腾讯研究院
Tencent
Research Institute

人工智能伦理：问题与策略
AI Ethics: Problems and Strategies

曹建峰/Jeff Cao
腾讯研究院研究员/Research Fellow at TRI

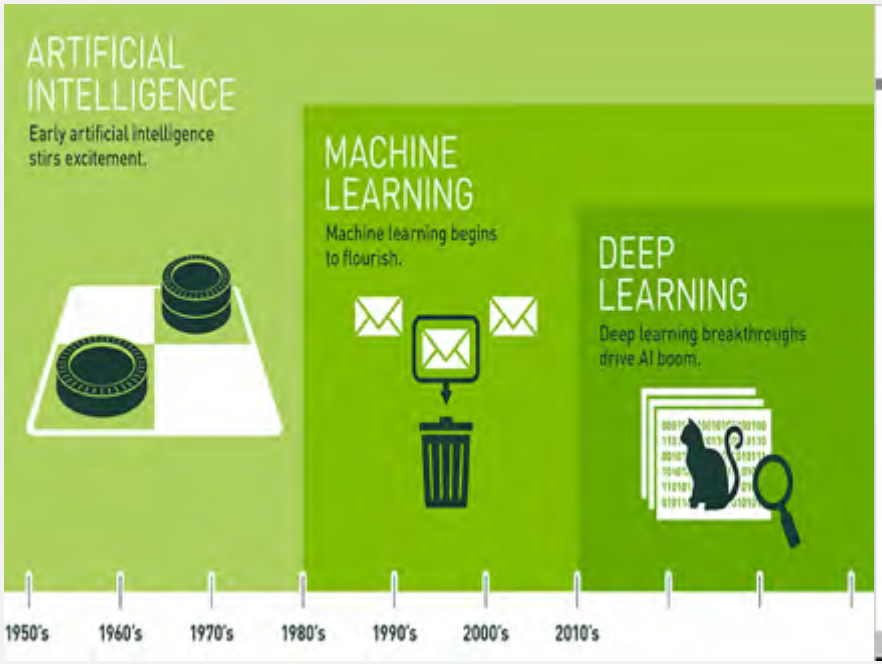
腾讯研究院
Tencent
Research Institute

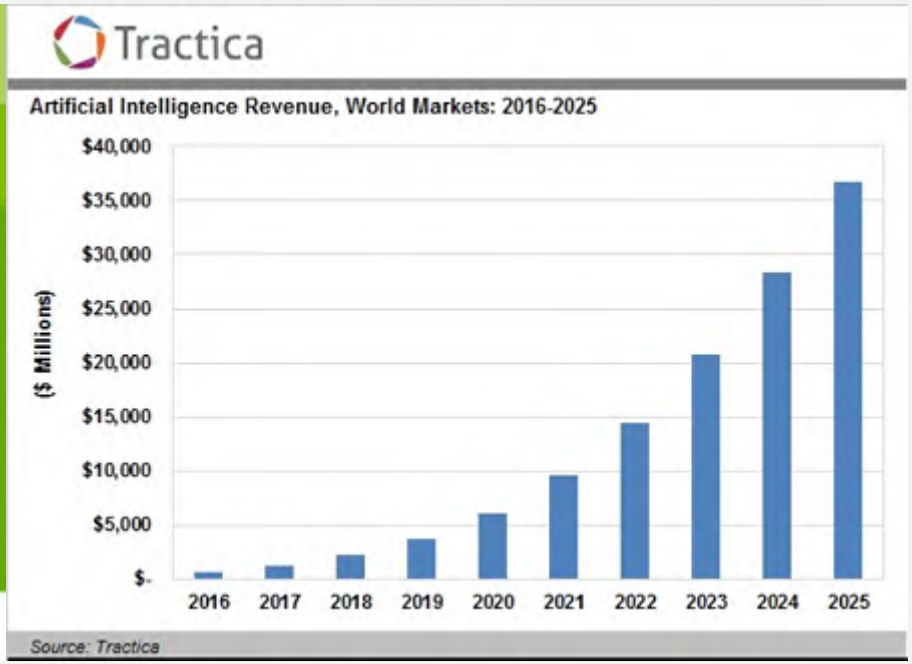
一、人工智能时代到来，算法决策开始主导

第三次AI浪潮开启人工智能时代

腾讯研究院
Tencent
Research Institute

- **技术进步：改进的机器学习（算法）**，更强大的计算能力，大数据
- **应用广泛：**图像识别、**语音识别、机器翻译、**虚拟助手、推荐引擎、**医疗诊断、自动驾驶**、智能机器人、计算机视觉、自然语言处理……
- **业界风向：**谷歌、微软、Facebook等科技巨头纷纷转向AI，AI创业和投资如火如荼，市场规模增长空间巨大





