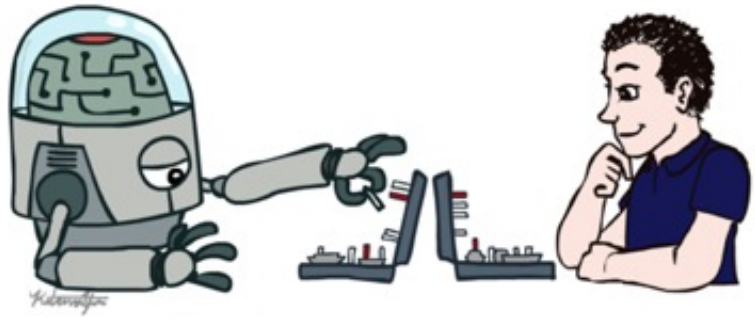


The Ethics of Artificial Intelligence

# 人工智能的伦理问题

中国地质大学（北京）
叶 山
yes@cugb.edu.cn

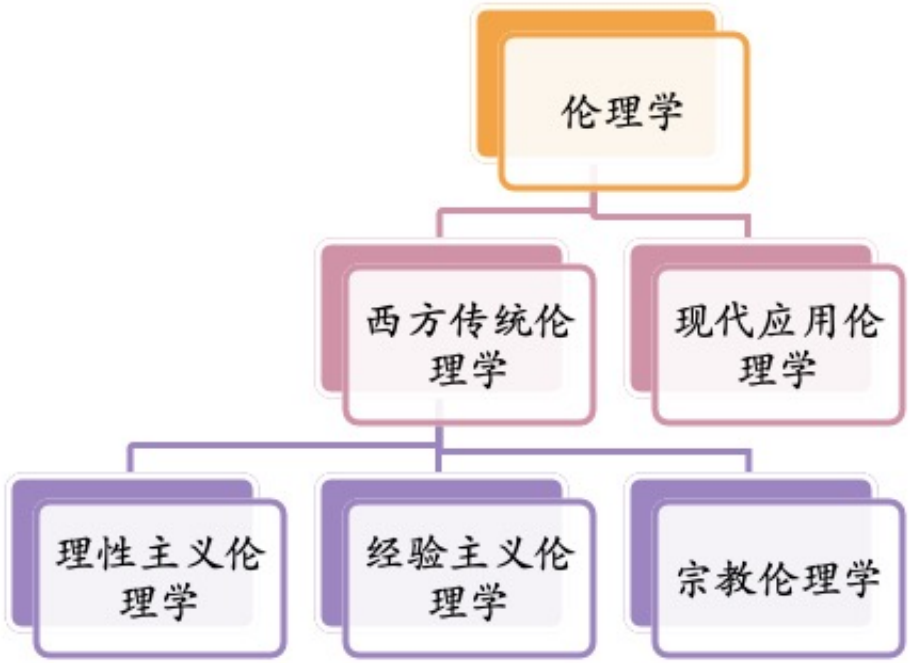
# 人工智能伦理

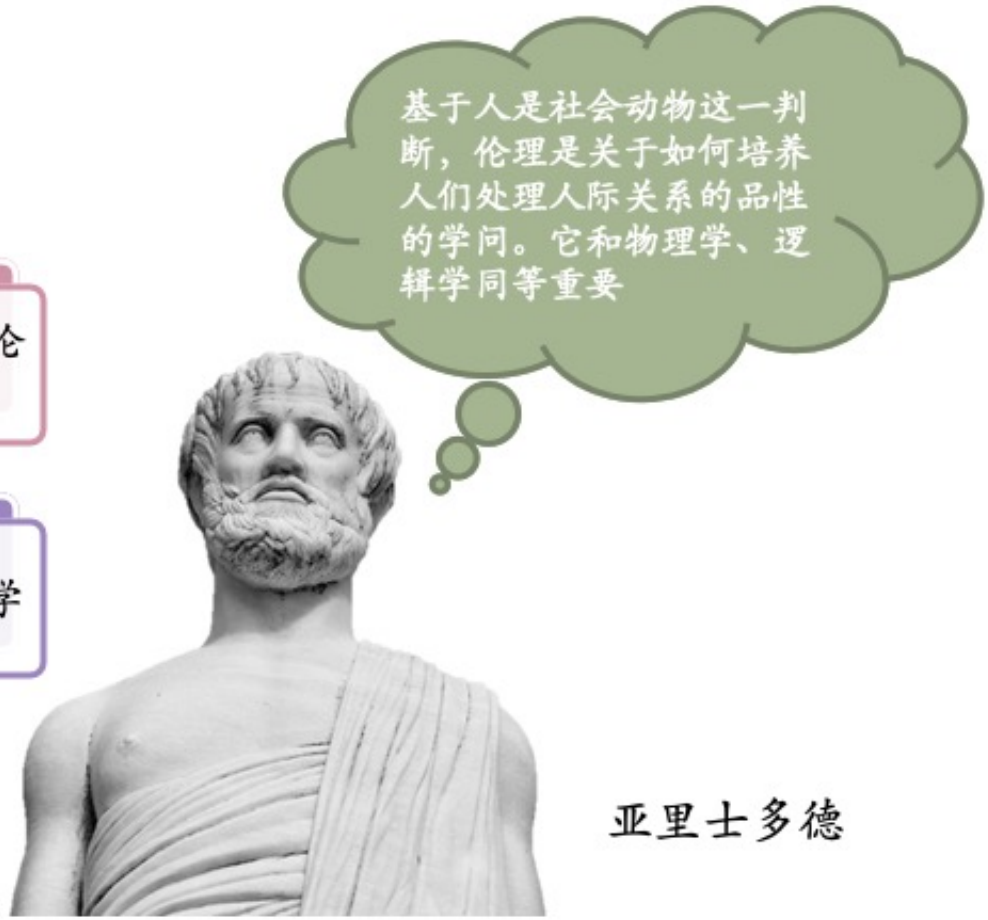
伦理（ethics）：本意是指事物的条理，引申为人伦道德之理，是一种社会公认的非强制性的行为规范（和强制性的法律相对）。

- 狭义的伦理：关注人与人、人与社会、人与自身的道德关系。
