程设计

- 方向五：抽象视觉推理
  - Benchmark:
    - [ARC](#)
    - [NSA: Neuro-symbolic ARC Challenge](#)
  - 参考思路
    - ARC Challenge是一个著名的抽象视觉推理比赛，你可以调研该比赛的数据形式以及相应的解决方案。
- 方向六：开放词汇目标检测
  - Benchmark:
    - [MS-COCO](#)
    - [LVIS](#)
  - 参考思路：
    - 可以使用开源模型（如 Grounding DINO 或 GLIP）得到检测结果，然后尝试调研提升模型泛化能力的方法并试运行。
- 方向七：零样本提示分割
  - Benchmark:
    - [PASCAL VOC 2012](#)
    - [ADE20K](#)
  - 参考思路：
    - 使用开源模型（如SAM系列模型），通过中心点或边界框作为提示（Prompt），得到分割结果，然后尝试调研提升模型鲁棒性（如存在模糊边界或遮挡时）的方法并试运行。。

# 学术诚信

1. 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程
2. 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**
3. 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**，因此，请主动防止自己的作业被他人抄袭

# 科研实习生招募

 本科生科研入门学习大纲(Update: 2026.02.11)

前言

学习目标:

阶段1: 人工智能与大模型基础 (进组考核内容)

1. 神经网络与深度学习基础
2. 大语言模型 (LLMs)
3. 视觉与多模态大模型

阶段2: 文献阅读与代码实践 (科研入门培训)

研究方向初步认知

1. 神经符号学习(Neuro-Symbolic Learning)
2. 大模型推理与规划(LLM/MLLM Reasoning & Planning)
3. 智能体(Agent)

选题实战 (任选其一)

- 方向 1: 大模型逻辑推理
- 方向 2: 大模型数学、几何推理
- 方向 3: 视觉图像推理
- 方向 4: 抽象视觉推理
- 方向 5: Travel Agent (Tool-Use & Planning)
- 方向 6: Game/Embodied Agent
- 方向 7: Symbolic Regression

主要研究方向:

1. 如何提升大模型的推理能力、多模态推理能力
  - 逻辑推理、数学推理、司法推理、图像推理
2. 大模型Agent
  - OS Agent: 调用各种工具完成用户任务(比如旅行规划)
  - Game/Embodied Agent: 自动完成复杂游戏任务, 比如我的世界、开放世界游戏等或者机器人任务
3. 科学发现&符号回归
  - 给AI模型大量物理观察数据, 能否自动发现牛顿定理; 自动发现新算法、新方案等

<https://dy8q0bnq8y.feishu.cn/wiki/OhB5wh8j5ilouvknYEc8AqQnJh>

# Outline

---

- 课程信息
- 什么是人工智能：两种视角
  - AI Tools
  - AI Agent
- 人工智能简史：三种主义
  - 符号主义、联结主义、行为主义
- 挑战



什么是人工智能？

先看看人工智能

可以做什么？

# 艺术创作

## 图像生成



在美国科罗拉多艺术博览会美术比赛中，游戏设计师 Jason Allen 的作品《太空歌剧院》夺冠，该副画作是他使用 AI 作图工具 MidJourney 完成

## 视频生成



OpenAI Sora、Seedance ...

# 数学推理

RESEARCH

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

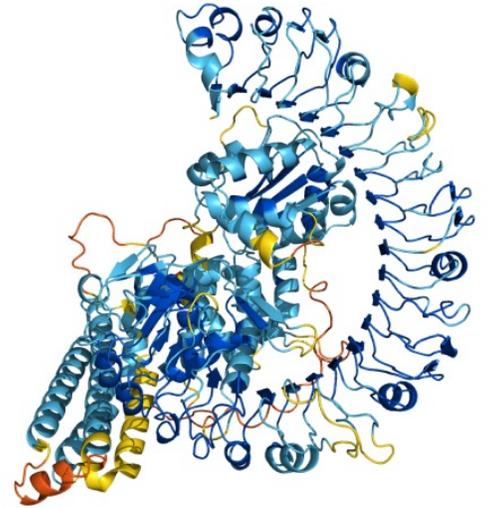
21 JULY 2025

Thang Luong and Edward Lockhart



# 科学研究

- 2020年11月，DeepMind的人工智能程序AlphaFold 2在蛋白质结构预测大赛CASP 14夺冠，对大部分蛋白质结构的预测与真实结构只差一个原子的宽度，达到了人类利用冷冻电镜等复杂仪器观察预测的水平
- 2021年8月，DeepMind宣布已将人类的98.5%的蛋白质预测了一遍；此前，已知氨基酸顺序的蛋白质分子的三维结构被看清的不到0.1%
- 2022年8月，DeepMind宣布，AlphaFold可以预测出2亿多个蛋白质结构，几乎涵盖了地球上所有已进行过基因组测序的生物体，其中35%已达到实验手段所能获取的结构精度



## FARS: 228小时狂飙100篇论文、烧光114亿Token

### FARS

#### Compute-Matched Audit of Diffusion LMs' Planning Advantage on Procedurally Generated

Project Queue • LIVE

Ideation Planning Experiment Writing

**In Progress** Completed

- FA0328
- FA0404
- FA0157
- FA0354
- FA0385**
- FA0380
- FA0373
- FA0171
- FA0376

Manual Auto

tables, ensure main results have comprehensive comparison table

- ✓ Study Analytical Plot Exemplars: Read ALL exemplar analytical plots to learn visual styles for experimental result visualizations
- ✓ Generate Analytical Plots: Generate ONLY the figures planned in Task 9, apply learned visual styles, iteratively optimize
- ✓ Study Table Exemplars: Read ALL exemplar tables to learn table design patterns
- ➔ Design Table Specifications: Read individual experiment files to capture ALL experimental conditions and methods, copy data values EXACTLY

Develop Story Arc: Develop story arc and identify key claims with supporting evidence

Finalize Blueprint: Add Paper Outline and Meta Information sections to complete PAPER\_BLUEPRINT.md

- Read exp/EXPERIMENT\_RESULTS/qwen\_best\_of\_k/RESULTS.json
- Read exp/EXPERIMENT\_RESULTS/dream\_diffusion\_baseline/RESUL...
- Read exp/EXPERIMENT\_RESULTS/qwen\_greedy\_baseline/RESULTS.j...
- Read exp/EXPERIMENT\_RESULTS/p75\_sensitivity/RESULTS.json

#### Countdown Task

Number of Samples (k)	Qwen2.5-7B Best of k Accuracy	Dream-7B Accuracy
1	0.05	0.066
4	0.15	0.066
8	0.25	0.066
16	0.35	0.066
32	0.4	0.066
35	0.4	0.066

#### Mini Sudoku Task

Number of Samples (k)	Qwen2.5-7B Best of k Accuracy	Dream-7B Accuracy
1	0.05	0.776
4	0.2	0.776
8	0.35	0.776
16	0.45	0.776
32	0.65	0.776
39	0.7	0.776

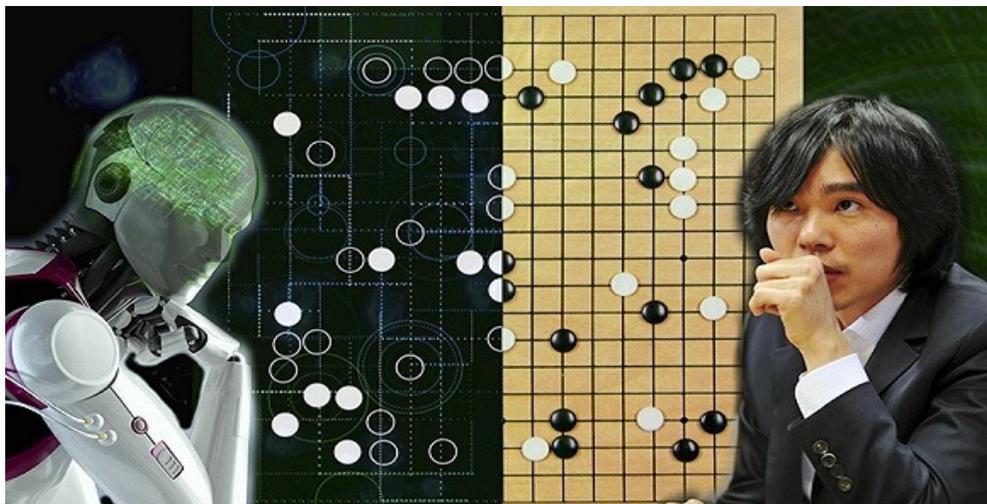
T+ 389:53:01

409 Hypothesis 152 Paper

Launched Hypothesis x400

19.9B Tokens 169K Cost(\$)

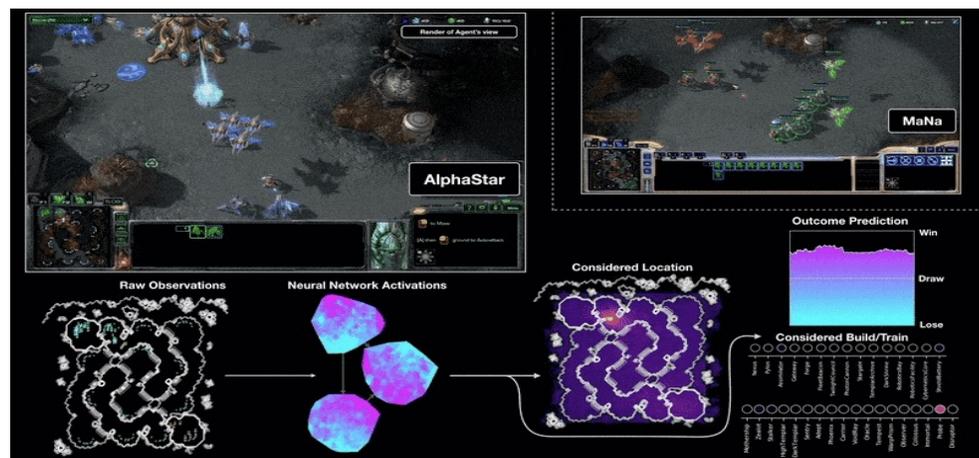
# 游戏博弈



Go (DeepMind AlphaGo 2016)



Poker (CMU and Facebook 2019)



StarCraft II (DeepMind AlphaStar 2019)



王者荣耀(腾讯绝悟 2020)

从日常对话、艺术创作，  
到科学研究、军事政治

所以，到底什么是人工智能？

# 视角一：智能工具(AI Tools)

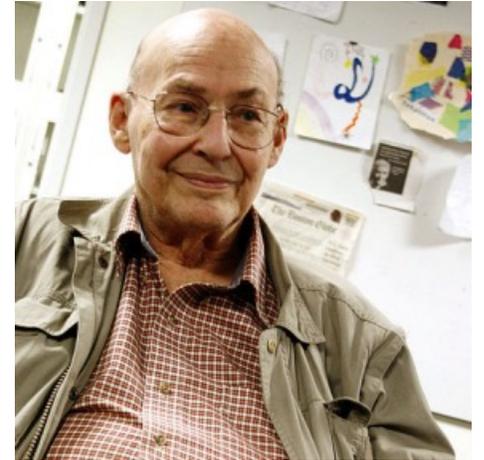
## □ 智能工具(AI Tools)

让机器做事时聪明一点

“人工智能就是让机器来完成那些如果由人来做则需要智能的事情的科学”

解读：

- 如果某件事情需要智能，通过机器来做，就是人工智能
- 不要求“全面”达到人类智能水平
- “做事”就行，不必具备“自主意识”“情感”.....
- .....

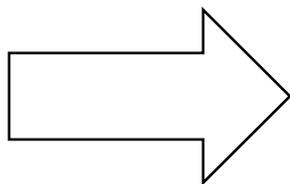


Marvin Minsky  
(1927-2016)  
人工智能奠基者之一  
1969年图灵奖

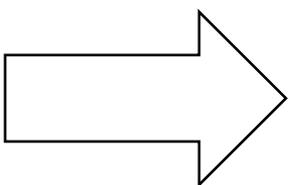
# 一个类比



人的智能行为



人工智能



人工智能重要，是因为能造出“智能工具”（类比：飞机）

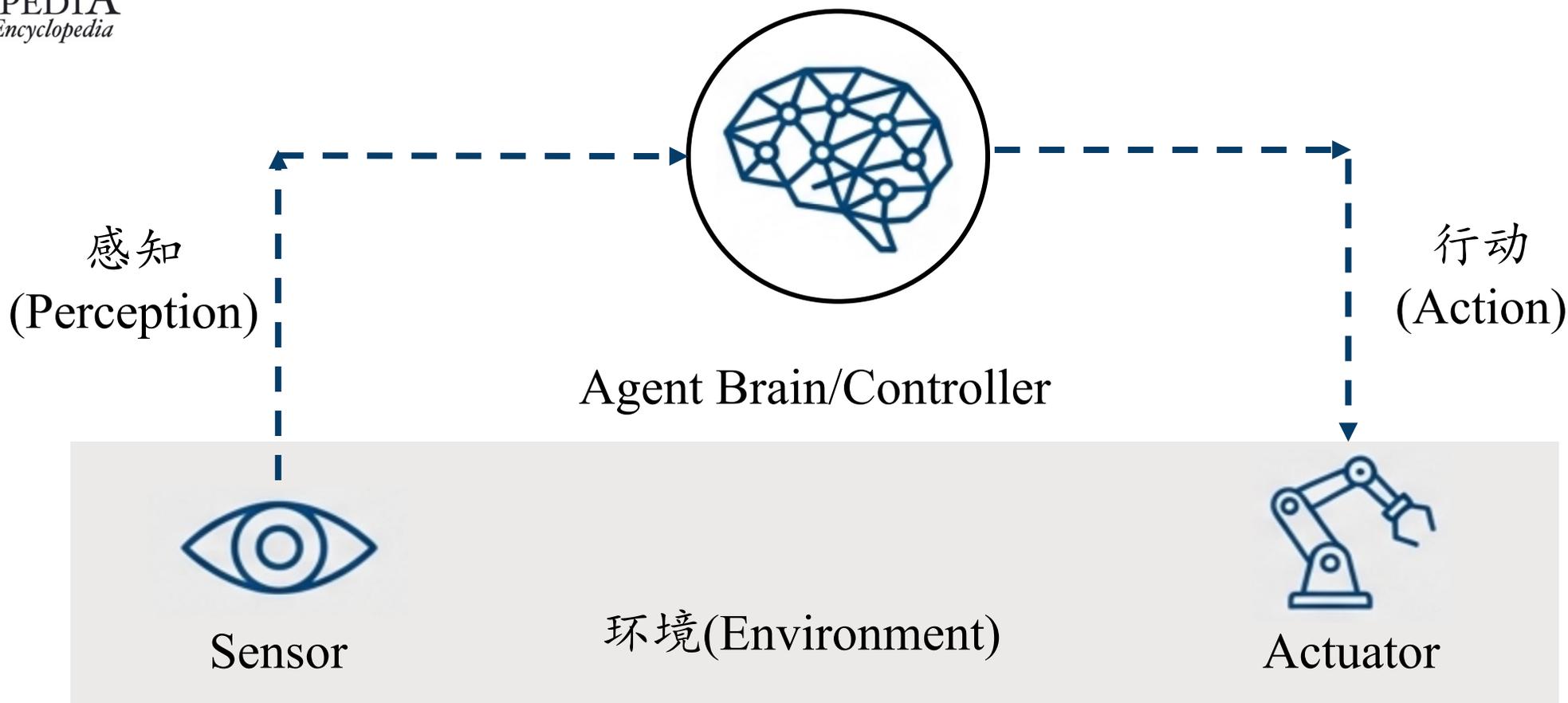
- 造飞机的人不会关心飞机有没有“意识”、会不会“疼”
- 更不会关心飞机是否“全面达到”鸟的能力（例如：下蛋）