生活在算法之下，算法决策开始主导

腾讯研究院
Tencent
Research Institute

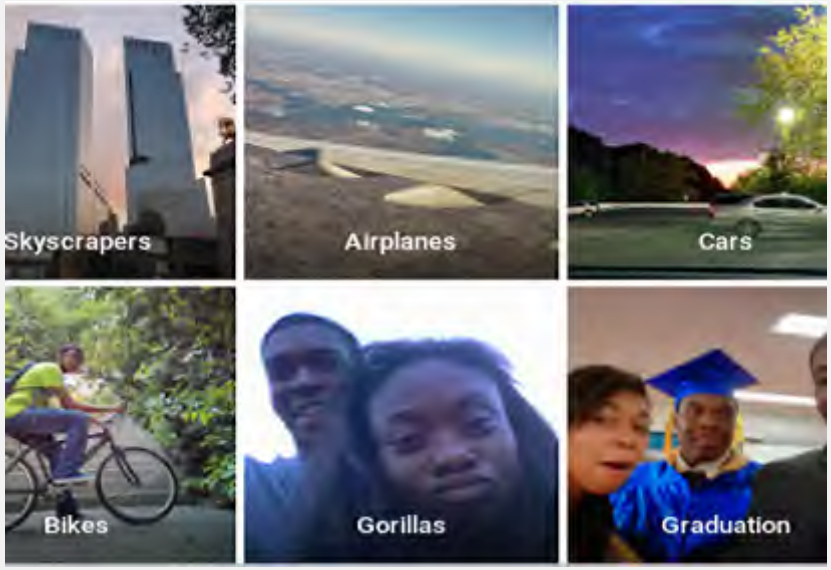
越来越多的决策、任务/工作将被部分或者完全自动化，由算法主导或者辅助

二、变革背后的阴暗面：不容忽视的伦理问题

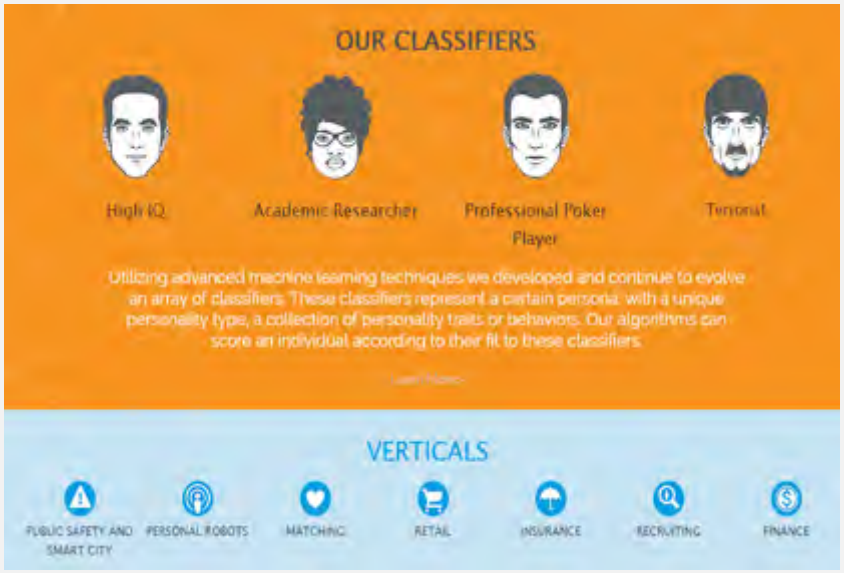
算法歧视越来越多

腾讯研究院
Tencent
Research Institute

图像识别

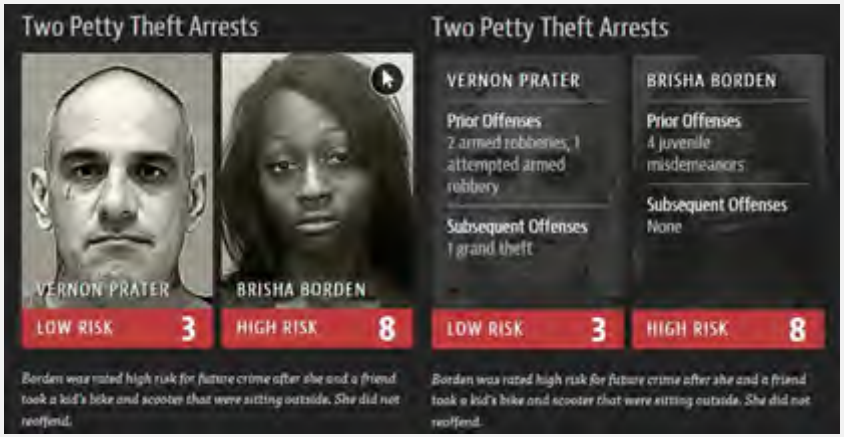


“看脸识人格”算法Faception



聊天机器人

犯罪风险评估算法COMPAS