- 广义的伦理：涉及人与自然、人与科技的道德关系，以及义务、责任、价值、正义等一系列范畴。
- 人工智能伦理：将伦理推广到人工智能系统中，制定人工智能与人之间的行为规范。

# 人工智能伦理

伦理学：以道德作为研究对象的科学，也是研究人际关系的一般规范或准则的学科。它是关于理由的理论，即做某事或不做某事的理由，认为某个行动、规则、做法、制度、政策和目标好坏的理由。

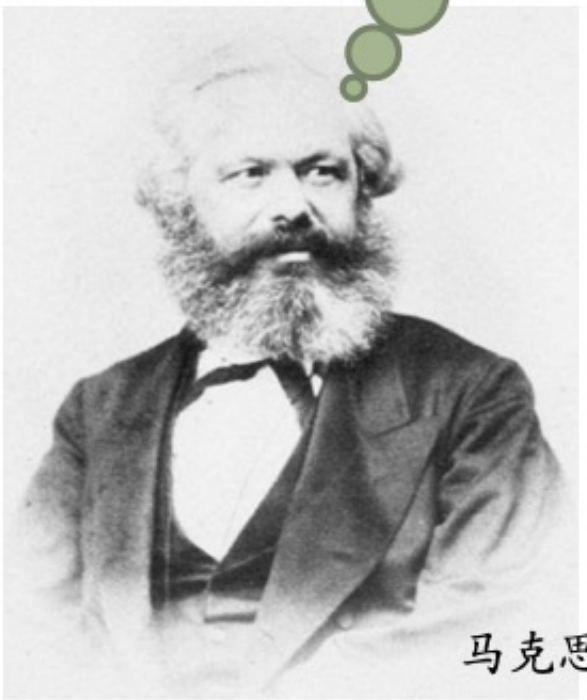



亚里士多德

# 人工智能伦理

每一种事物好像都包含有它自己的反面。我们看到机器具有减少人类劳动和使劳动更有成效的神奇力量，然而却引起了饥饿和过度的疲劳。技术的胜利，似乎是以道德的败坏为代价换来的。