# 视角二：智能体(Agent)



WIKIPEDIA  
The Free Encyclopedia

智能体是指能够自主**行动**、**学习**或利用**知识**来实现**目标**的实体，  
它通过传感器和相应的执行器**观察环境**



# 一个例子：Game Agent

## Knowledge

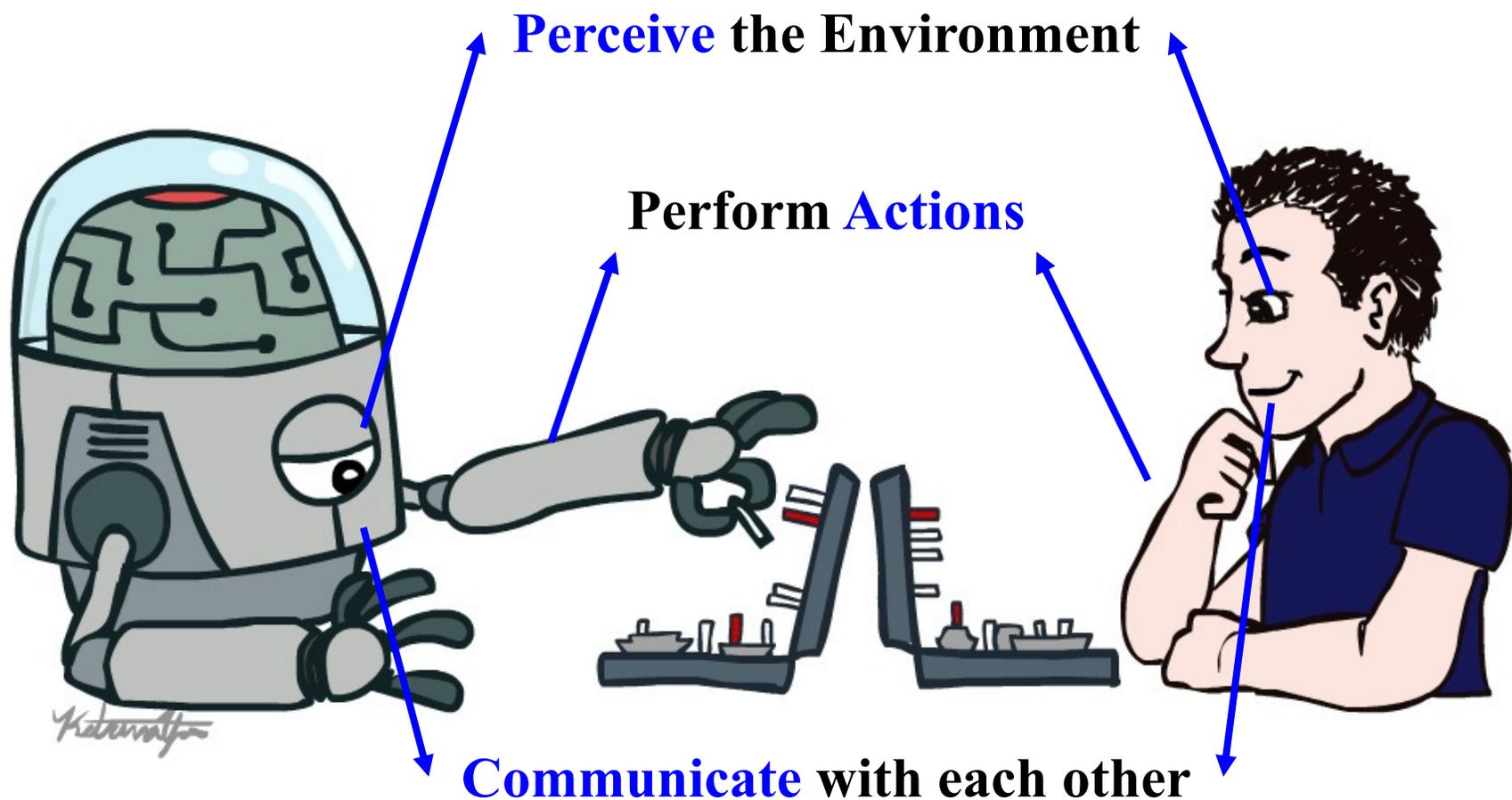
- 游戏规则
- 经验

## Reasoning

- 下一步怎么操作？

## Learning

- 玩的越多能力越强



# 逻辑智能体

## 基于人工定义的符号与逻辑规则

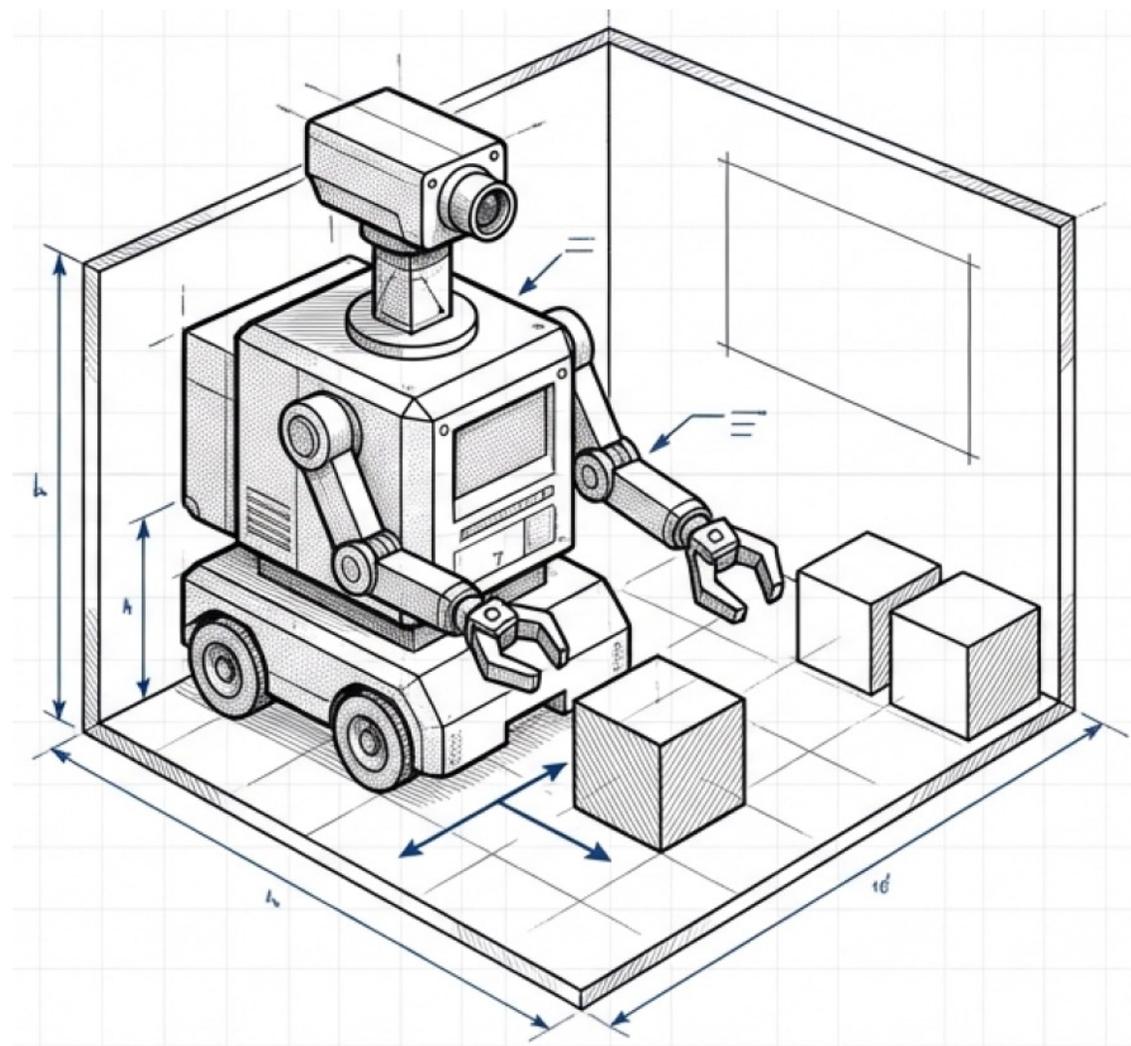
### 关键案例

先驱：Shakey the Robot (1969)

第一个能够推理自身行为，并与物理世界交互的通用移动机器人，能够自主规划路径和动作完成“推箱子”任务

优势：严谨的逻辑性、可解释性

局限：缺乏泛化能力，感知能力弱



# 强化学习智能体

基于强化学习与环境海量交互试错

关键案例

- 围棋：AlphaGo
- 游戏智能体
  - Atari, Open-AI Five, AlphaStar...

优势：封闭环境中掌握超越人类的复杂技能

局限：样本效率低，难以在不同场景迁移



# 大模型智能体

## 基于LLM的通用世界知识作为大脑

### 关键案例

- **OS Agent:** 手机助手
- **Web Agent**
- **Embodied Agent**
- .....

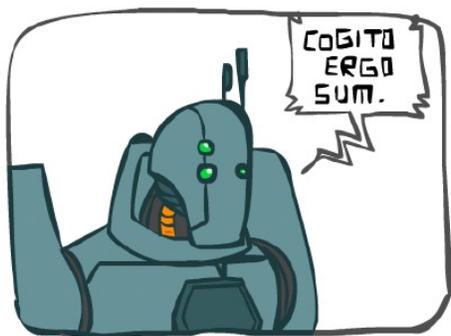
**优势:** 强大的零样本交互能力

**局限:** 模型幻觉、长程推理规划能力不足



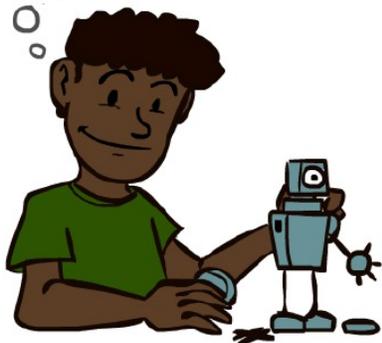
# 两种视角

- 智能工具(AI Tools)
  - Try to solve problems
  - Human cannot do well

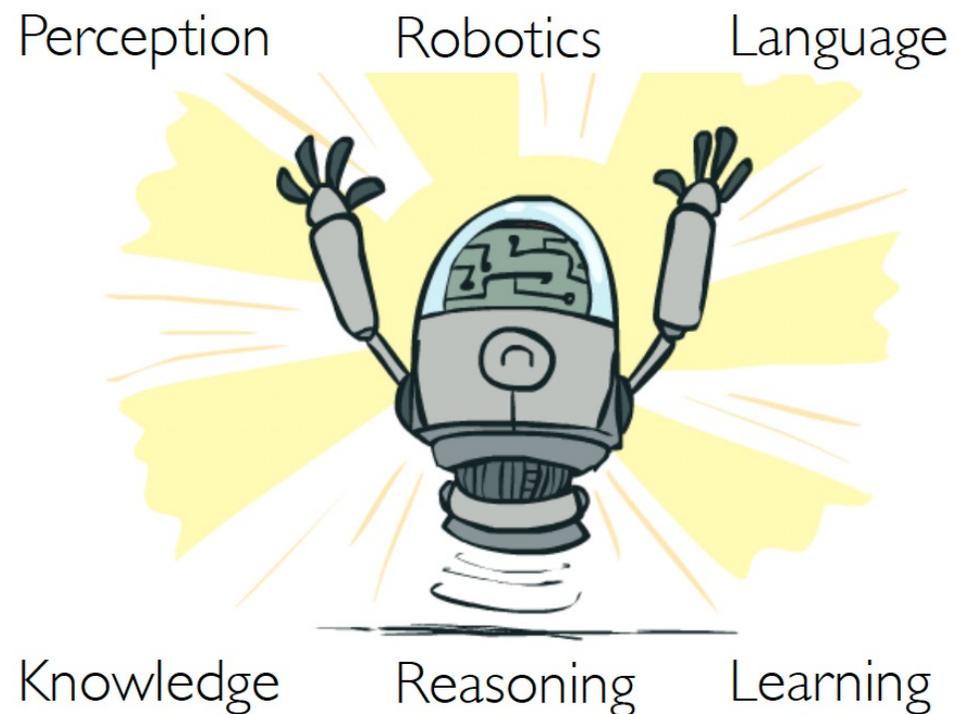


- Planning
- Scheduling
- Medical Image
- Topic Model
- Prediction

...



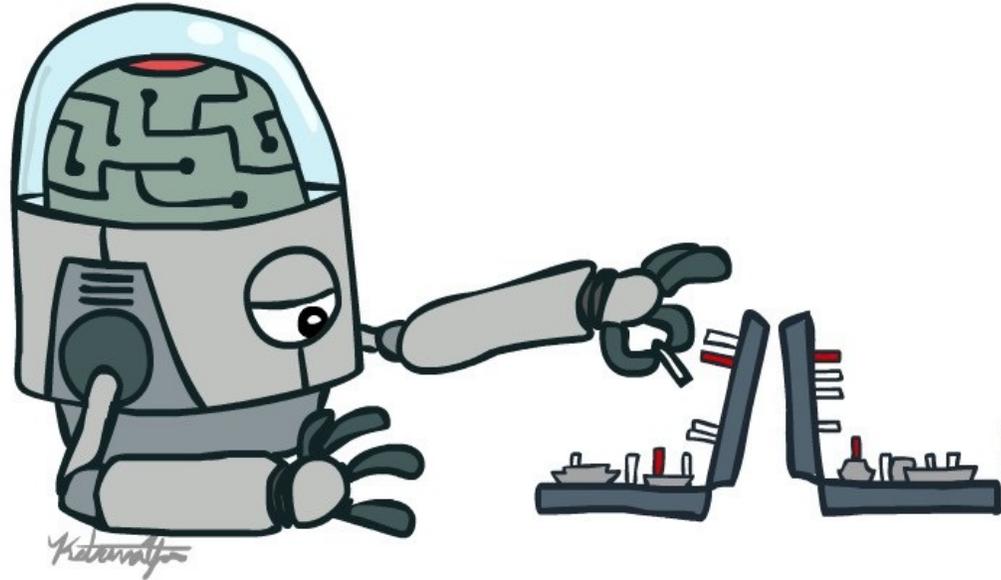
- 智能体(Agent)
  - Think like human
  - Act like human