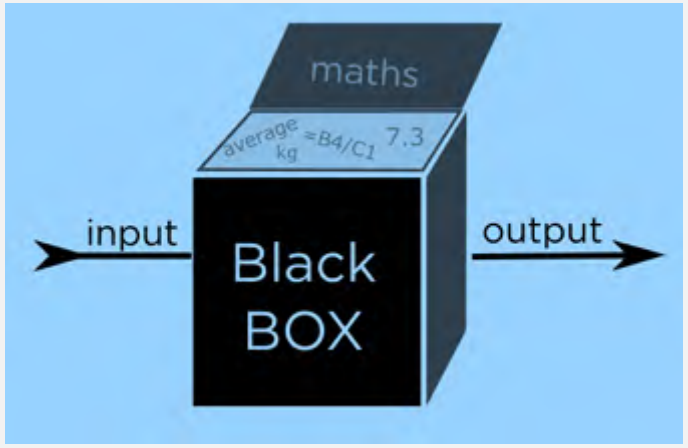
算法歧视具有危害

腾讯研究院
Tencent
Research Institute

规模
Scale

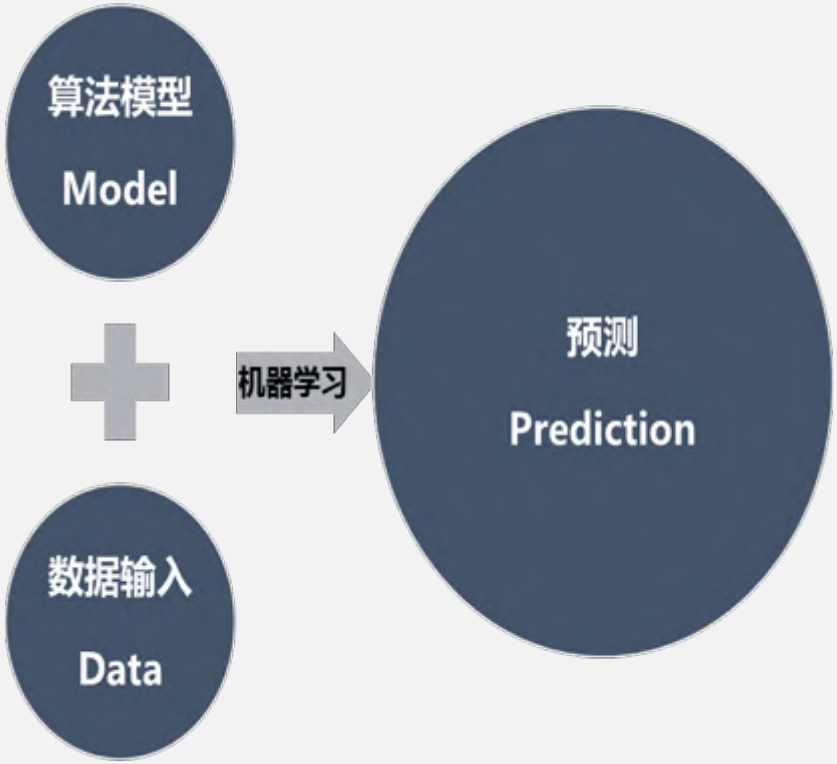
损害
Damage

黑箱
Black Box



算法并不客观，可能暗藏歧视

腾讯研究院
Tencent
Research Institute



- **算法模型**

  算法是“以数学方式或者计算机代码表达的意见”，包括其设计、目的、成功标准、数据使用等都是设计者、开发者的主观选择

- **数据输入**

  数据是社会的镜子，历史数据本身可能是歧视性的

  数据不准确、不完整、过时，所谓“垃圾进，垃圾出”