马克思

- 新技术的潜在负面影响
  - 物质享乐思想
  - 道德滑坡
  - 新的犯罪手段
  - 社会公平问题

# 人工智能伦理

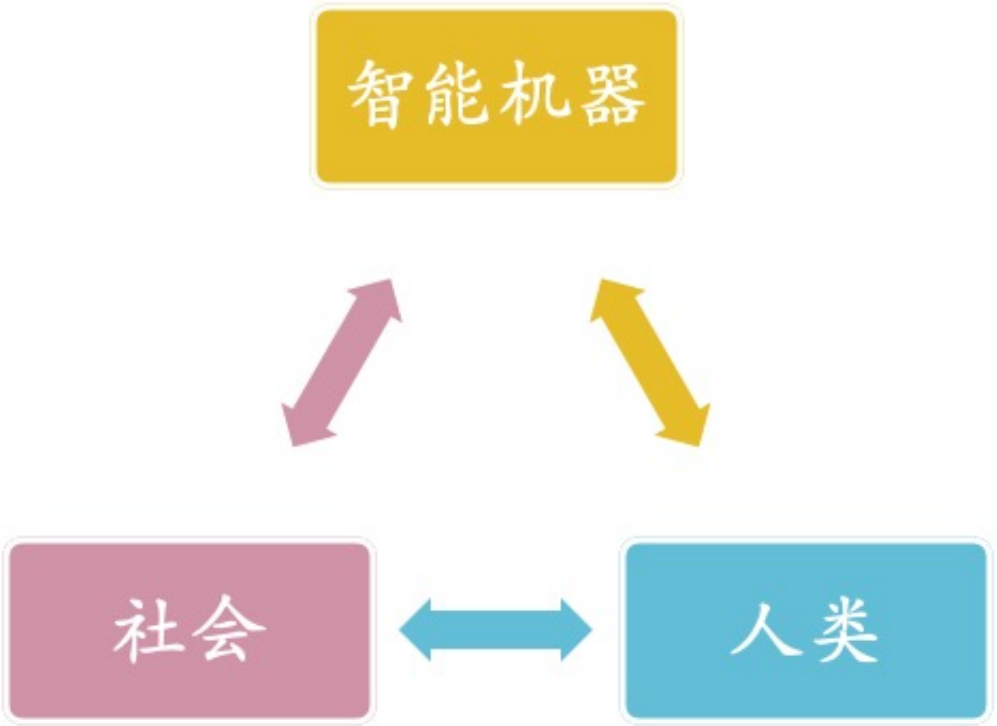
## 狭义人工智能伦理

- 考虑人工智能系统、智能机器及其使用所引发的涉及人个范畴的伦理问题。

## 广义人工智能伦理

- 考虑人与人工智能系统、人与智能机器、人与智能社会之间的伦理问题。
- 超现实的通用人工智能伦理问题，包括人工智能系统与智能机器对于人类的责任、安全等。

# 人工智能伦理学




核心任务：明确和智能机器行为相关的行动、动机、态度、判断、规则和目标的理由。

# 人工智能伦理学

## 狭义人工智能伦理学

- 研究关于人工智能技术、系统与机器及其使用所引发的涉及人类的伦理道德理论的科学。
- 主要关注和讨论关于人工智能技术、系统及智能机器的伦理理论。
- 因人工智能技术而出现的新伦理问题，如数据伦理、算法伦理、机器伦理、机器人伦理、自动驾驶伦理、智能医疗伦理等。

## 广义人工智能伦理学

- 研究人、智能机器与社会之间的新型道德伦理关系。
- 关注智能机器(包括人机结合)的道德规范体系。
- 智能机器道德水平与人工智能技术发展水平之间的关系、智能机器道德原则和道德评价的标准、智能机器道德的教育。
- 在智能机器超越人类的背景下，人生的意义、人的存在价值、生活态度等问题。