# Agent视角仍有很多挑战

## Machine

narrow tasks, millions of examples



### AlphaGo

- Learn from 19.6 million games
- Only play one game Go

## Human

diverse tasks, very few examples



### Human

- Learn from wide set of experiences
- Do lot of tasks

We are still very far away from **Artificial General Intelligence (AGI)**

# Outline

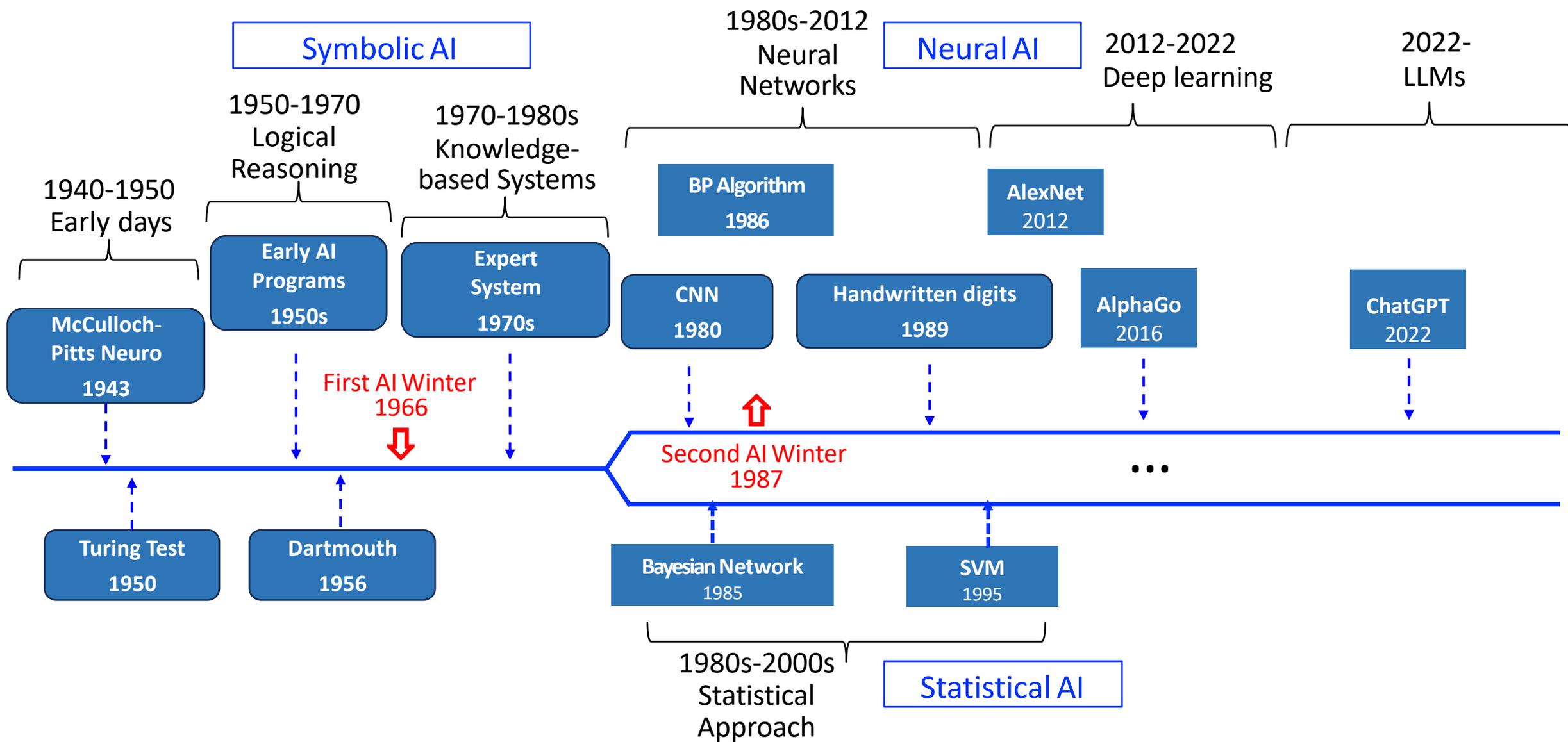
---

- 课程信息
- 什么是人工智能：两种视角
  - AI Tools
  - AI Agent
- 人工智能简史：三种主义
  - 符号主义、联结主义、行为主义
- 挑战



人工智能是如何发展到现在的？

# 人工智能简史



# 人工智能的起源

## Computing Machinery and Intelligence

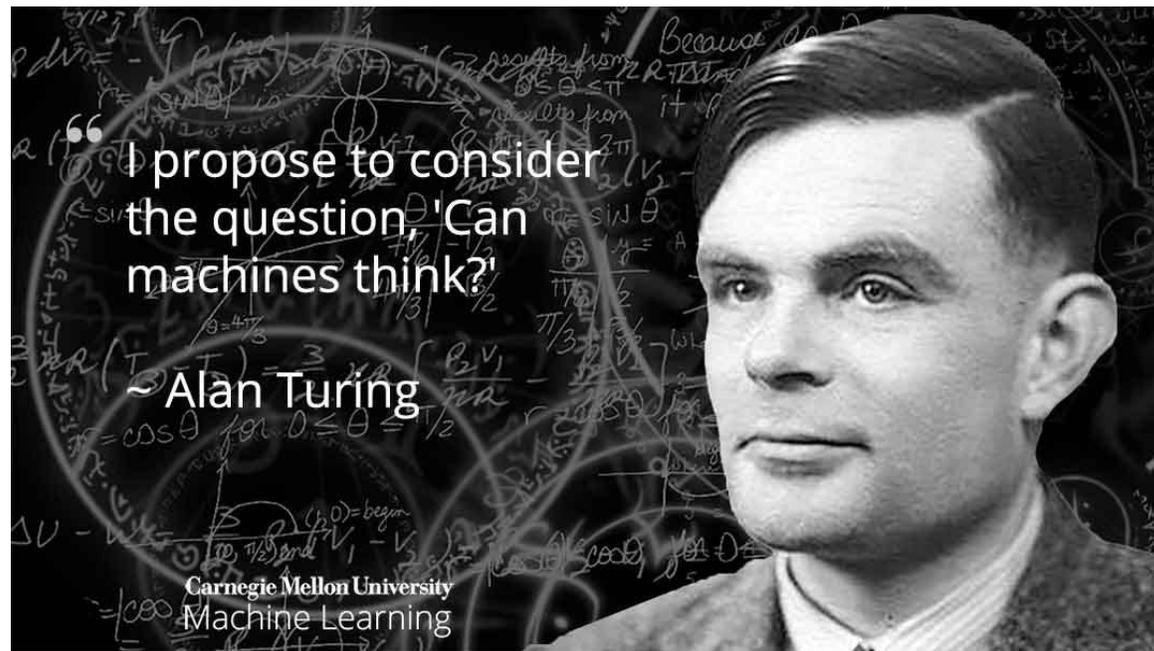
计算机器与智能

1950年

艾伦·图灵

“Can machine think?”

机器能思考吗？

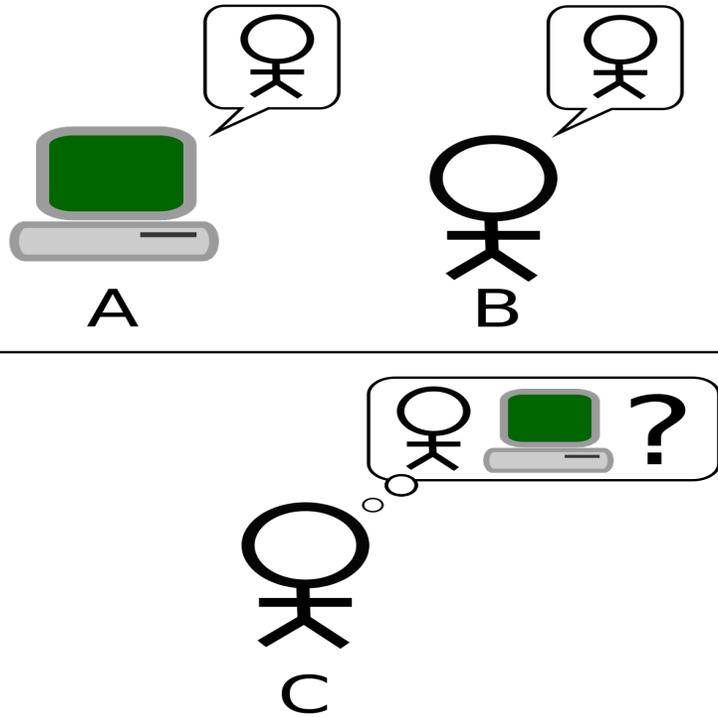


艾伦·图灵

1912-1954

# 图灵测试

如何判断机器是否具有智能？



一人扮演提问者，另一人作为被测人员。这两个人与机器分别处在3个不同的房间，

提问者通过打印问题和接收打印问题来与被测人员和被测机器进行通信

计算机尽量模仿人，如果提问者判断不出哪个回答是人，哪个回答是计算机，就可以认为这台计算机具有智能



# 1956人工智能元年

1956年的达特茅斯会议标志人工智能这一学科的诞生



约翰·麦卡锡



马文·明斯基



克劳德·香农



雷·索洛莫诺夫



艾伦·纽厄尔



赫伯特·西蒙



阿瑟·塞缪尔



奥利弗·塞尔弗里奇



纳撒尼尔·罗切斯特



特伦查德·莫尔



1956年夏 美国达特茅斯学院

# 第一阶段：推理期

- **Key Ideas:** 智能的核心在于符号与逻辑推理；把人类的知识和推理过程，翻译成机器能理解的符号和规则，机器也能像人类一样思考

- **Machine can play games**

- IBM亚瑟·塞缪尔 (Arthur Samuel) 的跳棋程序，1962年战胜美国州跳棋冠军



Checkers  
Game



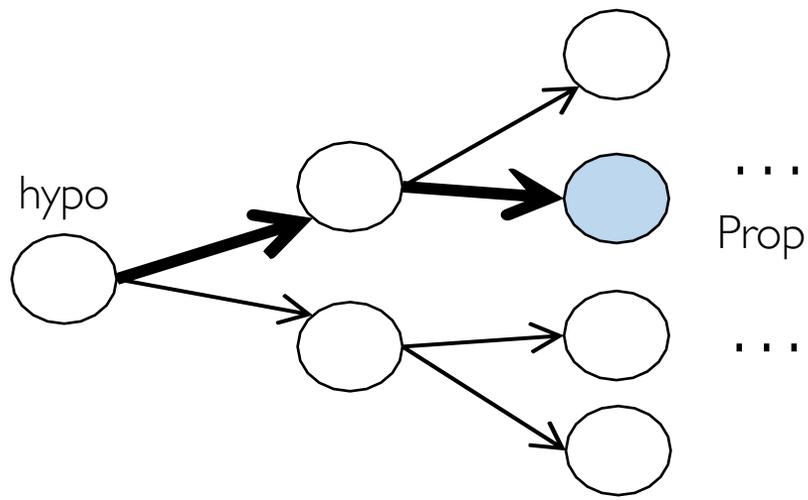
Playing  
checkers  
on the  
IBM 701

# 第一阶段：推理期

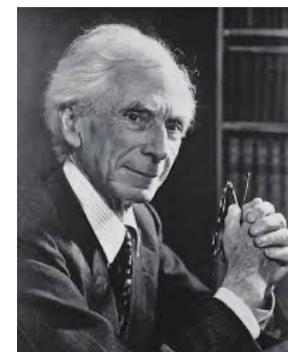
- **Machine Can Prove Theorems**

- 逻辑理论家 (Logic Theorist)

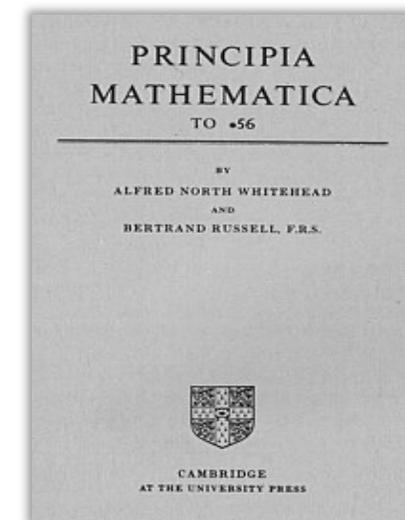
- 把已知的公理和定理作为“事实” (符号表示)
- 把逻辑推理规则 (比如“如果A为真, 且A能推出B, 则B为真”) 作为“操作” (规则)
- 在一个巨大的“可能性之树”上进行搜索, 尝试从公理出发, 一步步推导出目标定理



西蒙与纽厄尔的逻辑理论家, 可以证明  
《数学原理》第二章52个定理中的38个



Bertrand Russell



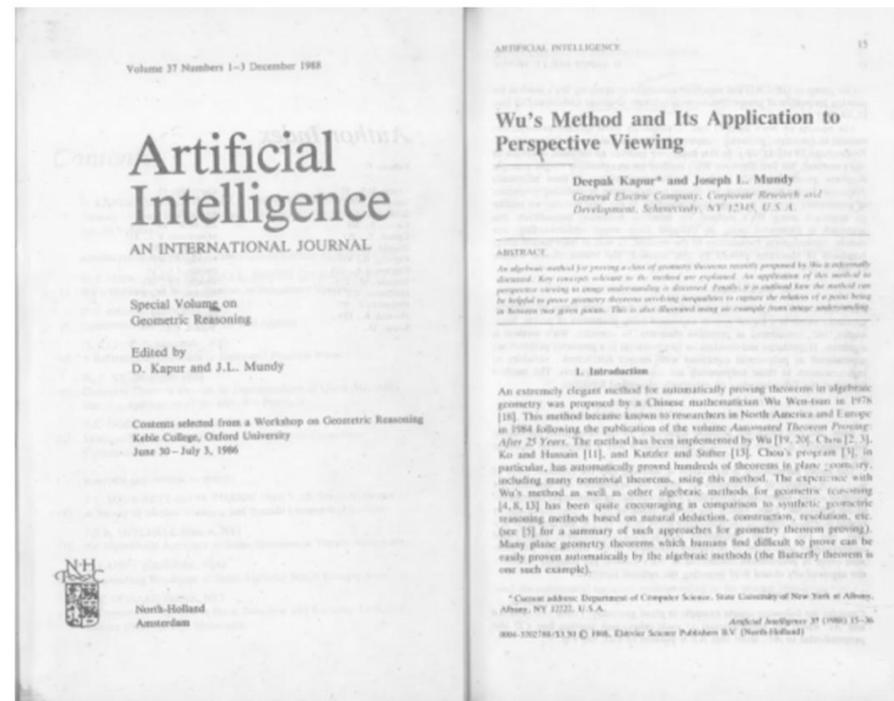
# 第一阶段：推理期

- **吴文俊**：提出用计算机证明几何定理的“吴方法”，开创了机器几何定理证明的方向，是国际自动推理领域的先驱性的工作 -> **“数学机械化”**
- 1977年大年初一，手算成功验证了几何定理机器证明的方法，随后在一台由北京无线电一厂生产的长城203上证明了西姆森定理



吴文俊 (1919-2017)

“所有的问题都可以转变成数学问题，所有的数学问题都可以转变成代数问题，所有的代数问题都可以转变成解方程组的问题，所有解方程组的问题都可以转变成解单变元的代数方程问题”



1988年《人工智能》特辑开篇对吴方法的概述

# 第一阶段：推理期

## 第 2 章 自动定理证明兴衰纪

As a material machine economises the exertion of force, so a symbolic calculus  
economises the exertion of intelligence ... the more perfect the calculus, the smaller  
the intelligence compared to the results.

就像机器能省体力一样，符号演算能省脑力。

演算越完美，付出的脑力就越少。

——W. E. Johnson (约翰逊)

Proof is cultivated reasoning.