  样本大小差异：依赖多数（majority）学习

- **习得的歧视**

  自我学习、适应、改进的算法可能从交互中习得人类社会中既有的歧视

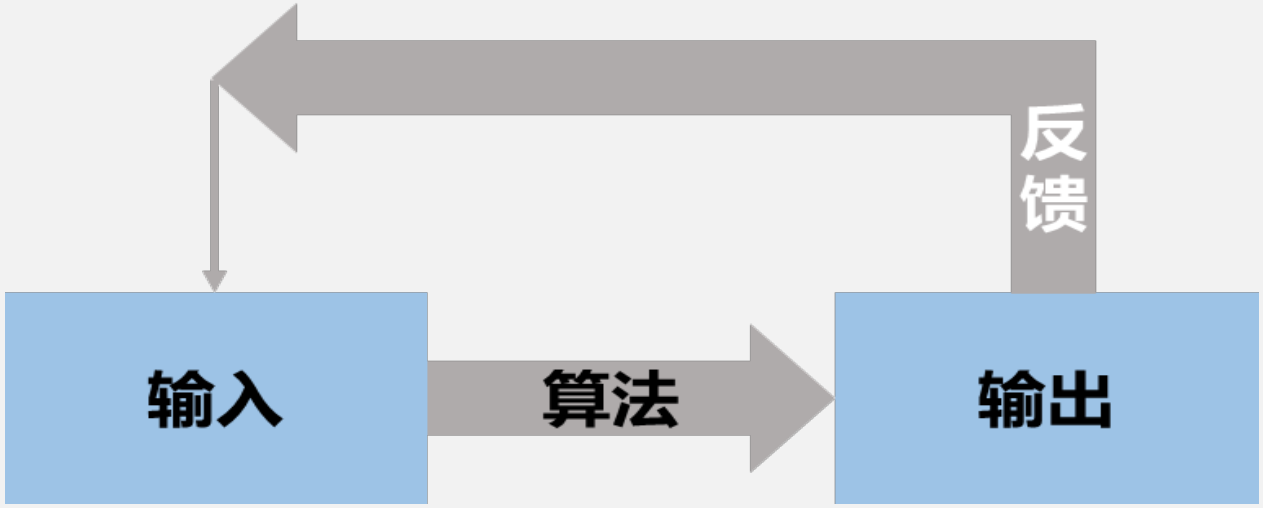
算法固化、放大歧视，可使歧视自我长存

腾讯研究院
Tencent
Research Institute

“谁掌握过去，谁就掌握未来；谁掌握现在，谁就掌握过去。”——奥威尔《1984》

算法的“自我实现的歧视性反馈循环（feedback loop）”

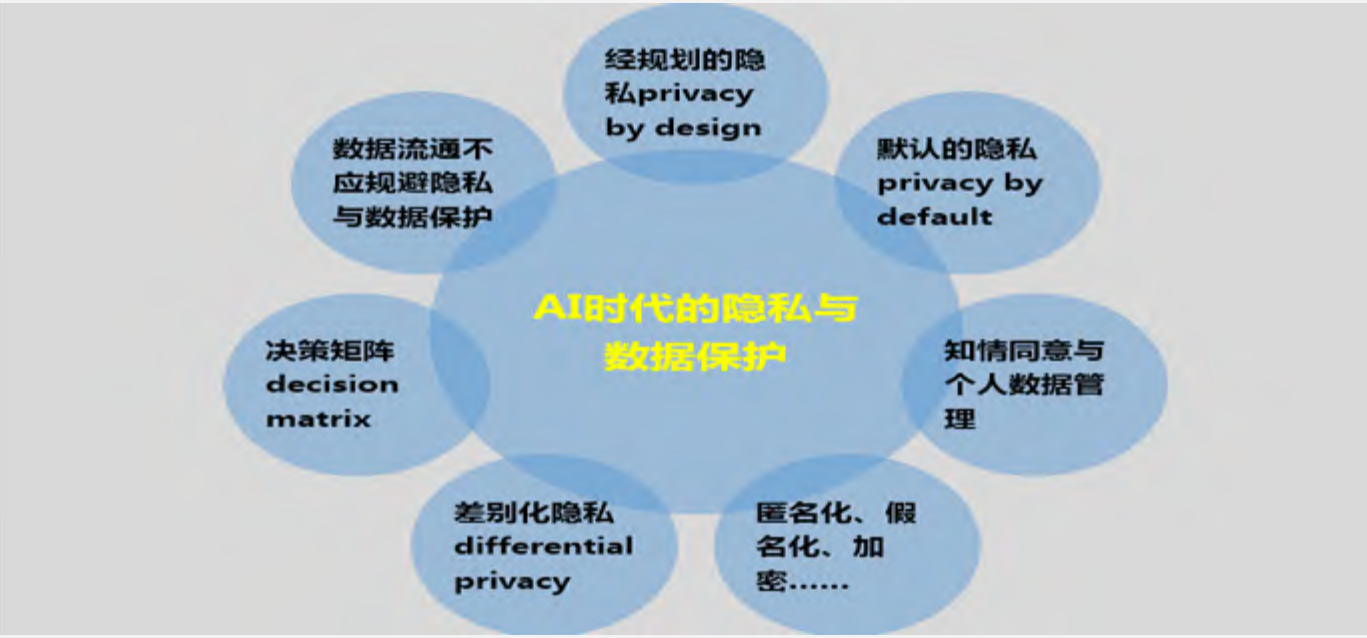
- Predictive policing
- 犯罪风险评估
- E-score（比如信用评估）