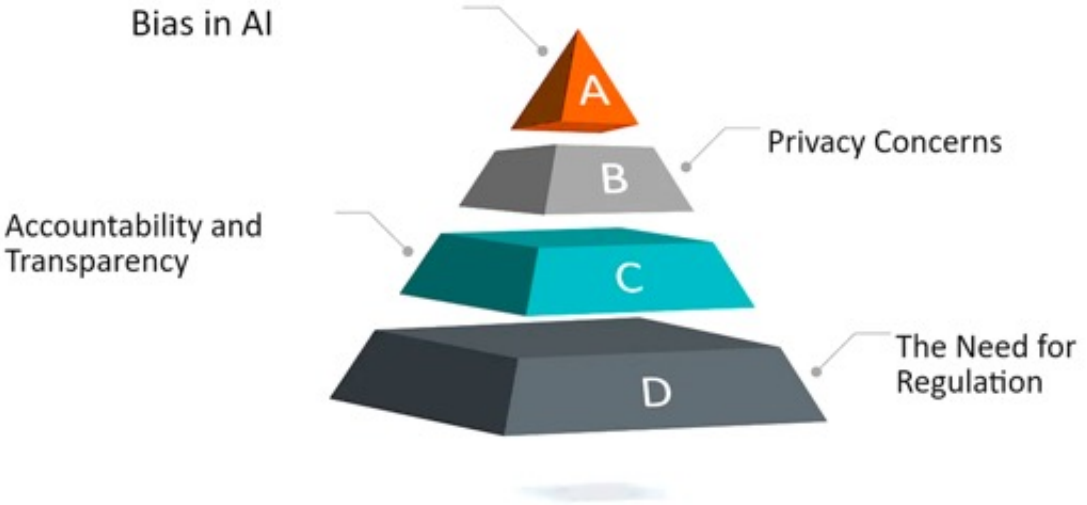
# 人工智能伦理问题的来源

# 人工智能伦理问题

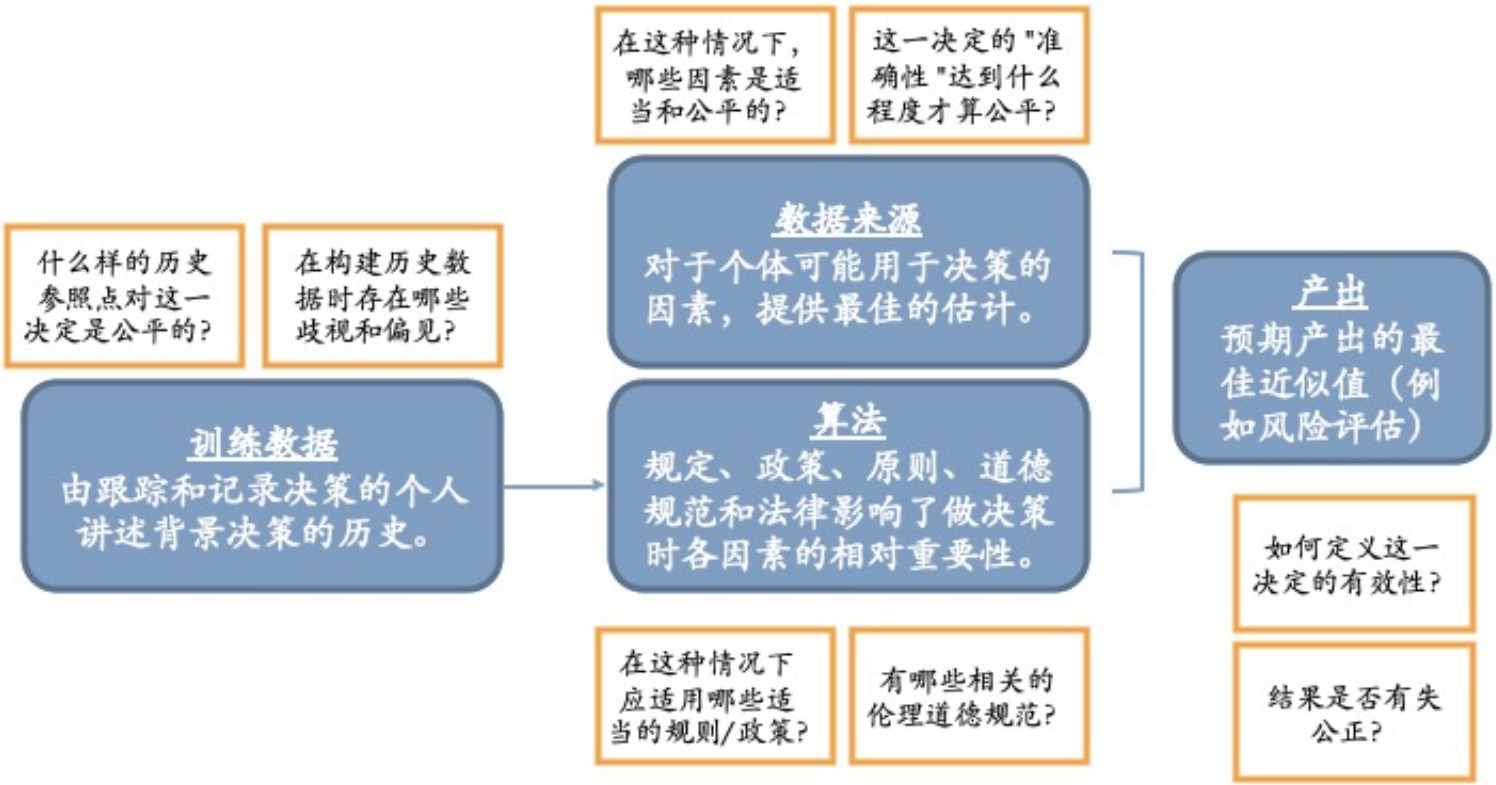
| | |
|---|---|
| 数据伦理 | 人工智能与大数据结合产生的伦理问题，比如隐私泄露、“大数据杀熟”等。 |
| 算法伦理 | 算法和模型可能带有偏见性、算法自主性造成的不确定风险、算法信任危机、算法评价滥用、算法对人的认知能力影响。 |
| 机器伦理 | 机器的伦理属性、机器具备一定的道德主体地位之后产生的伦理问题、智能机器人伦理、自动驾驶伦理等。 |
| 人机混合伦理 | 因脑机接口等技术的发展，使人类体能、感知、记忆、认知等能力在神经层面得到增强或提升，由此引发的伦理问题。 |
| 设计伦理 | 如何让机器遵循人类的道德原则。 |
| 全球伦理 | 人工智能的开发和训练对能源的消耗、对生态的影响，以及人工智能如何让欠发达地区受益。 |
| 超现实伦理 | 超级人工智能实现之后，它们的人权、道德地位乃至法律上的人格地位等问题，以及“机器人造反风险”的问题。 |

