

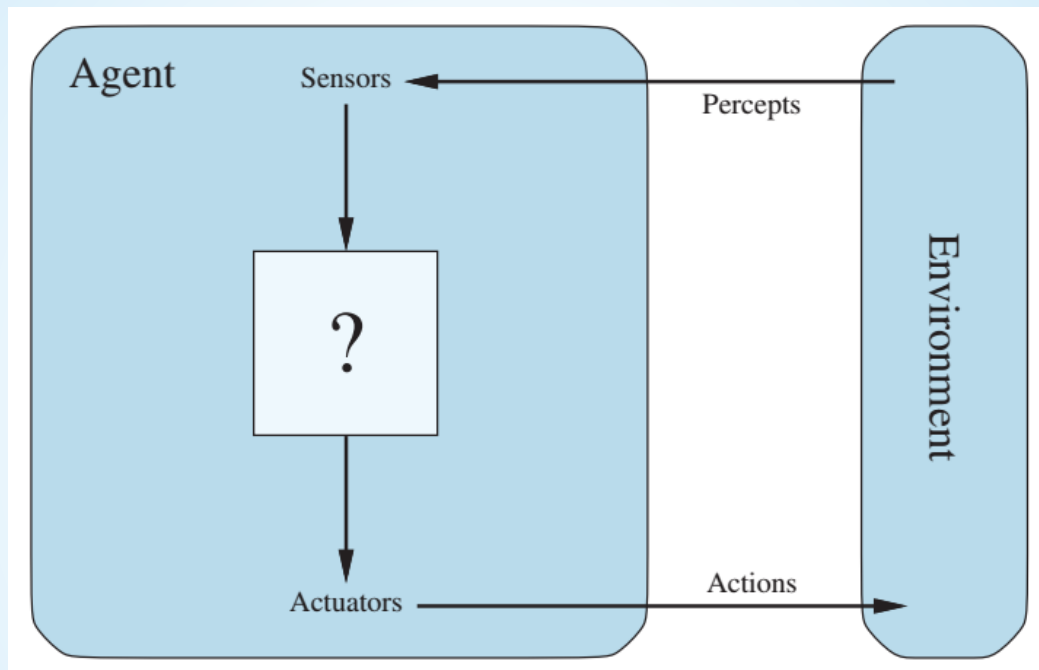
# 第15讲 AI应用和伦理



**15.1 Agent基本概念和AI体系复习**

**15.2 应用: ChatGPT**

**15.3 AI伦理**



自主智能体 (Agent) 是一个理论抽象,  
机器人、无人车等大量的自主智能装置与系统是其物化形态

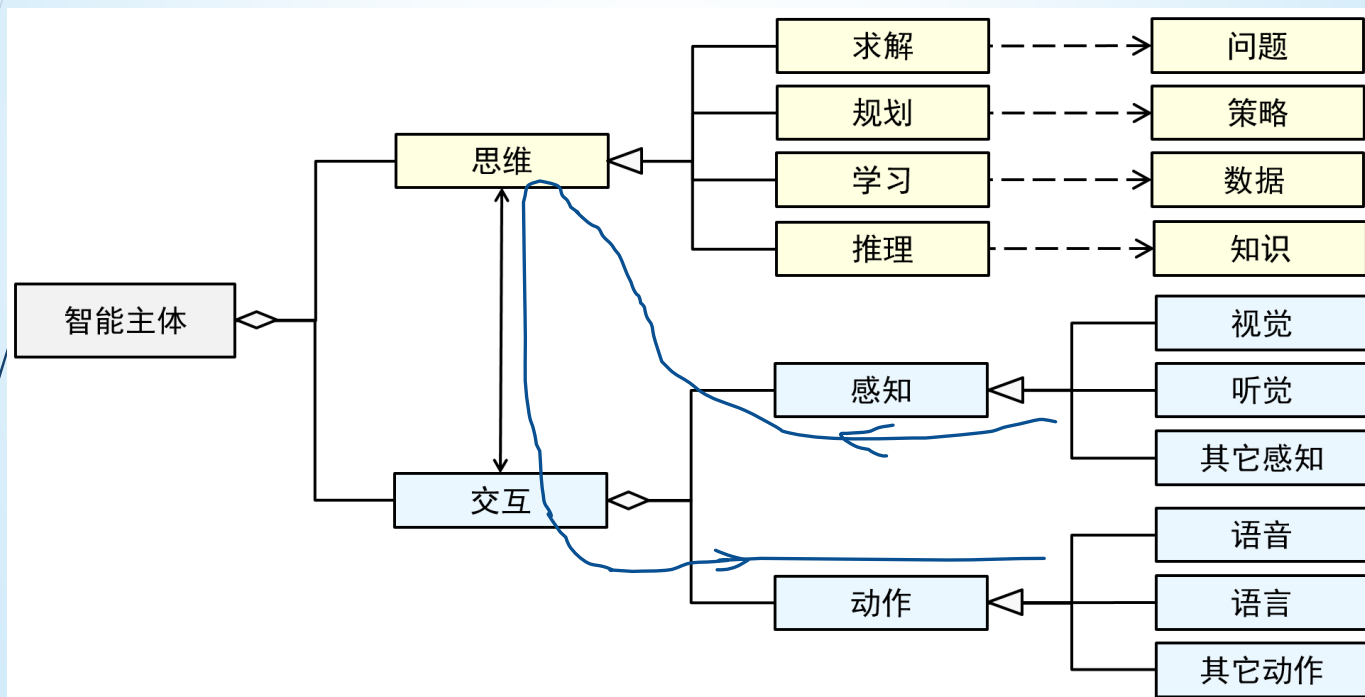
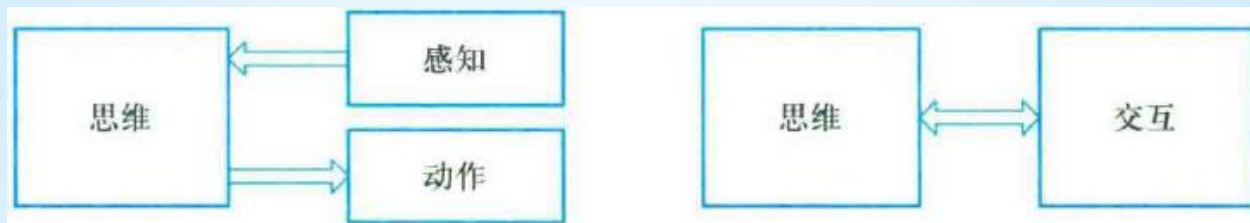