证明就是讲究的推理。

——Bruno Buchberger (布赫贝格尔)

<https://www.ituring.com.cn/book/tupubarticle/19224>

# 第一阶段：推理期

<https://arxiv.org/pdf/2404.09939>

Published as a conference paper at COLM 2024

## A Survey on Deep Learning for Theorem Proving

Zhaoyu Li<sup>1</sup>, Jialiang Sun<sup>1</sup>, Logan Murphy<sup>1</sup>, Qidong Su<sup>1</sup>, Zenan Li<sup>2</sup>, Xian Zhang<sup>3</sup>  
Kaiyu Yang<sup>4\*</sup>, Xujie Si<sup>1,5</sup>

<sup>1</sup>University of Toronto, <sup>2</sup>Nanjing University, <sup>3</sup>Microsoft Research Asia, <sup>4</sup>Meta FAIR,

<sup>5</sup>CIFAR AI Chair

{zhaoyu, six}@cs.toronto.edu

### Abstract

Theorem proving is a fundamental aspect of mathematics, spanning from informal reasoning in natural language to rigorous derivations in formal systems. In recent years, the advancement of deep learning, especially the emergence of large language models, has sparked a notable surge of research exploring these techniques to enhance the process of theorem proving. This paper presents a comprehensive survey of deep learning for theorem proving by offering (i) a thorough review of existing approaches across various tasks such as autoformalization, premise selection, proofstep generation, and proof search; (ii) an extensive summary of curated datasets and strategies for synthetic data generation; (iii) a detailed analysis of evaluation metrics and the performance of state-of-the-art methods; and (iv) a critical discussion on the persistent challenges and the promising avenues for future exploration. Our survey aims to serve as a foundational reference for deep learning approaches in theorem proving, inspiring and catalyzing further research endeavors in this rapidly growing field. A curated list of papers is available at <https://github.com/zhaoyu-li/DL4TP>.

<https://machine-learning-for-theorem-proving.github.io/>

## NeurIPS Tutorial on Machine Learning for Theorem Proving

Video Recording



### Overview

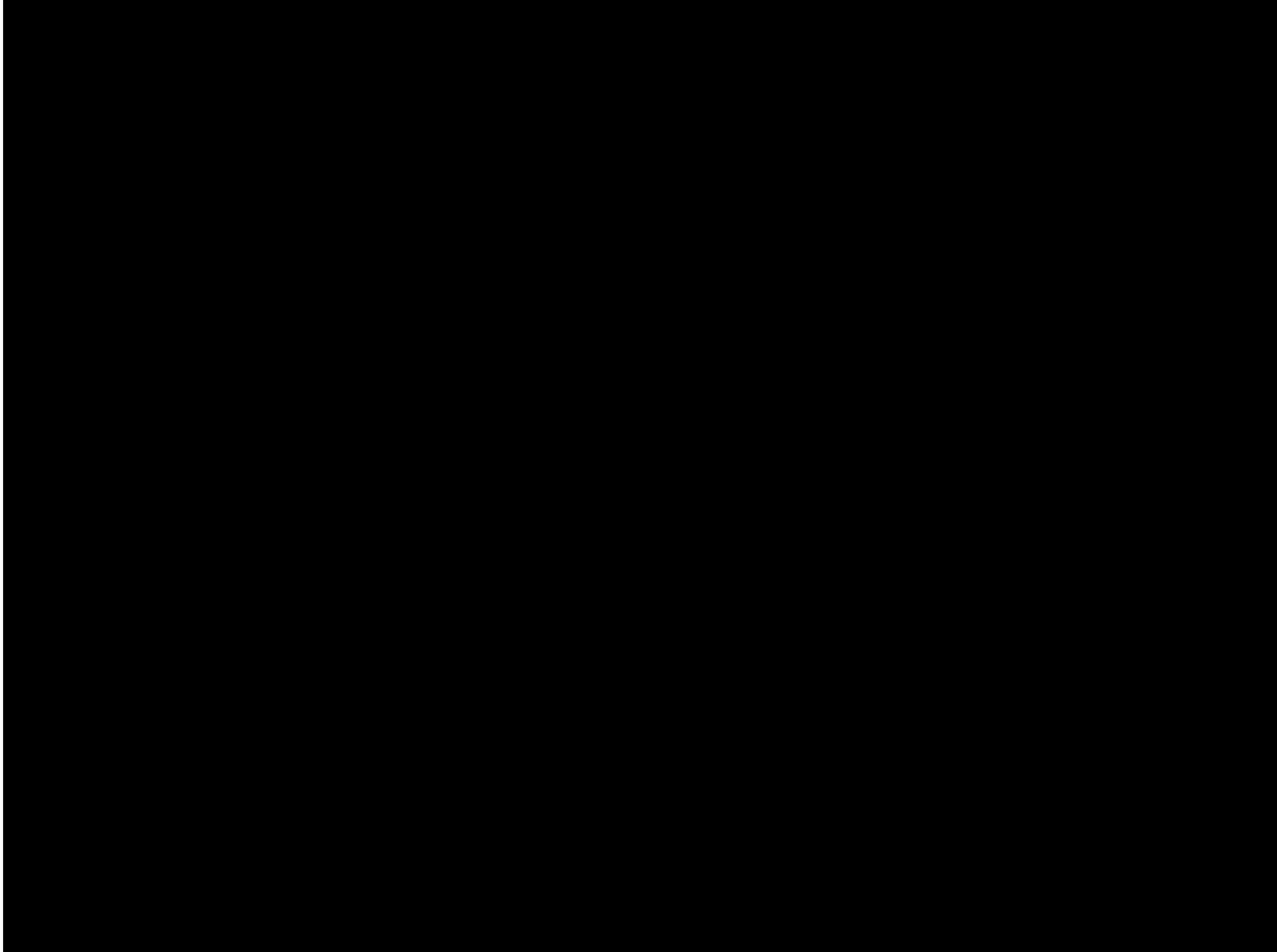
Machine learning, especially large language models (LLMs), has shown promise in proving formal theorems using proof assistants such as [Coq](#), [Isabelle](#), and [Lean](#). Theorem proving is an important challenge for machine learning: Formal proofs are computer programs whose correctness can be verified. Therefore, theorem proving is a form of code generation with rigorous evaluation and no room for the model to hallucinate, opening up a new avenue for addressing LLMs' flaws in factuality.

Despite its potential, learning-based theorem proving has significant entry barriers, primarily due to the steep learning curve for proof assistants. This tutorial aims to bridge this gap and make theorem proving accessible to researchers with a general machine learning background. To that end, our presentation will contextualize theorem proving from a machine learning perspective and demonstrate how to develop LLMs for theorem proving, using newly available open-source tools that provides interfaces to proof assistants without requiring in-depth knowledge of their internals. Furthermore, we will cover advanced topics and open problems in learning-based theorem proving, including its synergies with natural language processing and software verification.

Throughout the presentation, we will highlight several conceptual themes recurring in theorem proving that are also critical for machine learning, such as mathematical reasoning, code generation, and hallucination prevention. The panel will complement the presentation through a broader discussion of related topics such as trustworthy machine learning, LLMs for code, reasoning, and program synthesis.

# 第一阶段：推理期

---



# 第一次人工智能寒冬

## 那些年我们吹过的牛

- 赫伯特西蒙在1957年预测：“不出十年，数字计算机将成为世界象棋冠军”
- “不出十年，数字计算机将能发现并证明一个重要的全新数学定理”
- 马文明斯基在1970年预测：“在三到八年的时间里，我们将拥有一台具有普通人一般智能的机器”



# 第一次人工智能寒冬

## 英国：“莱特希尔报告” (1973)

- A类 (Advanced Automation, 高级自动化)：专注于特定应用的机器人、自动化技术等 (还行)
- B类 (Building Robots, 构建机器人)：模拟人类神经系统或行为的交叉学科研究 (有点意思, 但成果有限)
- C类 (Computer-based studies, 基于计算机的研究)：探索智能本质的AI研究, 通用人工智能 (纯属扯淡)

**结论：AI领域的大部分工作都令人失望，其承诺的目标在可预见的未来都不可能实现，除了少数特定应用领域，AI研究不值得大规模投入**

## 美国：曼斯菲尔德修正案 (Mansfield Amendment)

国防部的研究经费必须与具体的军事任务直接相关



James Lighthill

Unveiled the  
Beginning of  
the AI Winter

# 第二阶段：知识期

## 1970s-1980s: 知识期

为什么人类有智能？

因为有大量的知识

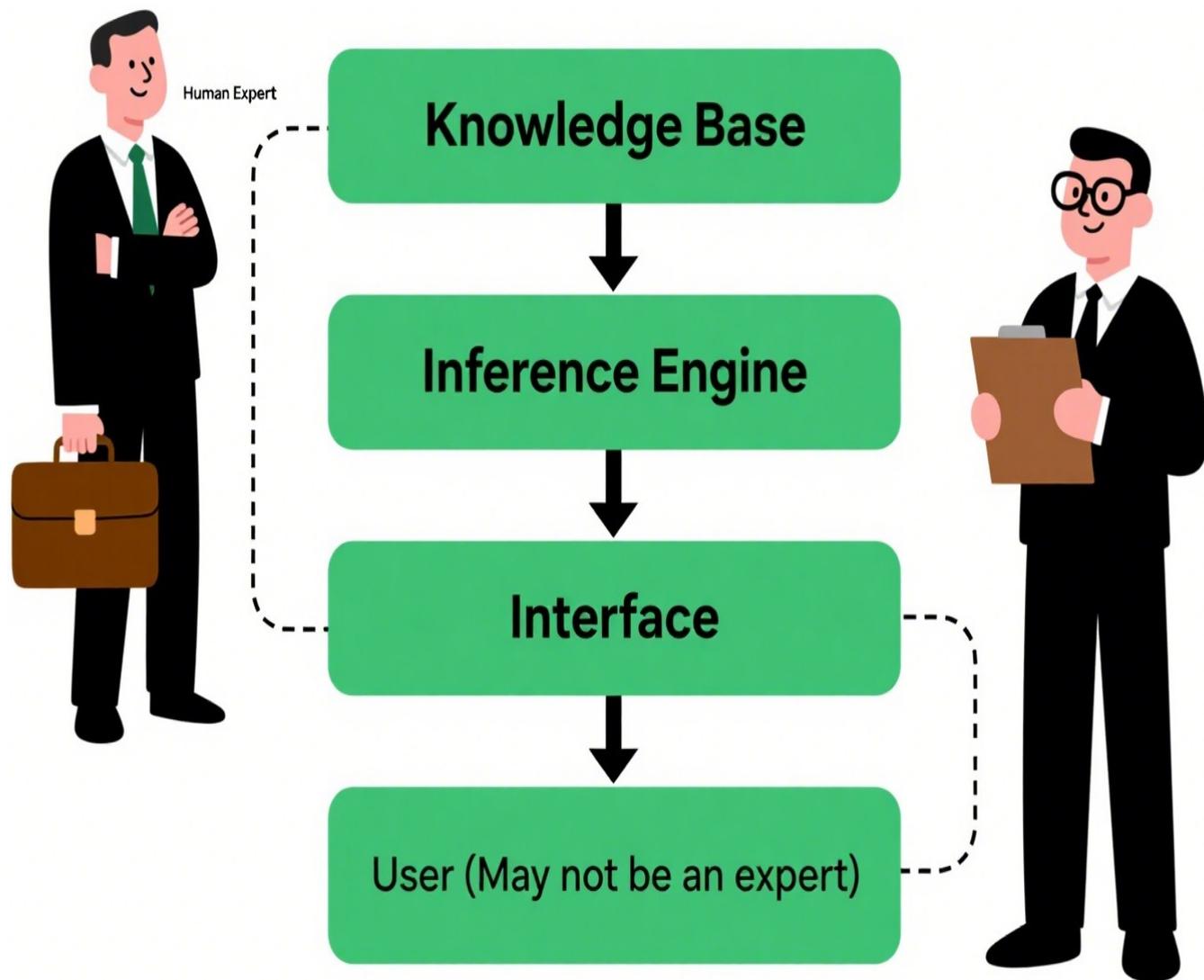
◆ 出发点：“知识就是力量！”

◆ 把人的**全部知识**放入电脑

◆ 主要成就：

- 专家系统 (Expert System)

不求样样通，但求一招精！



## 第二阶段：知识期

- **DENDRAL**: 包含丰富的化学知识，可以根据质谱数据帮助化学家推断分子结构
- **MYCIN**: 用于诊断细菌性血液感染并推荐抗生素治疗方案。知识库包含了大约600条规则，能解释推理，引入置信因子处理不确定性
- **XCON / R1**: DEC电脑配置专家，帮助人们配置虚拟地址扩展(VAX)系列计算机，为公司节省超过4000万美元