# 数据伦理

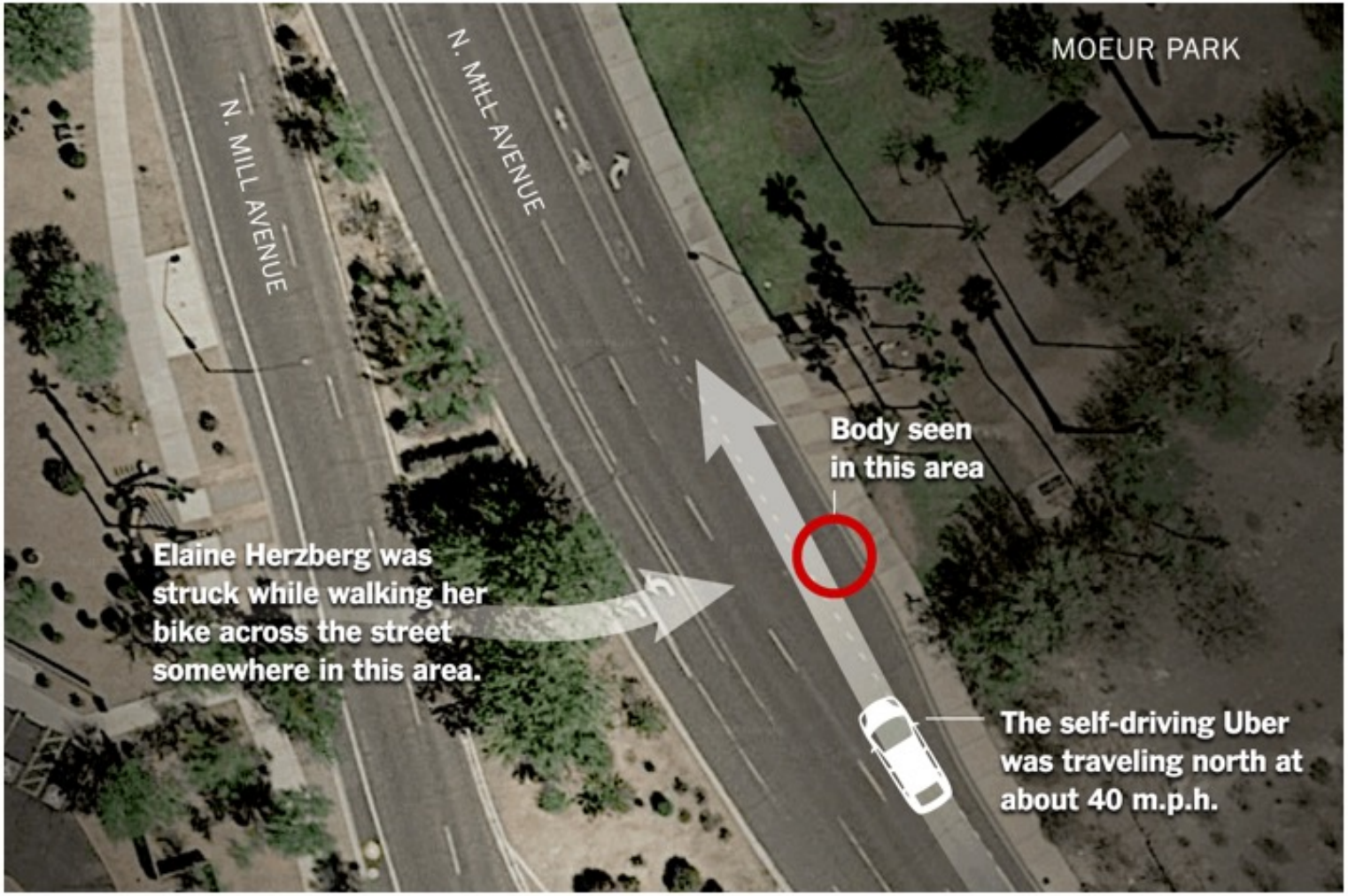
The Impact of Artificial Intelligence on Data Ethics