数据是AI时代的新石油，加深隐私忧虑

腾讯研究院
Tencent
Research Institute

- **训练AI需要大量训练数据：**数据收集无处不在，AI对数据（包括敏感数据）的大规模收集、使用，可能威胁隐私
- **自动化画像、决策：**越来越广泛的应用可能给个人权益产生不利影响
- **数据流通（currency）**：各种服务之间交易数据，数据流动更加频繁，极大削弱个人对其数据的控制和管理



AI安全与责任问题不容回避

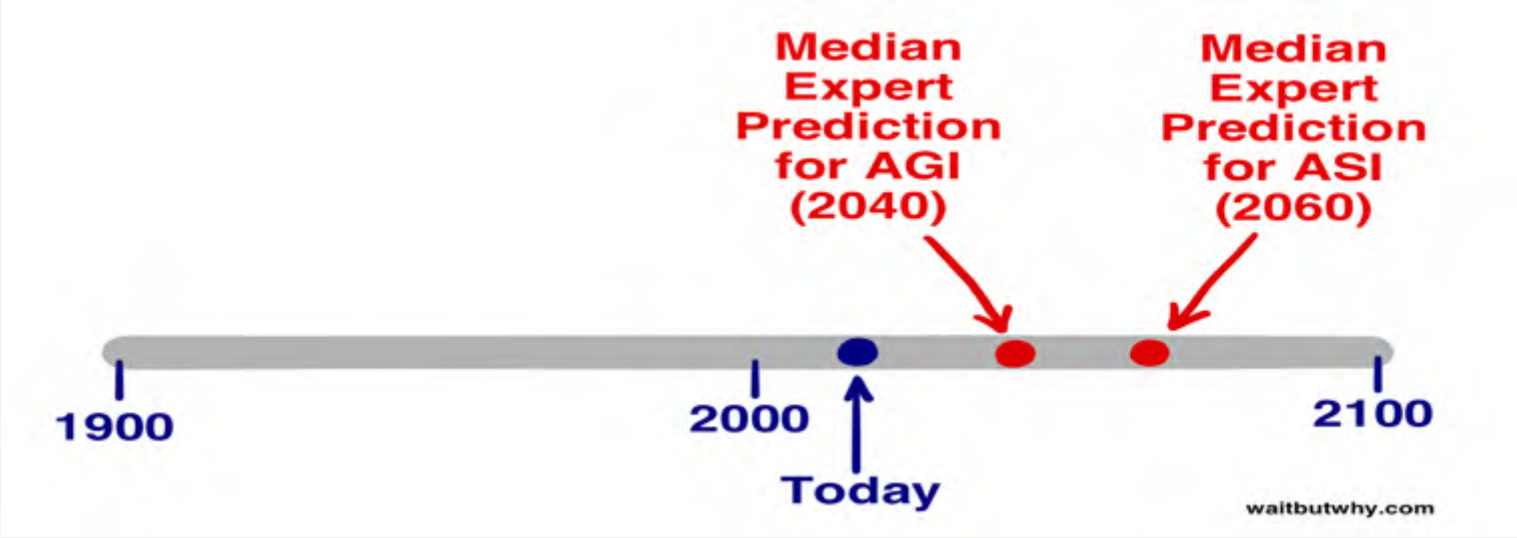
腾讯研究院
Tencent
Research Institute

确保智能机器人安全、可控

- **行为安全+人类控制**
- 阿西莫夫：机器人三定律
- 2017年阿西洛马会议：提出23条人工智能原则，构建有益AI

智能机器人造成人身、财产损害的责任承担