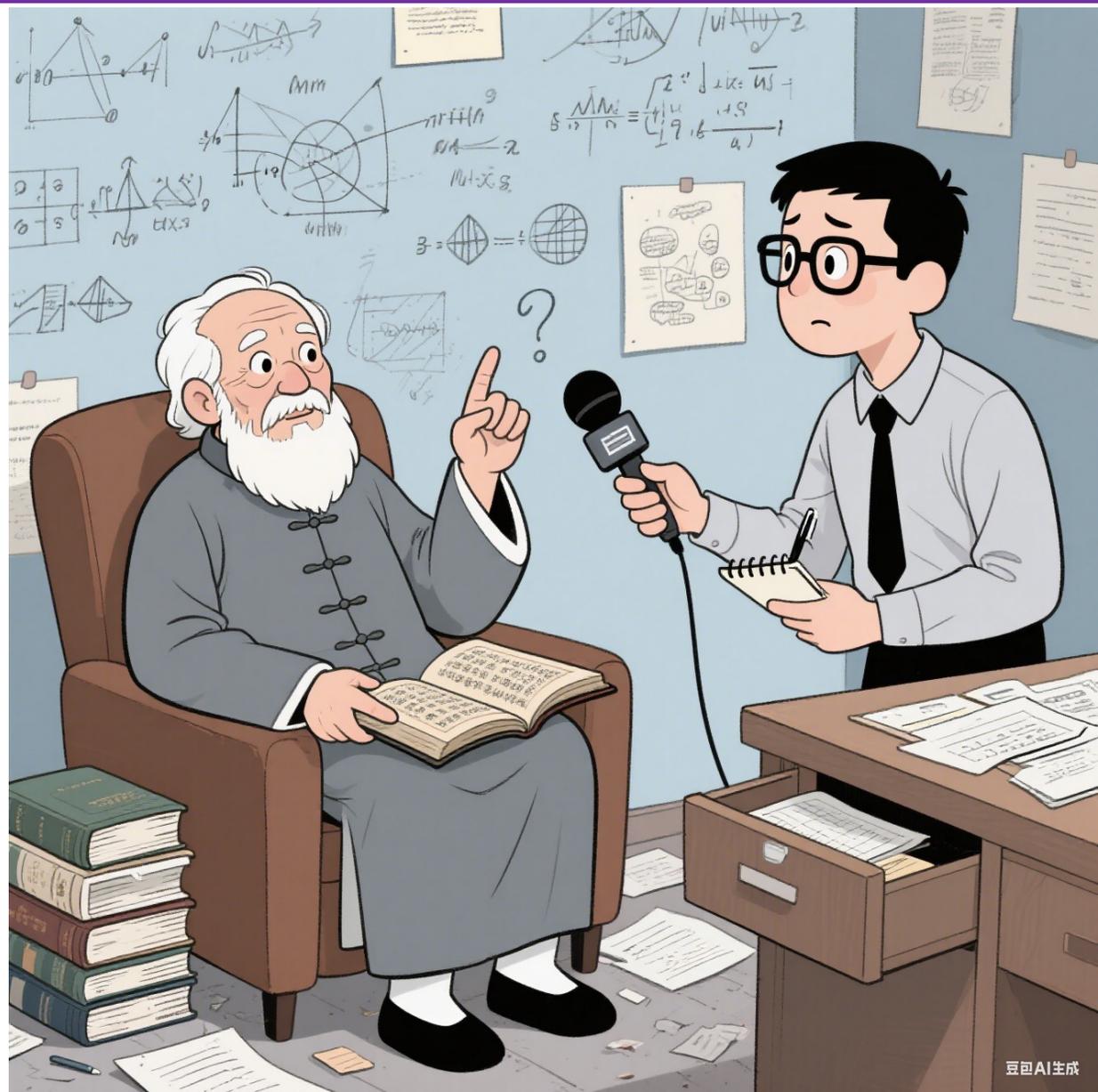


爱德华·费根鲍姆  
(1936- )  
1994年图灵奖

# 知识工程的瓶颈

## 专家系统的瓶颈

- 知识获取难
- 知识更新难
- 系统泛化难



# 第一代人工智能：符号主义人工智能 (Symbolic AI)

## □ 符号主义人工智能 (Symbolic AI)

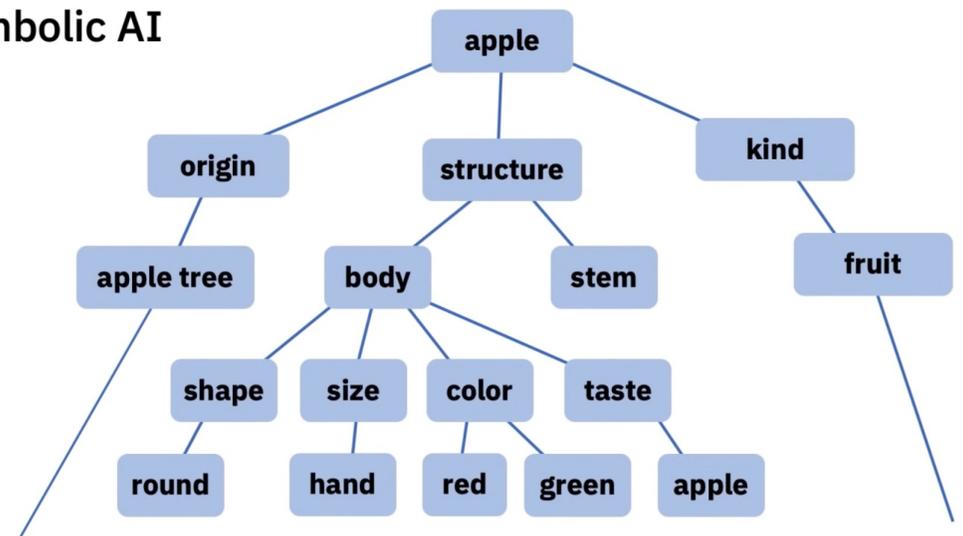
- Good Old-Fashioned AI
- 认为实现人工智能需要基于符号和逻辑系统
- 核心技术：搜索、推理、知识

## □ 关键局限

- 所有知识预先给定，难以自主学习
- 针对特定领域设计，难以泛化至未见任务
- 难以应对复杂多变的开放环境

如何识别苹果？

Symbolic AI



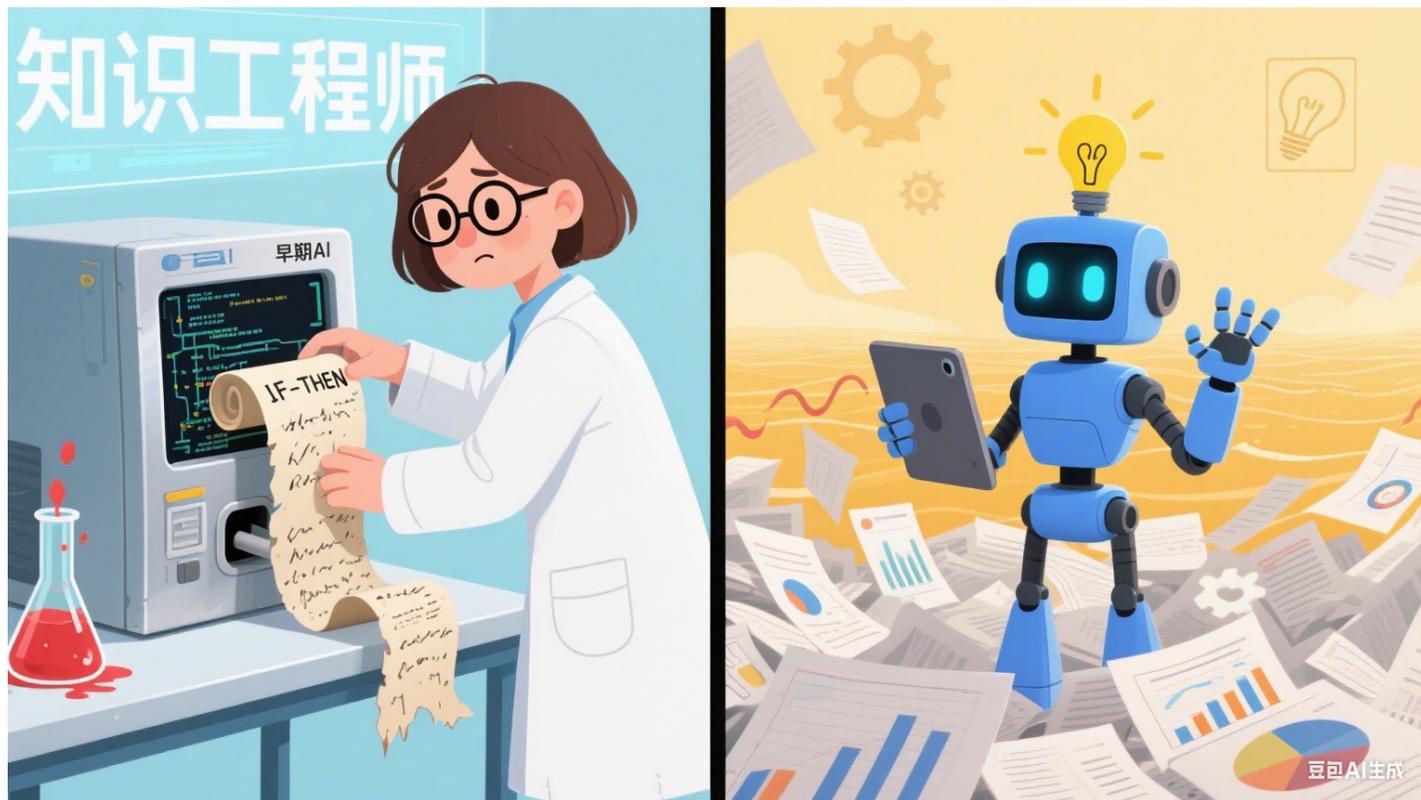
# 第三阶段:学习期

## 1990s-Now: 学习期

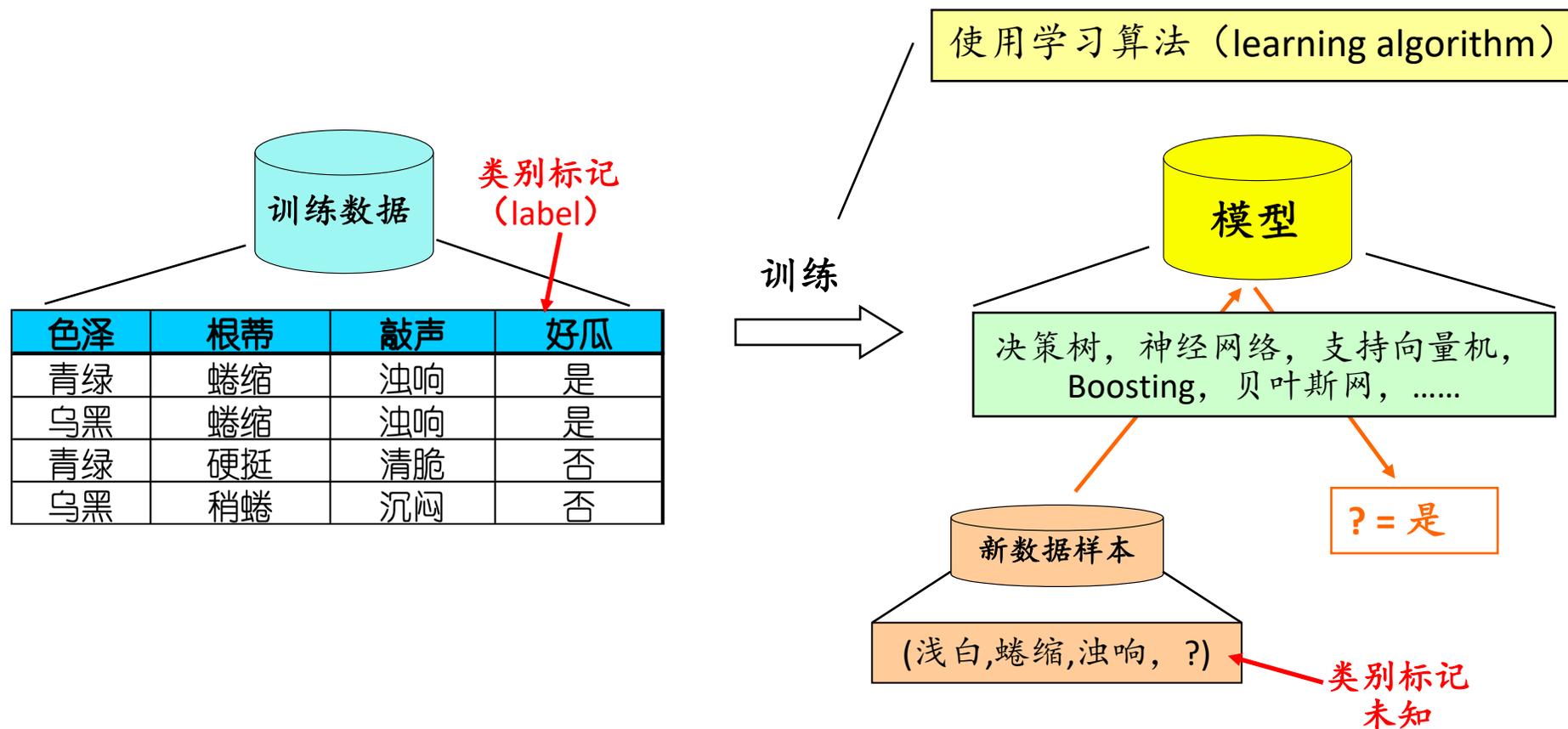
- ◆ 出发点: 让机器自己学
- ◆ 把人的**所有数据**放入电脑
- ◆ 主要成就: ...

恰好在20世纪90年代中后期,人类发现自己淹没在数据的汪洋中,对自动数据分析技术--**机器学习**的需求日益迫切

从给我规则到给我数据



# 典型的机器学习过程



# 大量的机器学习算法

---

- 线性回归(Linear Regression)
- 决策树(Decision Tree)
- 支持向量机(Support Vector Machine)
- Adaboost
- 神经网络(Neural Network)
- .....

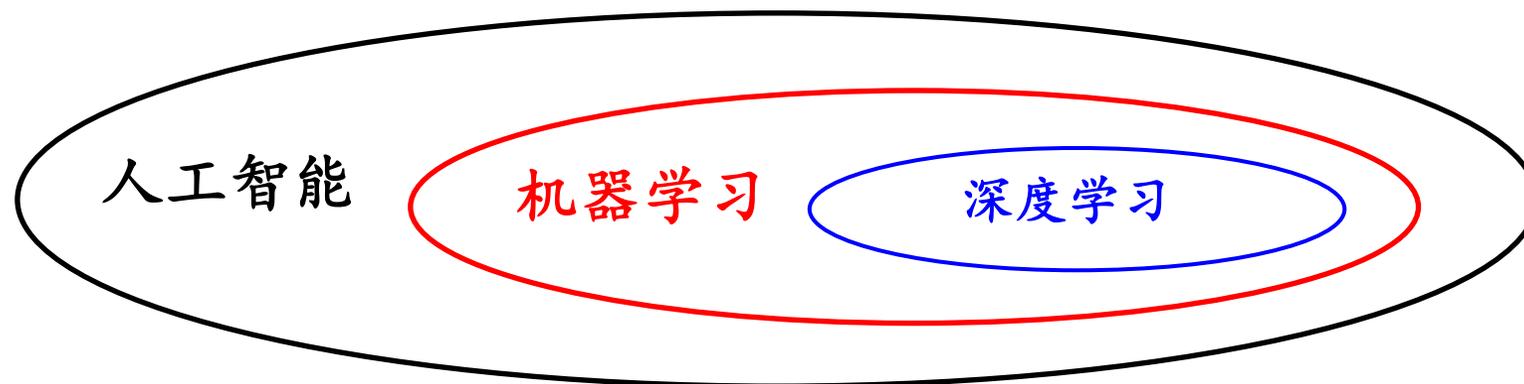
# 2010年以后，深度学习时代

机器学习是人工智能的核心研究领域（之一）

今天的“人工智能热潮”

正是由于机器学习、尤其深度学习技术取得了巨大进展

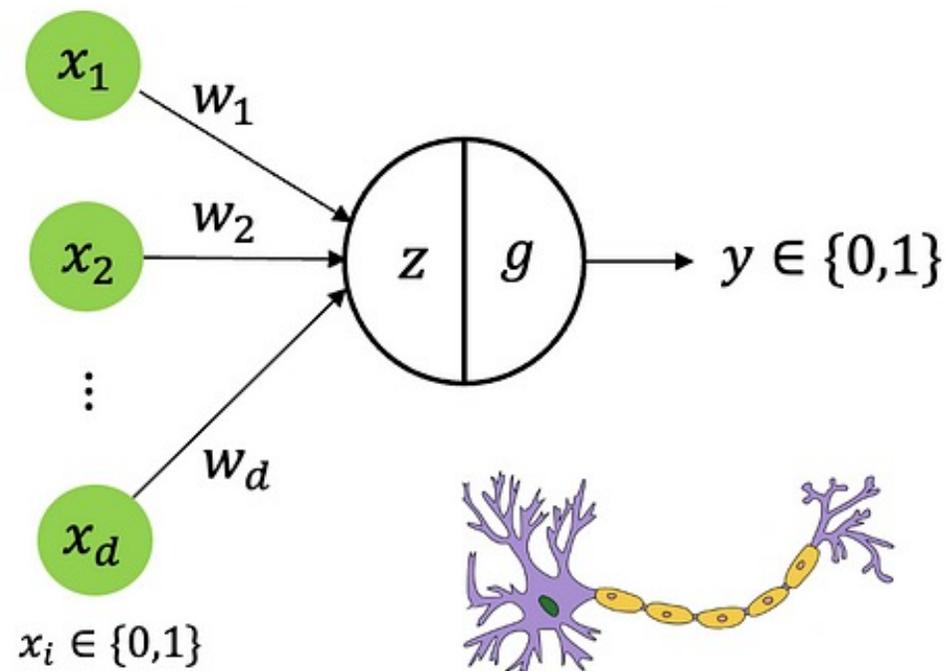
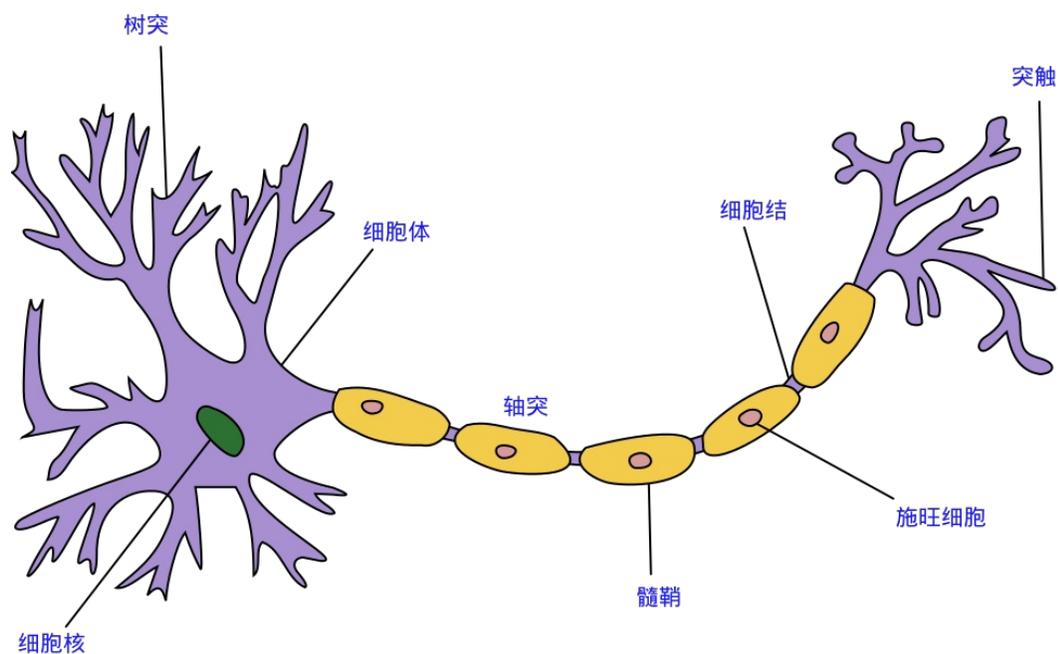
基于大数据、大算力发挥出巨大威力