Agent

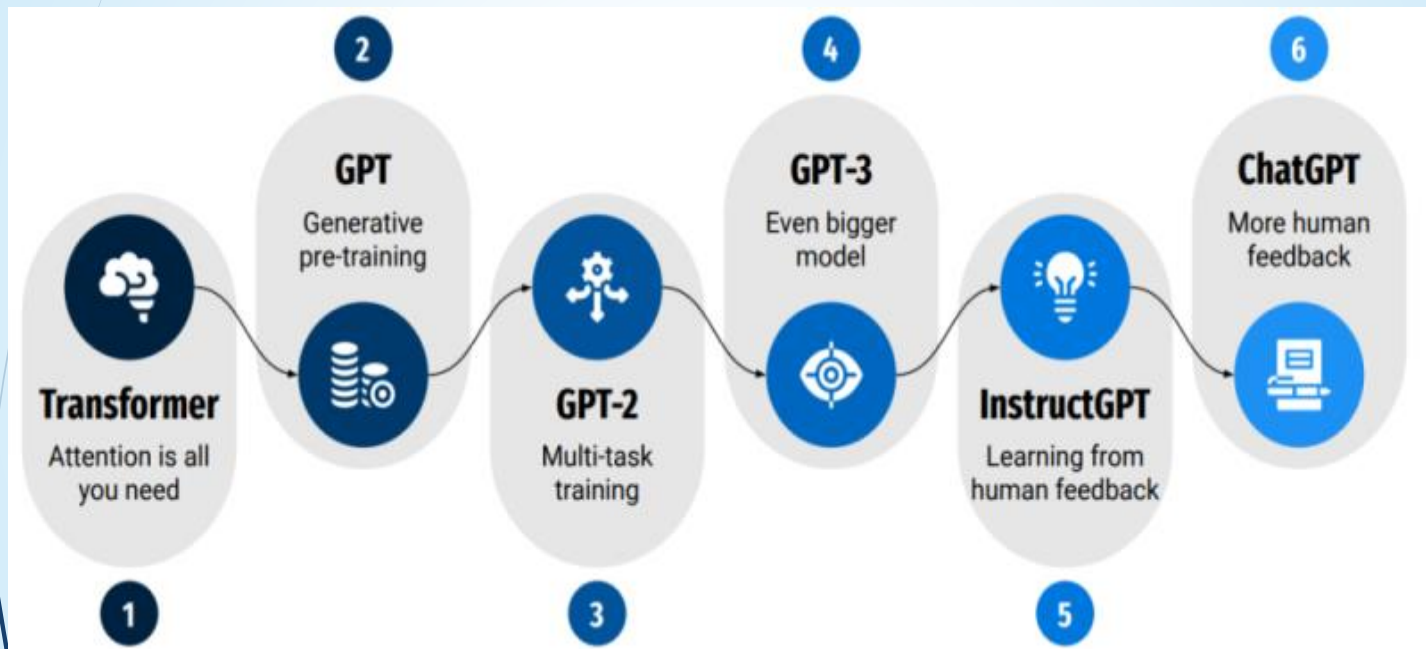


- 人
- 机器人
- 无人车/智能车
- 无人机
- 智能传感器/智能控制单元
- 智能加工单元
- 自主交易系统
- .....