# 算法伦理


在这种情况下，哪些因素是适当和公平的？
这一决定的"准确性"达到什么程度才算公平？
数据来源
对于个体可能用于决策的因素，提供最佳的估计。
什么样的历史参照点对这一决定是公平的？
在构建历史数据时存在哪些歧视和偏见？
训练数据
由跟踪和记录决策的个人讲述背景决策的历史。
算法
规定、政策、原则、道德规范和法律影响了做决策时各因素的相对重要性。
产出
预期产出的最佳近似值（例如风险评估）
如何定义这一决定的有效性？
结果是否有失公正？
在这种情况下应适用哪些适当的规则/政策？
有哪些相关的伦理道德规范？

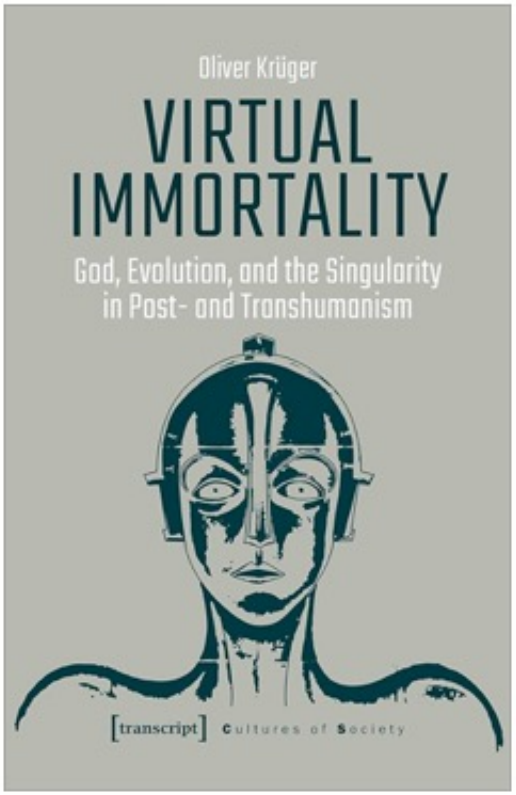
# 机器伦理




2018年3月：亚利桑那州Uber无人车事故
第一例导致死亡的自动驾驶事故

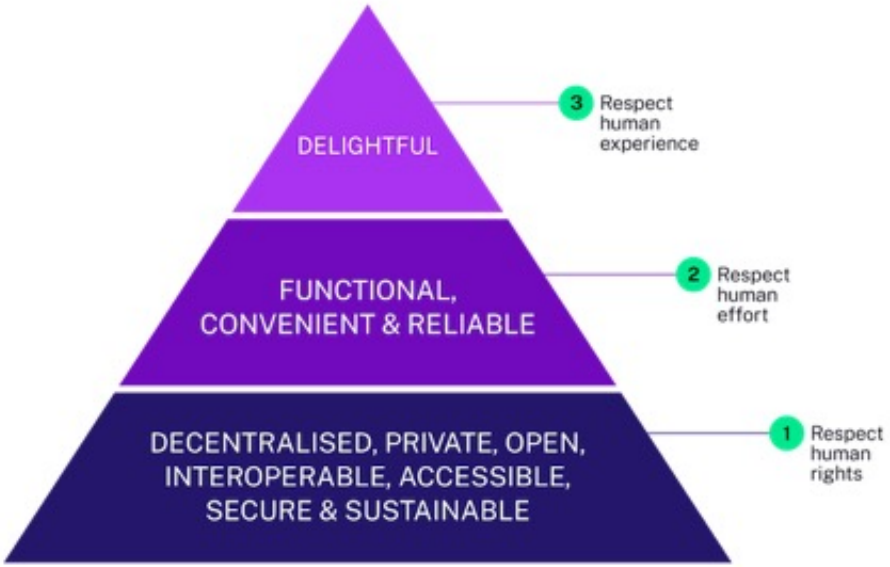
# 人机混合伦理


Brain-Computer Interface 脑机接口

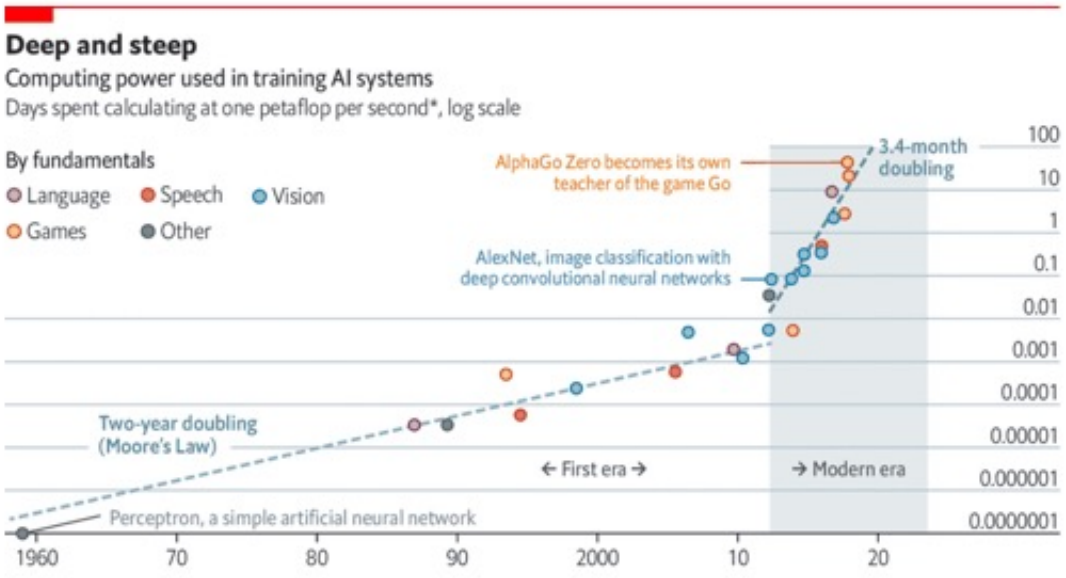



虚拟永生/数字永生

# 设计伦理


DELIGHTFUL
3 Respect human experience
FUNCTIONAL, CONVENIENT & RELIABLE
2 Respect human effort
DECENTRALISED, PRIVATE, OPEN, INTEROPERABLE, ACCESSIBLE, SECURE & SUSTAINABLE
1 Respect human rights