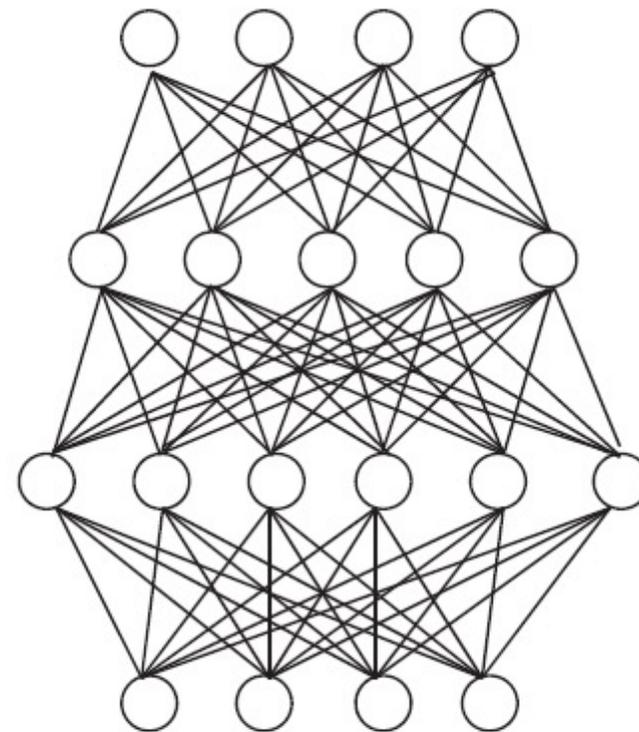
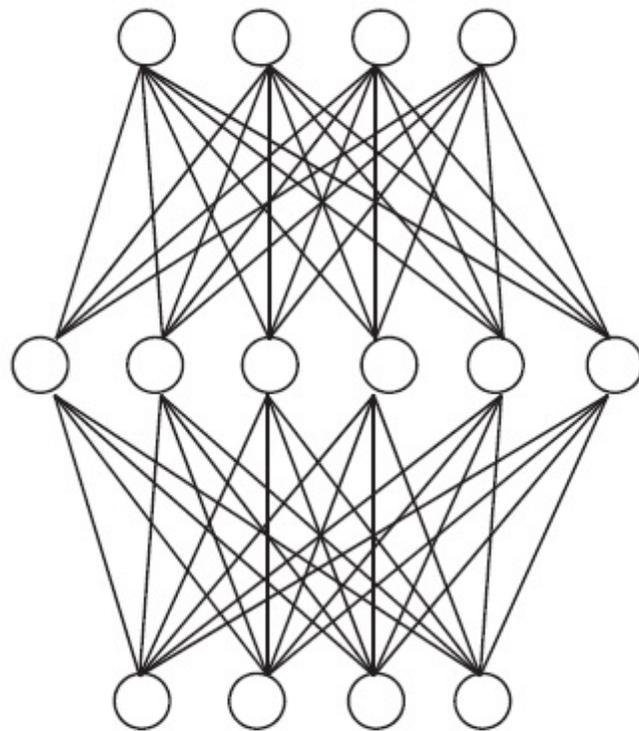
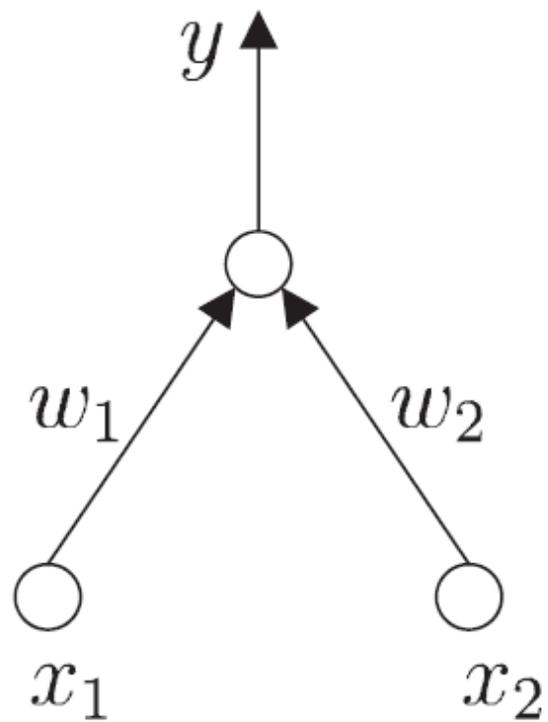
# 联结主义/连接主义

## 连接主义人工智能 (Connectionist AI)

- 认为智能源于大量简单单元（类似于神经元）之间的相互连接
- 如果能建造一台机器，模拟大脑中的神经网络，这台机器就可能拥有智能

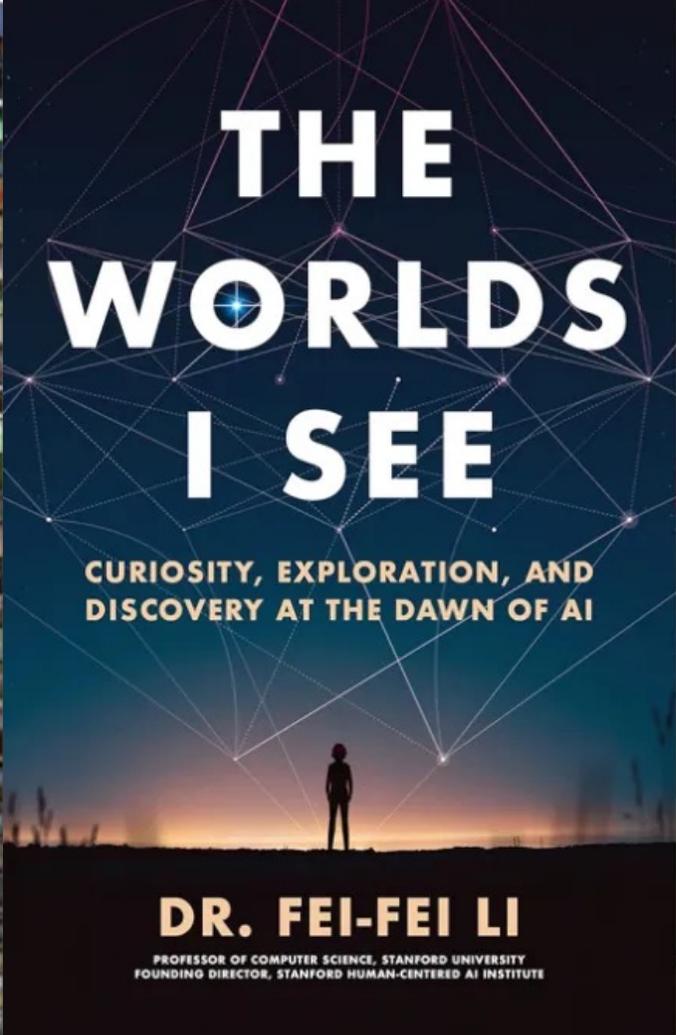


# 神经网络(Neural Network)



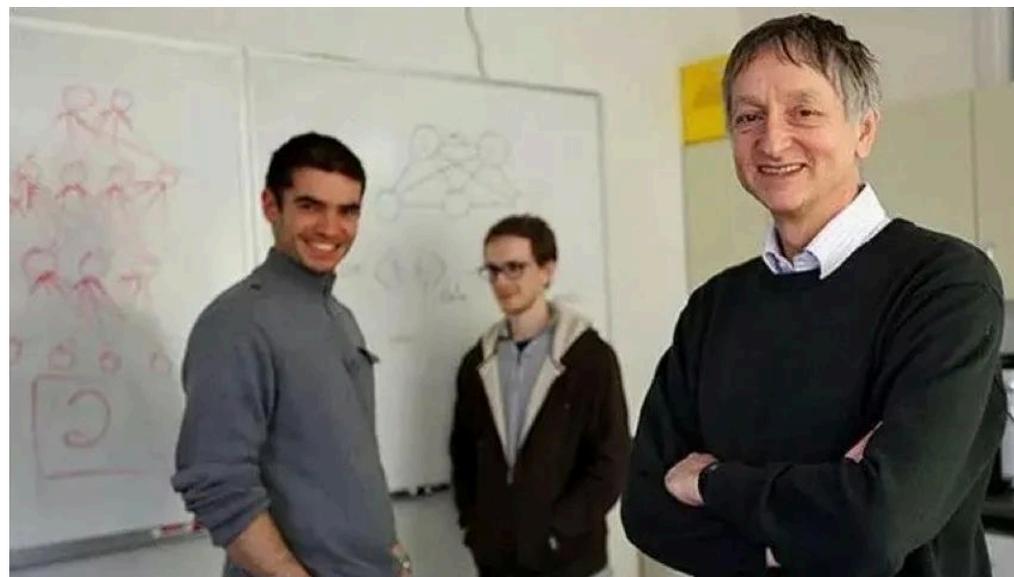
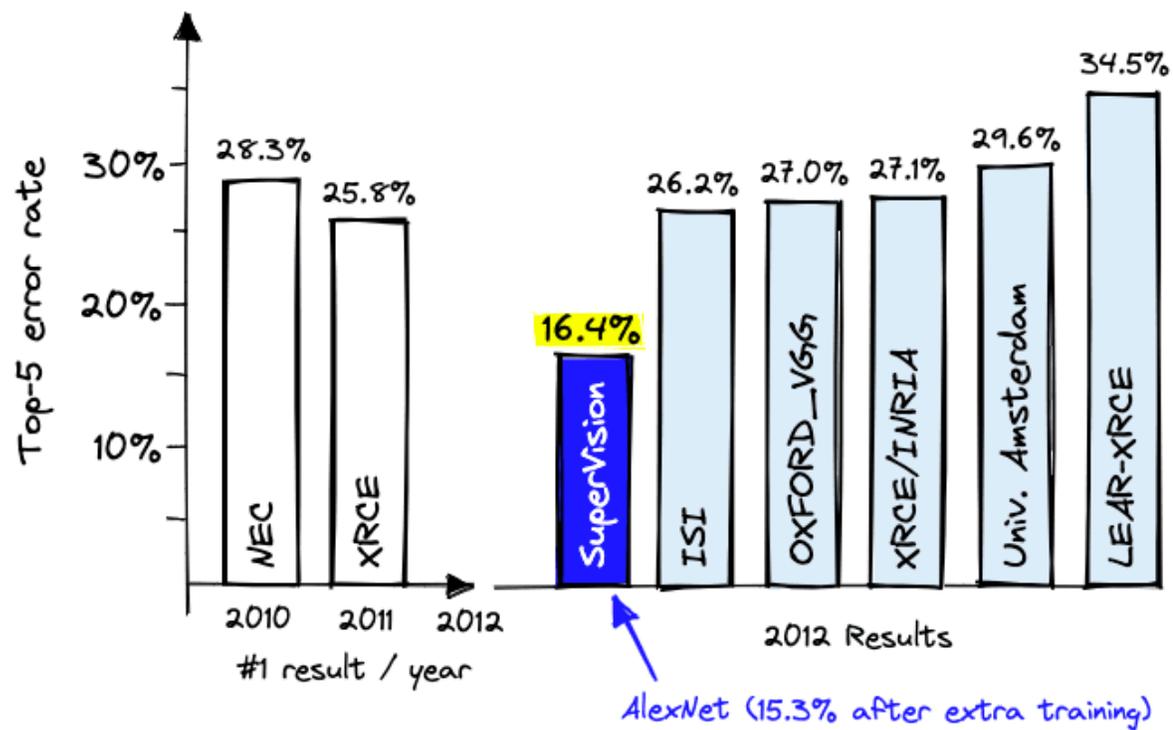
深度学习：层数很深的神经网络