# 15.1 AI体系 - (Review) 人工智能的体系



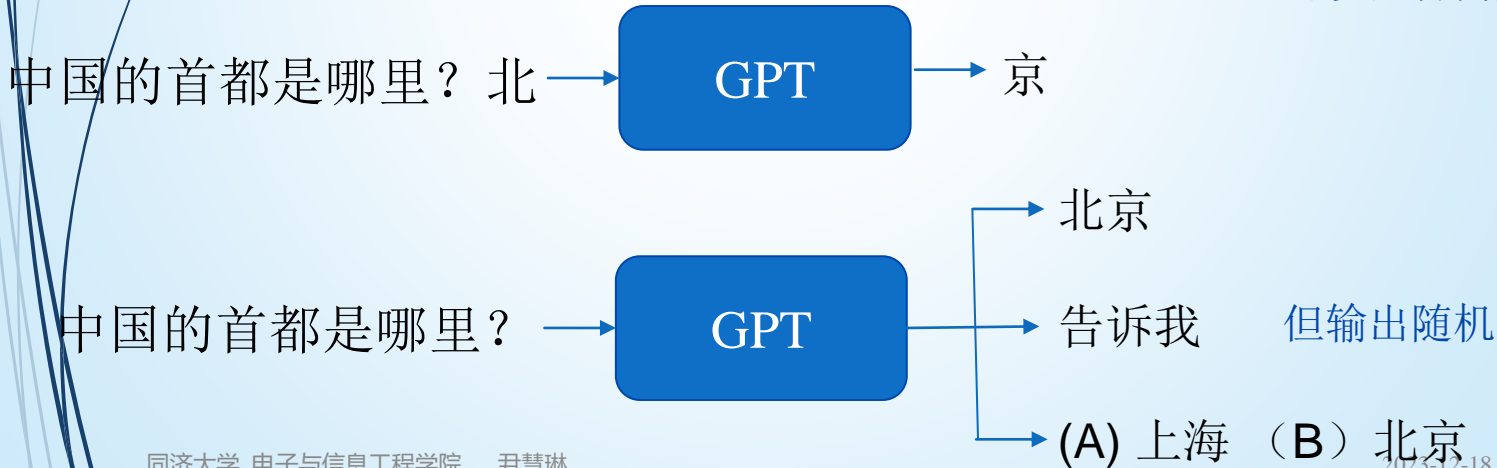
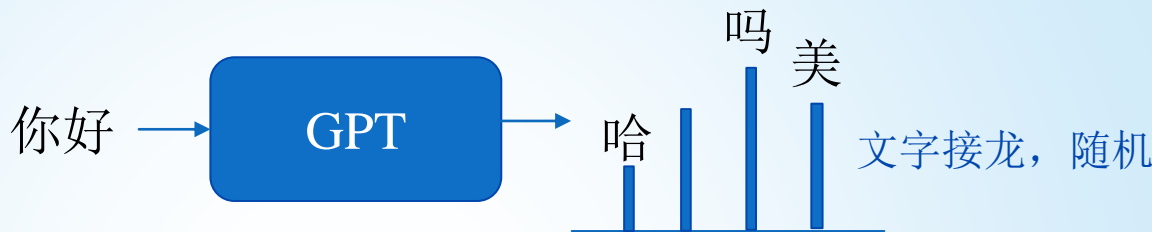
## 15.2 AI应用 - ChatGPT的发展历程



## 15.2 AI应用- ChatGPT的发展历程



# 1. 无监督学习







## 2. 监督学习

中国的首都是哪里？  → 北京

## 3. 模仿人类的喜好

中国的首都是哪里？ →  GPT  → 北京  
✓  
谁回答我？  
✓  
.....

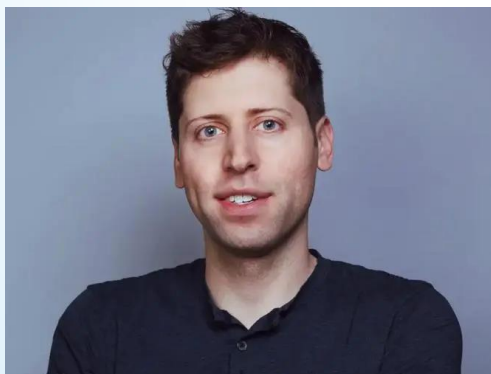
中国的首都是哪里？ 北京 →  奖赏模型 → 高分

中国的首都是哪里？ 谁回答我？ →  奖赏模型 → 低分

## 4. 强化学习



- 数据：恶意数据、偏见、误导数据
- 模型：不可解释性、知识涌现、测试难以穷举；信息泄露
- 用户：黑客、误导APP、造假诈骗



阿尔特曼：“我特别担心这些模型可能会被用于大规模的虚假信息，现在他们在编写计算机代码方面做得越来越好，可以用于进攻性网络攻击。”