- 影响责任承担的三大问题：**可预测性，可解释性，因果关系**
- 谁来承担责任？



机器人权利，如何界定AI的人道主义待遇

腾讯研究院
Tencent
Research Institute

机器人权利的伦理基础

- Moral Agent（道德主体）
- Moral Patient（道德受体）

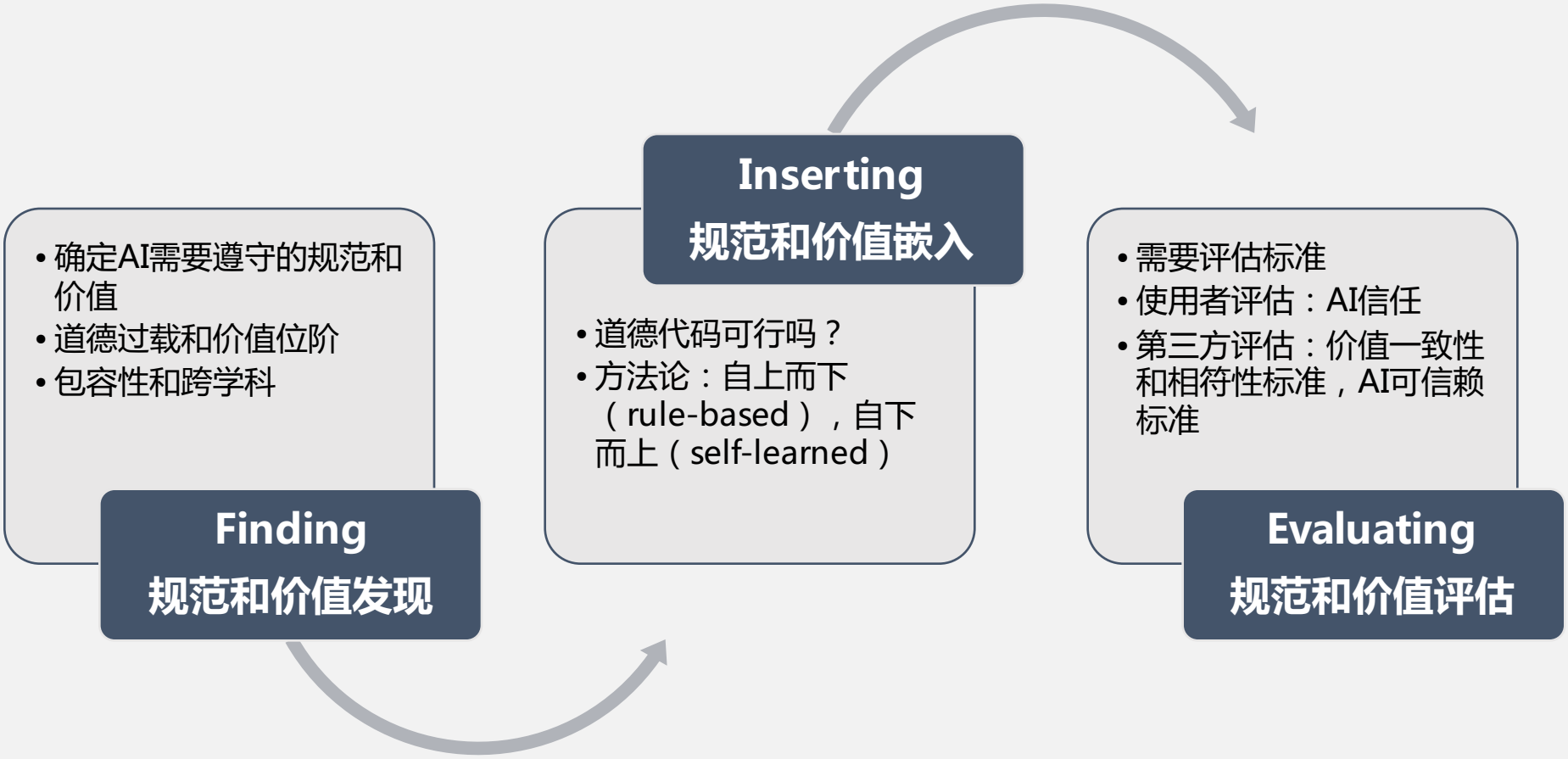
自主智能机器人在法律上是什么？是否可以享有一定的法律地位？

- 自然人？法人？动物？物？
- 欧盟考虑赋予自主智能机器人新的法律人格（电子人electronic person）

三、构建算法治理的内外部约束机制

合伦理设计：将规范和价值嵌入AI

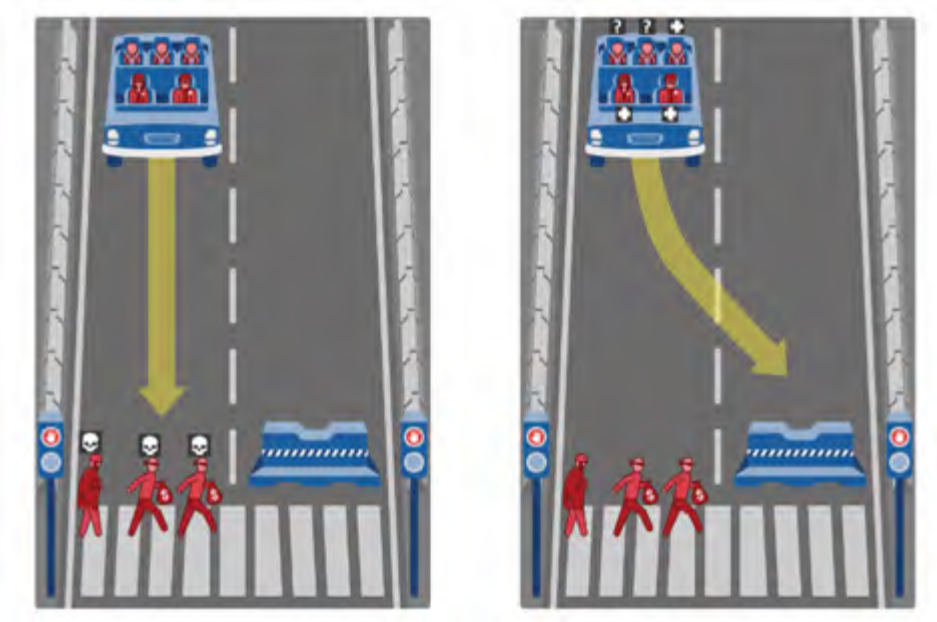
腾讯研究院
Tencent
Research Institute