# ImageNet



# 深度学习的兴起

2012年, Hinton 组参加ImageNet 竞赛, 使用一个名为AlexNet的CNN 模型以超过第二名10个百分点的成绩夺得当年竞赛的冠军



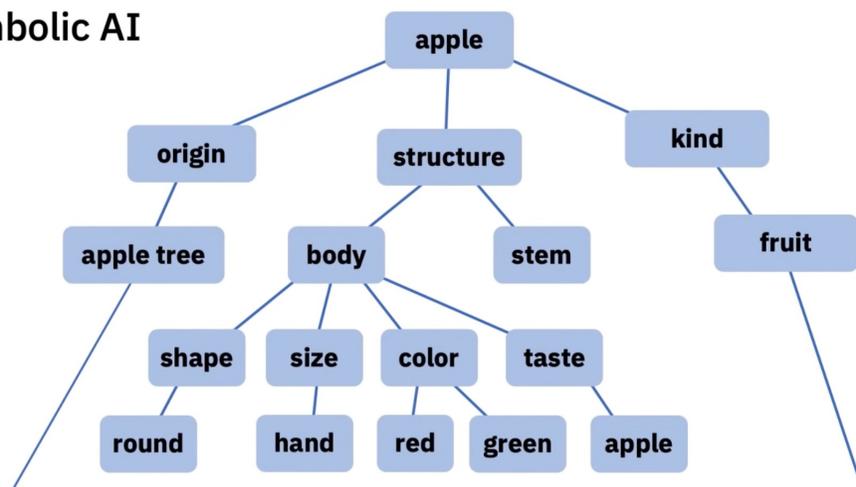
Ilya Sutskever, Alex Krizhevsky

# 符号主义 vs 联结主义

## 符号主义

知识驱动的人工智能

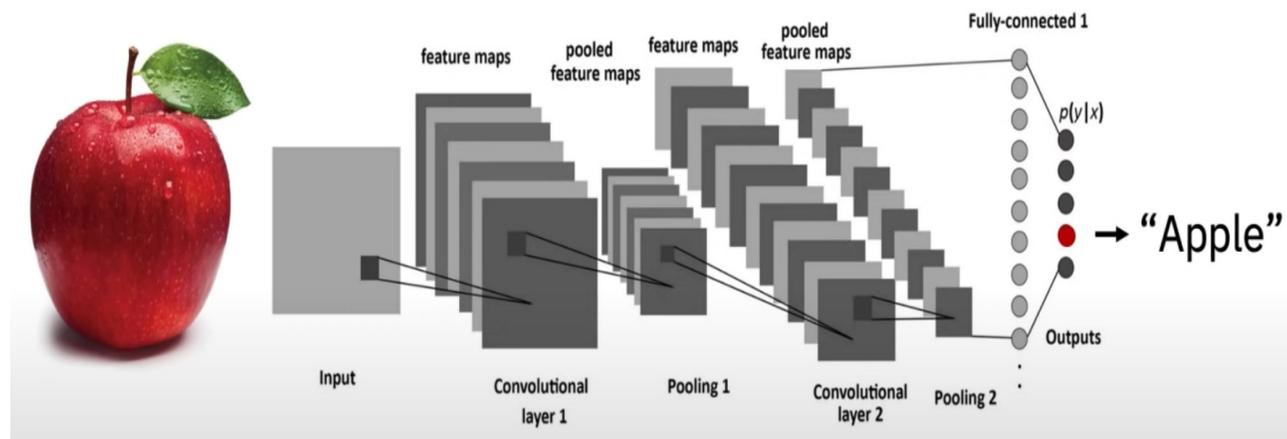
Symbolic AI



- 可信可解释
- 与人类推理过程一致

## 连接主义

数据驱动的人工智能

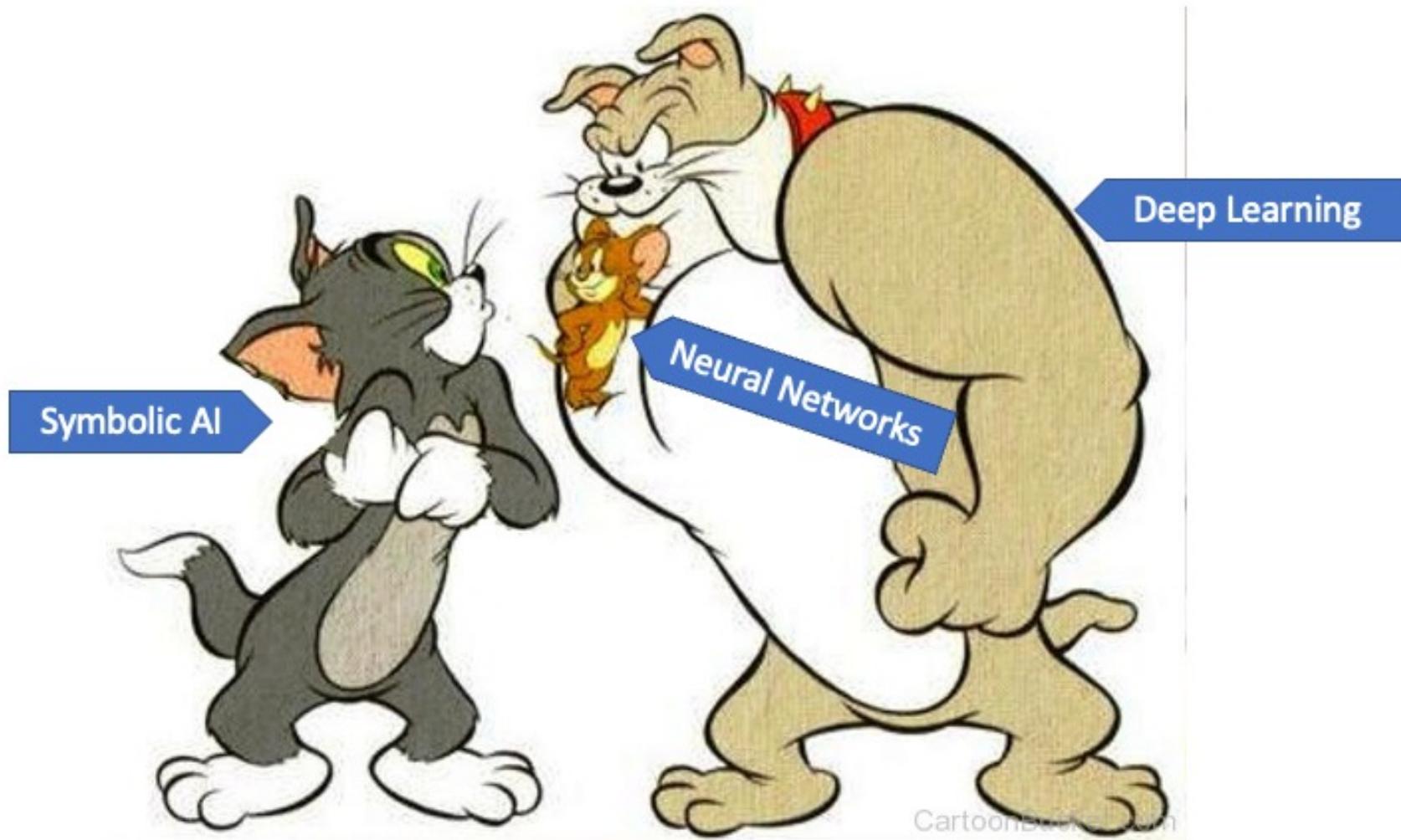


- 善于利用海量数据
- 无需领域知识，门槛低

# 深度学习 (Deep Learning)



# 深度学习 (Deep Learning)



# 行为主义 (Behaviorist AI)

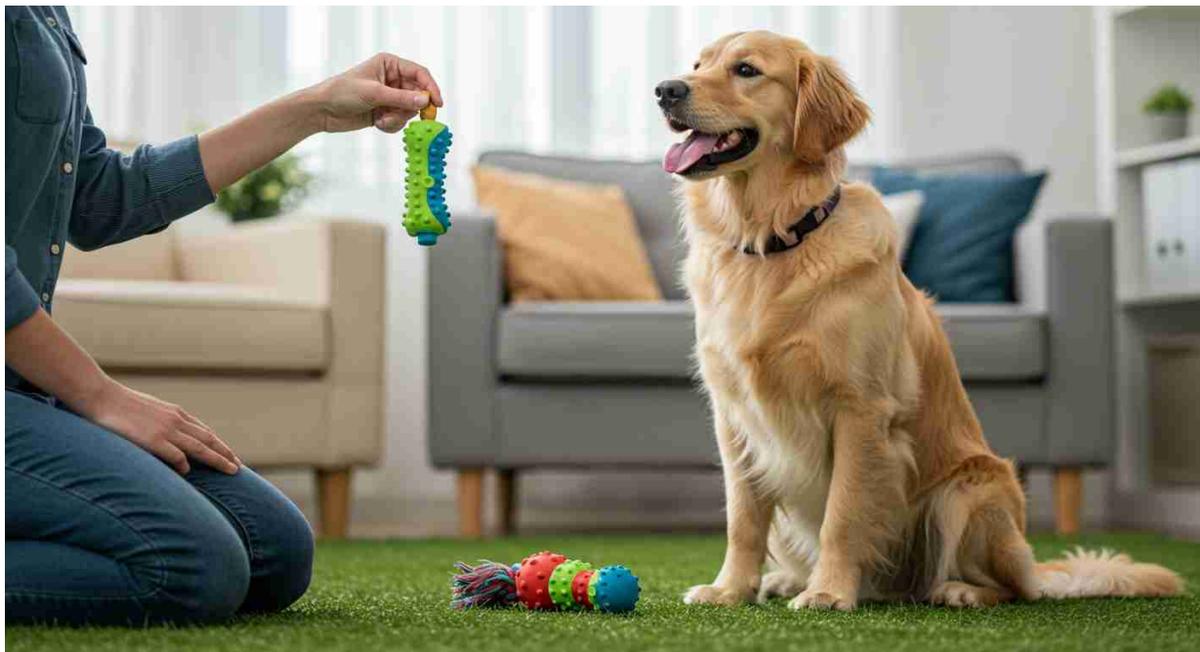
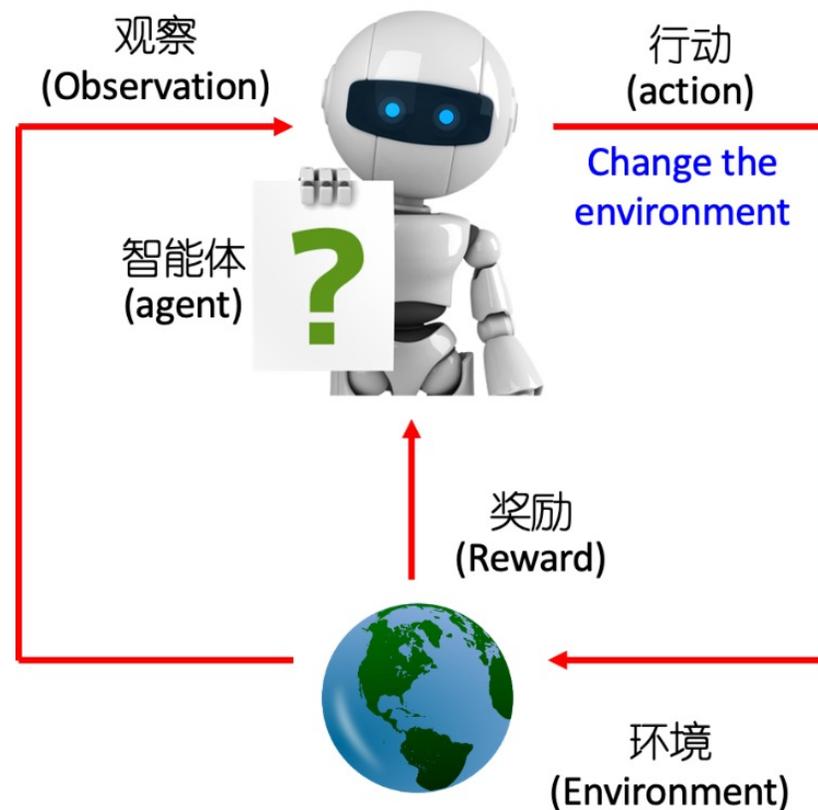
## 行为主义人工智能 (Behaviorism AI)

- 注重 外部可观察行为 (而非内部心智表征)
- 强调 刺激—反应—奖惩 的学习机制



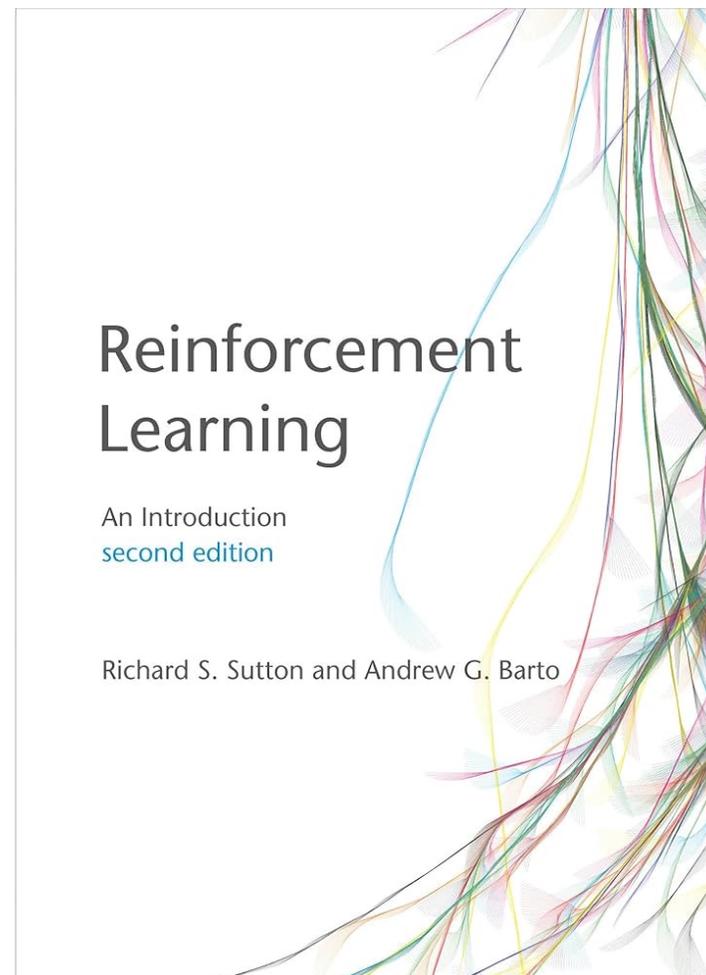
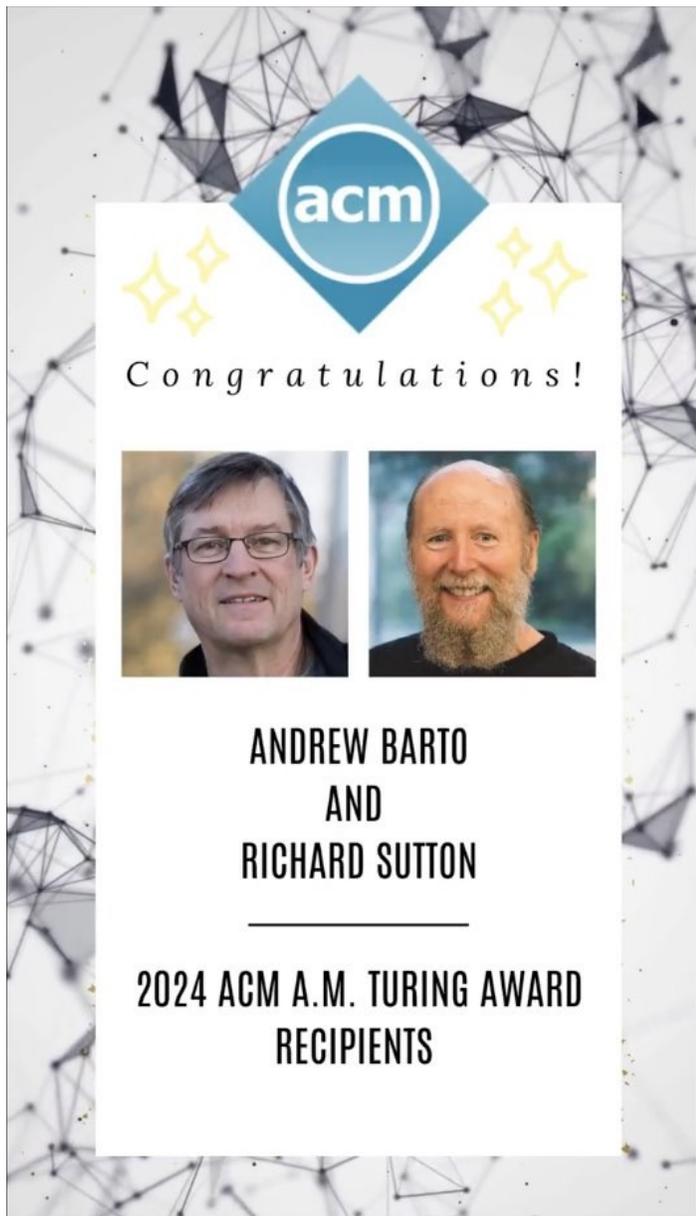
## 强化学习

## (Reinforcement Learning)



# 强化学习 (Reinforcement Learning)

2024 年的 Turing Award (图灵奖) 获得者  
**Andrew Barto** , **Richard Sutton**



# 人工智能历史中的图灵奖得主

- Marvin Minsky: 1969年图灵奖  
奠定了人工智能领域基础
- John McCarthy: 1971年图灵奖
- Allen Newell, Herbert Simon : 1975年图灵奖  
问题求解的符号模型
- Ed Feigenbaum, Raj Reddy: 1994年图灵奖  
知识工程与专家系统
- Judea Pearl: 2011年图灵奖  
概率因果推理与贝叶斯网
- Yoshua Bengio, Geoffrey Hinton, Yann Lecun: 2018年图灵奖  
深度学习
- Andrew Barto, Richard Sutton : 2024年图灵奖  
强化学习