- 微软推出生成式AI 安全产品Microsoft Security Copilot, 帮助网络安全专家了解关键问题并找到解决方案。
  - 阿尔特曼呼吁 “监管机构和社会需要进一步参与这项技术, 以防止对人类造成潜在的负面影响。”
  - OPEN AI在其官网发布《Our approach to AI safety》, 就构建安全、可靠的AI 产品, 尊重用户隐私, 保护儿童及提高生成数据准确性等多方面制定政策。
  - 马斯克等联名叫停GPT训练, 强调潜在伦理风险
- “只有当我们确信它们的影响是积极的, 并且它们的风险是可控的时候, 才能开发更强大的人工智能系统。”



马克.扎克伯格  
Facebook 创始人

“人类制造机器就是为了  
让机器在某些方面强  
于人类，但是机器在  
某些方面超越人类不  
意味着机器有能力学  
习其他方面的能力，  
或者将不同的信息联  
系起来而做超越人类  
的事情，而这一点非  
常重要”。

**VS**



埃隆.马斯克  
SpaceX 太空探  
索技术公司CEO

只要你认可AI技  
术会不断发展，  
我们会在智力上  
远远落后于AI，  
以至于最终成为  
AI的宠物。



- 社会问题
- 安全问题
- 法律问题
- 道德问题

《人类简史》系列三部曲作者尤瓦尔·赫拉里：

- 19世纪工业革命创造了城市工人阶级
- 21世纪人工智能革命将可能创造 “无用阶级”



人工智能在什么情况下会危害人类？需要同时满足三个条件：

1. 有行为能力。Alpha Go 是下棋机器人，不能动，所以不会危害人类；
2. 有足够破坏力。扫地机器人不具有破坏的动能，所以不会危害人类；
3. 具有自主能力。完全听命于人类的系统，不会主动伤害人类，但会误伤人类。

■ 第一，可以动的问题已解决；第二，有破坏力的机器人也存在；第三，关键就是能不能自主。

■ 我们还不能确认机器人不会自我进化到危害人类的程度，所以对它预先要有约束。



➤ 新的阿西莫夫的机器人定律为：

**第零定律：**机器人必须**保护人类的整体利益**不受伤害。

**第一定律：**机器人**不得伤害人类个体**，或者目睹人类个体将遭受危险而袖手不管，除非这违反了机器人学第零定律。

**第二定律：**机器人必须**服从人给予它的命令**，当该命令与第零定律或者第一定律冲突时例外。

**第三定律：**机器人在不违反第零、第一、第二定律的情况下要尽可能保护自己的生存。





**人工智能的法律地位：**是否要赋予人工智能以法律主体地位或有限的法律主体地位？

- 有学者认为：高度自主人工智能具有脱离人类初始算法和规则预设的可能性，这意味着当人工智能发展到一定阶段时将很难受到人类控制，而且也会具备了像人一样**自主创新，解决问题的能力**。**建议赋予人工智能主体资格。**“电子人格说”
- 反对者：目前，人工智能并不具备人类理性，并不具有自我意志和自我行动能力的“（类）人性”判断。虽然人工智能在模仿人类方面取得了重大进展，初步具备了“人的智能”，但是与真正的“人脑”相比，尚存在较大的差距，**是模仿人类的思维逻辑处理问题，作为人的智能的延伸。**“工具说”



- 科幻电影《她》、《机械姬》和《机械管家》
- “跟一个真人谈恋爱，成本还是太大了”
- “定制一个AI伴侣，不仅能够完美地躲过一切矛盾，完美地满足你的一切需求”
- “和机器人谈的恋爱，是没有灵魂的。只是人和工具的关系。”



## 发展方向：

### ■ 可信赖的技术

### ■ 科技向善的初衷

- ✓ Google: “Don’ t be evil”
- ✓ 李飞飞: “人性” 化AI
- ✓ 马云: 技术必须向善
- ✓ IBM 大中华区董事长陈黎明: 科技向善与时代责任

....

The Google logo, consisting of the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red).The phrase "Don't be evil" written in a multi-colored font, similar to the Google logo, on a light background.

## 15.3 AI伦理 – 科技向善



**李开复：李飞飞是人工智能的“良心”**

- 斯坦福“以人为本”人工智能研究院  
Stanford Human-Centered AI Institute (HAI)
- 让智能机器更加以人为本，怀有善意，帮助人类解决一些最有意义难题
- AI 赋能人类，而非取代人类
- 安全，公平和善意