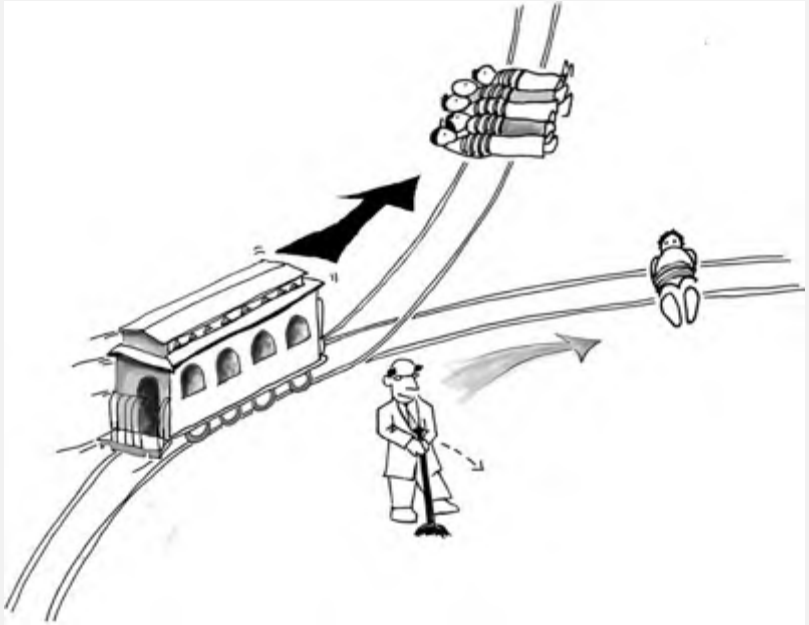


道德机器的伦理困境

腾讯研究院
Tencent
Research Institute

- MIT的“道德机器（moral machine）”调查
- 电车困境中的价值冲突：**功利主义（结果）VS绝对主义（行为）**





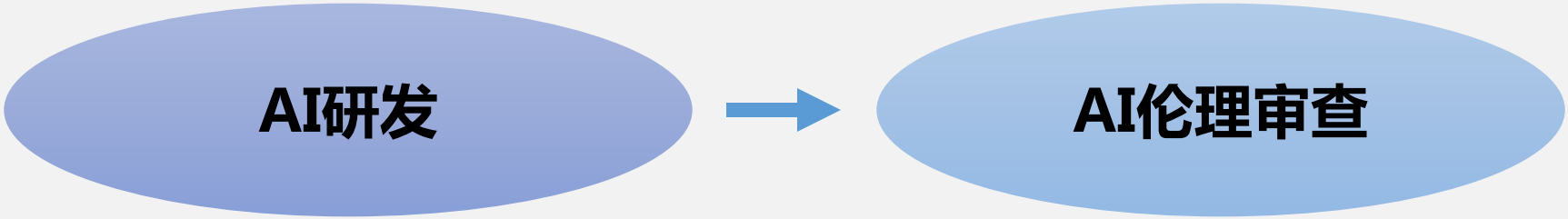
价值对接（value alignment）问题

腾讯研究院
Tencent
Research Institute

- **“一心一意”的机器人的行为，可能并非我们人类想要的**
  - 服务机器人、扫地机器人……
- **Stuart Russell：兼容人类的AI（human-in-the-loop）**
  1. 机器人的唯一目标是最大化人类价值的实现
  2. 机器人一开始不确定人类价值是什么
  3. 人类行为提供了关于人类价值的信息