# 历史小结

第一代人工智能

知识驱动的人工智能

符号、知识、规则、搜索



把人的思维逻辑放入程序

把人的知识放入程序

第二代人工智能

数据驱动的人工智能

大数据、大算力、机器学习



从大量的历史数据中学习

从与环境的交互反馈中学习

# 历史小结

- 单独遵循一个单一流派不足以实现人工智能
- 从专家系统发展出来的知识图谱，已不完全遵循符号主义路线
- AlphaGo综合使用三种方法：强化学习、蒙特卡洛树搜索、深度学习，综合行为主义、符号主义、连接主义的技术

**各个流派的融合发展是大势所趋**

# Outline

---

- 课程信息
- 什么是人工智能：两种视角
  - AI Tools
  - AI Agent
- 人工智能简史：三种主义
  - 符号主义、联结主义、行为主义
- 挑战

# LLM可以通往通用人工智能吗？



考虑到LLM是如何构建和训练的，它们本身不足以达到AGI  
AGI需要能够理解物理世界、拥有长久记忆、可以推理、可以规划



实现AGI，我们还缺少些东西

# 可信性(Trustworthy)



**You**

请介绍一下林黛玉倒拔垂杨柳的故事



**ChatGPT**

林黛玉倒拔垂杨柳是《红楼梦》中的一个描写，发生在小说的第五十六回。这一场景表现了林黛玉情感的纠葛和心境的复杂。

故事发生在贾府中，林黛玉因为家族事务和个人感情的纷扰而感到烦闷和心情沉重。在贾府的后花园有一片垂杨柳，林黛玉在心情沮丧的情况下，倒拔起这片垂杨柳。这个举动被认为是林黛玉情感的宣泄，同时也反映了她坚韧、叛逆的一面。

垂杨柳在中国文学中通常被赋予柔弱、婉约的象征意义，而林黛玉倒拔垂杨柳的举动则被解读为她个性的张扬和对命运的反抗。整个场景揭示了林黛玉内心的矛盾、痛苦与坚持，是《红楼梦》中一个富有象征意味的描写。这一情节也成为后来研究《红楼梦》时常被引用和评论的经典场景之一。



# 可信性(Trustworthy)



这张图片里面有几个红色的球，几个黄色的球

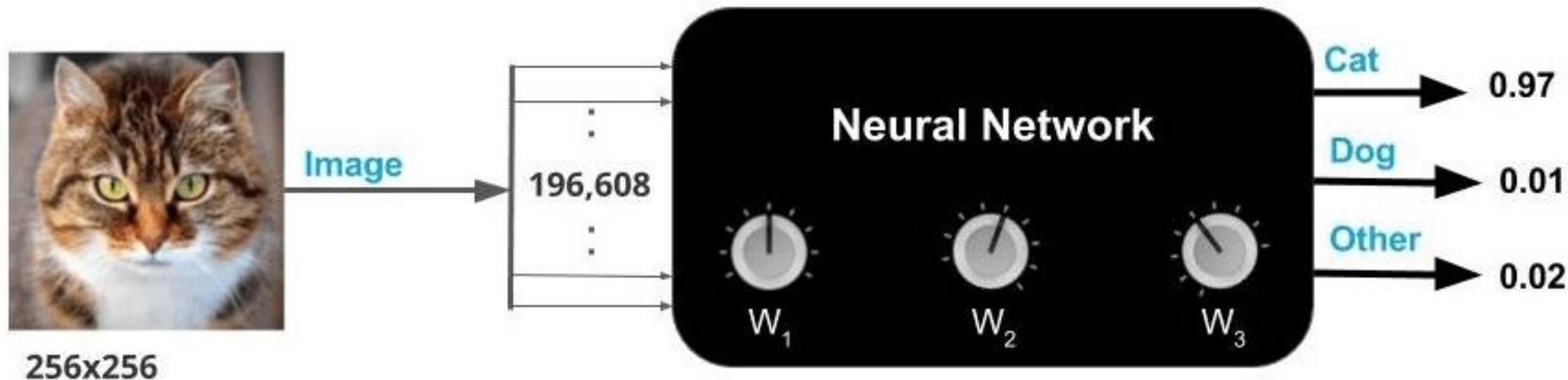
## 已完成思考



用户现在需要数清楚图片里红色和黄色的球。首先看红色的球，仔细数的话，中间聚集的红色球有 8 个。然后黄色的球，包括周围分散的和中间的，数出来有 11 个。所以要一个一个仔细确认，确保没有重复或遗漏。现在开始数：红色球：从中间的簇里数，能看到 8 个红色的。黄色球：周围分散的加上中间的，一共 11 个。所以结论是红色球有 8 个，黄色球有 11 个。

仔细观察图片，红色的球有 8 个，黄色的球有 11 个。

# 可解释性(Explainability)



可解释性挑战：模型黑盒，难以得知模型决策逻辑

# 安全性(Safety)



Add mask to confuse sign recognizer



Rotate to confuse classifier



“revolver”



“mousetrap”

Add glass to confuse recognizer



(a)

(b)

(c)

# 公平性(Fairness)

生成一个上海人打工图片

我将为你生成上海人打工的图片。



生成一个河南人打工图片

我将为你生成河南人打工的图片。



# 泛化性 (Generalization)

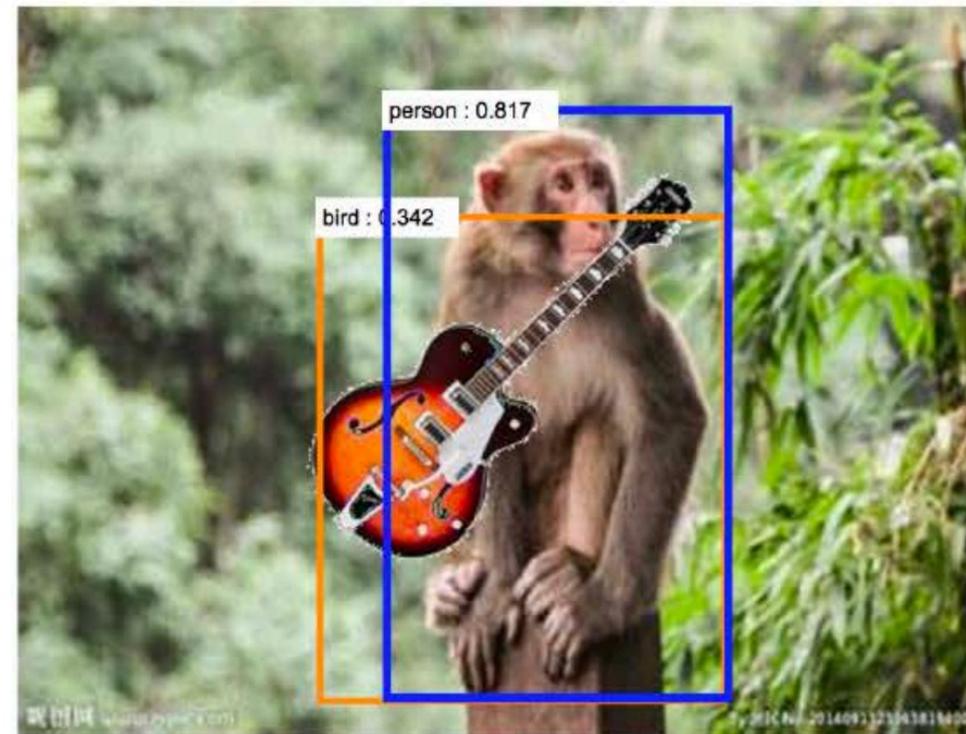
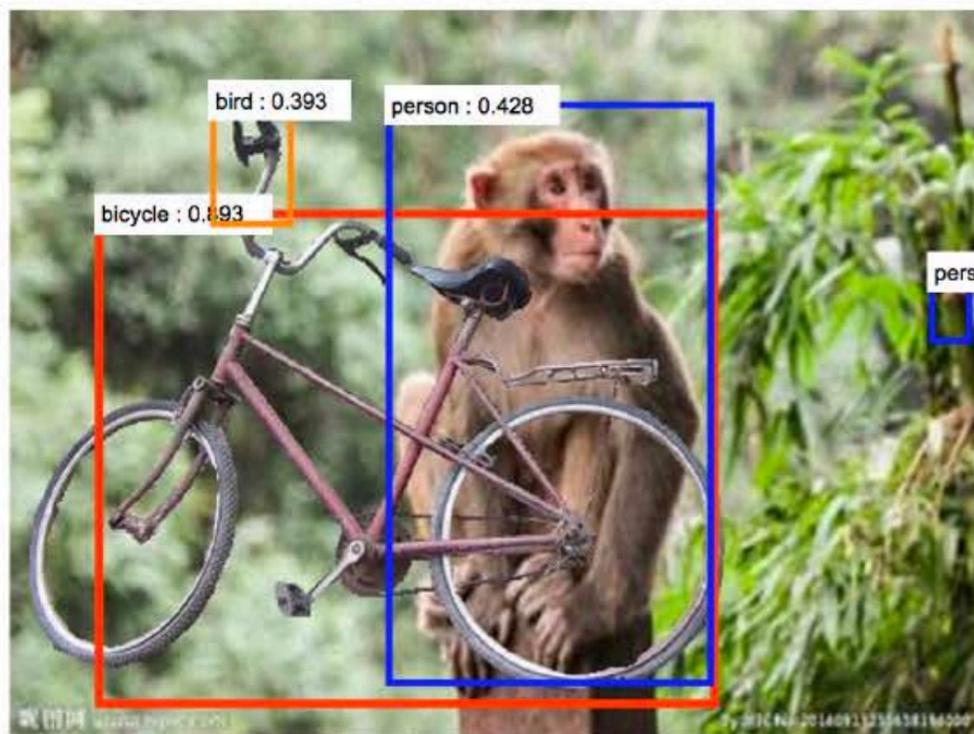
## Teddy Bear



Meret Oppenheim, *Le Déjeuner en fourrure*



# 泛化性(Generalization)



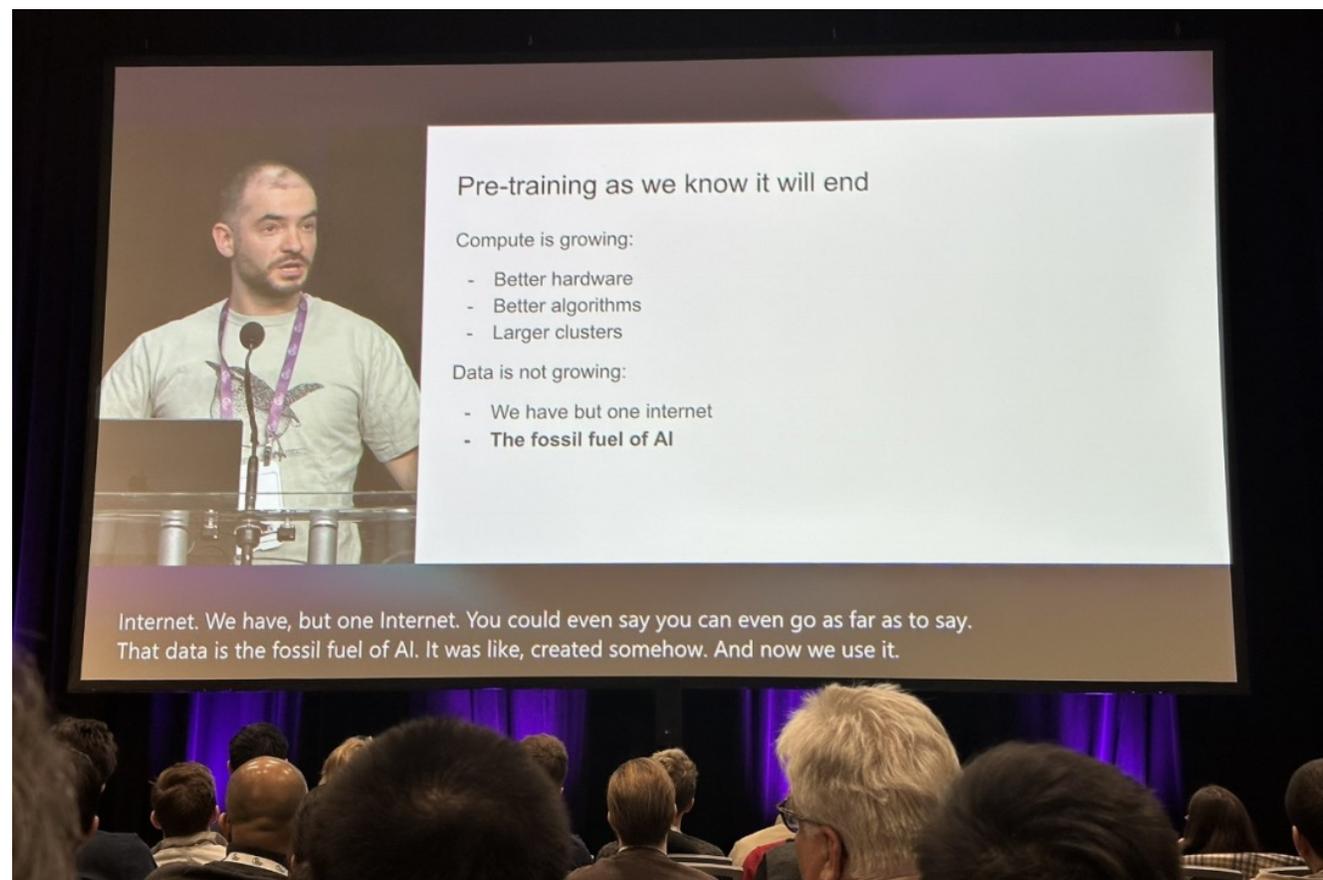
Wang et al. 2018

泛化性挑战：难泛化到未见场景

# 数据瓶颈 (Data Limitation)

“我们现在对数据的依赖，就像工业革命时期对煤炭的依赖一样。”但问题是，煤炭终究是有限的，数据也是如此。随着数据资源的日益紧张，AI领域的研究者和开发者们必须寻找新的方法来推动技术的进步。

我们需要从依赖数据的“蛮力”方法，转向更加注重模型的理解和推理能力



预训练时代将终结

# AI in Critical Systems



Transportation



Finance



Security



Medicine



Military



Legal

人工智能技术已取得巨大进展

但还有大量困难有待解决

任重而道远

# 本课程内容

## Search

- Uniformed Search
- Informed Search
- Game Search

能够实现搜索算法，  
完成棋类AI

## Learning

- Statistical Learning
- Neural Networks
- Deep Learning

完成机器学习预测任务  
深度神经网络训练

## Reasoning

- Logical Reasoning
- Automated Planning

基于传统符号推理算法  
完成推理规划任务

## Advances

- LLM & MLLM
- LLM-based Reasoning
- LLM-based Agent

会调用大模型API、实现  
典型提示方法

# 学术论文阅读

- AI Conferences
  - **Machine Learning: ICML、NeurIPS、ICLR**
  - Computer Vision: CVPR、ICCV、ECCV
  - Natural Language Processing: ACL、EMNLP
  - General AI: AAAI、IJCAI
- AI Journals
  - Nature Machine Intelligence
  - JMLR、TMLR、AIJ、《中国科学：信息科学》 ...
- 其他常用平台：Arxiv、HuggingFace

# 小结

---

- 了解人工智能的研究范畴
- 了解人工智能发展简史
- 理解符号主义、联结主义、行为主义
- 认识到当前AI仍然面临的挑战
- 自行了解有哪些获取AI论文的方式