# 全球伦理



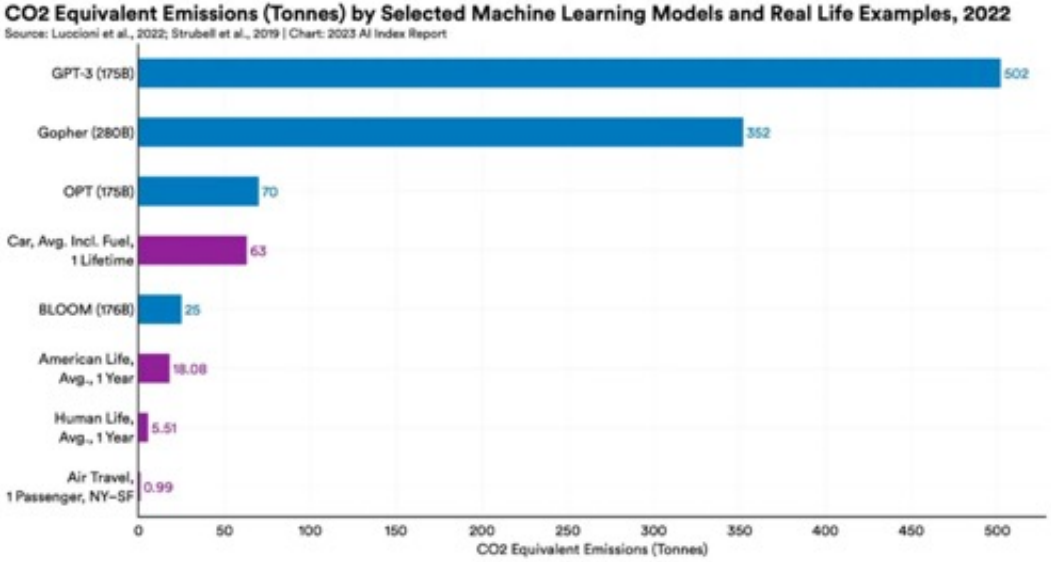




Figure 2.8.2

# 超现实伦理


HUMAN
ARTIFICIAL
INTELLIGENCE
STOP
the
ROBOTS
STOP THE
ROBOTS


HOW A.I
WILL KILL
HUMANITY

# 设计可信任的人工智能

## 提高人工智能产品被信任程度的“五大支柱”

| 公平性 | 可解释性 | 可靠性 | 透明度 | 安全性 |
| --- | --- | --- | --- | --- |
| • 如何保证产品对每个人来说是公平的，不存在歧视和偏见。 | • 能否向用户清楚地解释算法模型设计思路以及数据来源。 | • 是否容易被恶意攻击，是否只服务于某一个群体而有损其他人的利益。 | • 是否允许用户查看元数据和决策过程。 | • 能否保证数据隐私的安全。 |

# 设计可信任的人工智能产品

“如何让AI获得信任”不是普通的技术问题（不要用工具和技术的局限性当借口），而是社会学问题（社会学的核心是人，而非技术）。

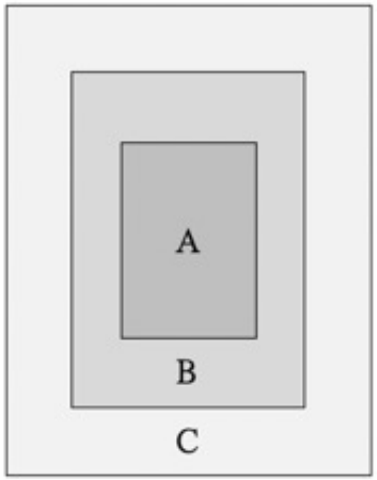
可信任人工智能产品开发的准则：

- 开发团队的人员背景需要多样化（不同性别、族裔、国别、文化背景）——多样性越高的团队犯错的几率越小（“群体智慧理论”）。
- 建立正确的项目管理流程，在设计过程中采用系统的同理心框架。
- 确保工具的正确性——工具本身不带有偏见和其他风险。
- 尊重人类的文化和人文主义思想——人工智能应被设计来增强人类的能力，而非取代人类的能力。

# 设计可信任的人工智能产品

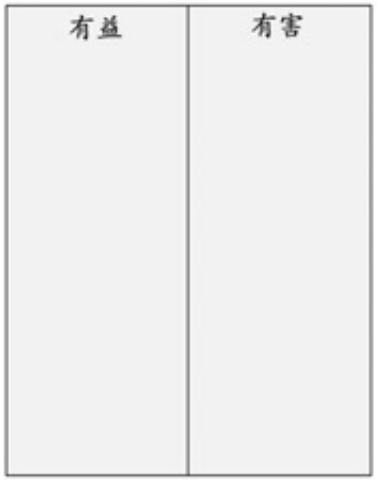
## 人工智能产品设计时的伦理评估

### 效果层分析
### Layers of Effects



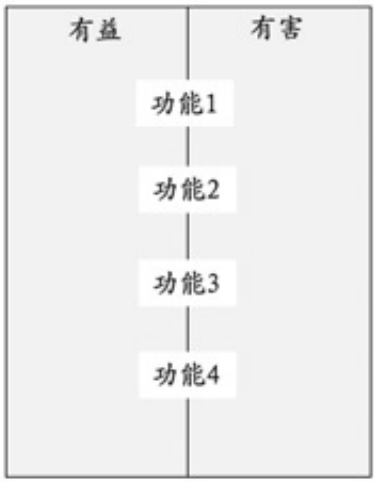

A：主要效果——AI产品设计时的目标效果。
B：次要效果——AI产品设计时的已知额外效果。
C：第三效果——AI产品投入使用后可能出现的意外效果，包括潜在的伦理问题和其他风险。

### 二分映射
### Dichotomy Mapping




将第三效果层可能出现的意外效果列举出来，让后将它们分为有益的效果和有害的效果。

### 伦理破解
### Ethical Hacking




思考：如何避免二分映射时发现的有害效果。同时应该思考，如果有益效果或设计时的其他产品亮点和功能被发挥到极致（即：被滥用），是否会出现物极必反或其它可能引起负面效果的情况。

# 期末考试

2024年 1月16日 14:00--16:00

综合楼 602