在AI研发中贯彻伦理原则

腾讯研究院
Tencent
Research Institute



针对AI研发人员的伦理准则

- **AI研发活动应遵循伦理原则，**包括有益性、不作恶、自治、**正义**、基本权利、预防措施、**包容性、透明性、多样性、**责任、安全、可逆性、**隐私**……

负责任的设计

1. 设计的目的是什么？
2. 为谁而设计？
3. 不为谁设计？

建立AI伦理审查制度

- 伦理审查应当是跨学科的，多样性的，对AI技术和产品的伦理影响进行评估并提出建议

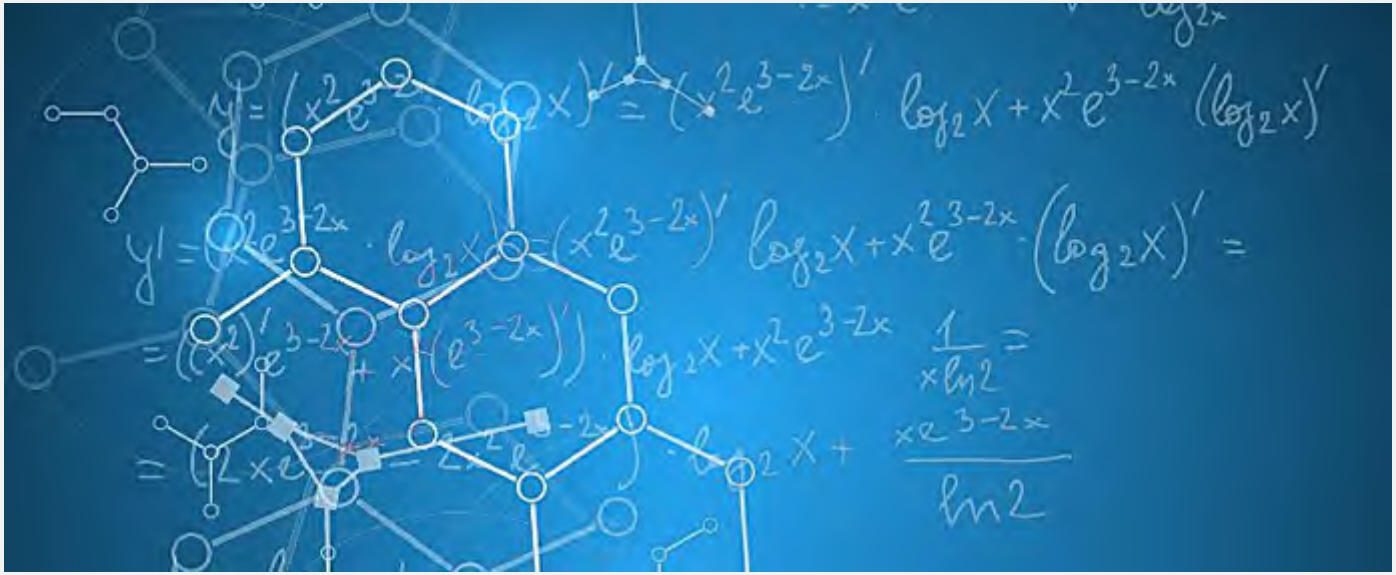
业界的伦理审查实践

- DeepMind的伦理委员会
- **DeepMind Health的独立审查委员会**
- IBM的伦理委员会

算法监管，避免算法作恶

腾讯研究院
Tencent
Research Institute

- 算法自身的设计越来越复杂，算法决策的影响越来越大，未来可能需要对算法进行必要的监管
- **可能的监管措施包括：**
  - 标准制定（分类，性能标准，设计标准，责任标准）
  - 透明性
  - 审批
- 欧盟：呼吁成立统一的机器人和人工智能监管机构



法律救济，弥补算法之害

腾讯研究院
Tencent
Research Institute

- **针对算法决策（自动化决策）**
  - 确保透明性：知情+解释
  - 提供申诉机制：向算法问责，对算法决策进行审查
- **针对智能机器人造成的人身、财产损害**
  - 救济原则：无辜受害人应当得到救济
  - 新的责任规则：严格责任，差别化责任，强制保险和赔偿基金，智能机器人法律人格

腾讯研究院
Tencent
Research Institute

谢谢聆听！

扫一扫上面的二维码图案，加我微信