

CHAPTER 07

智能时代的伦理、法律与未来

当机器学会行动，人类如何抉择？

通识选修课《具身智能：从看见到行动》

为什么技术需要"刹车片"？

前六章

机器"能做什么"

感知 → 认知 → 行动 → 仿真 → 现实

→

第7章

人类"应该怎么做"

伦理 → 法律 → 治理 → 责任

7.1 电车难题 → 7.2 机器换人 → 7.3 全球治理 → 7.4 公民责任 → 7.5 可持续发展

"AlphaGo可以下出人类想不出的妙手，却无法理解'为什么要下棋'。

Atlas可以完成后空翻，却无法回答'翻给谁看'。

能力不等于价值，智能不等于智慧。"

核心问题：技术之上，是人性

7.1 电车难题的"具身化"

1967年，菲利帕·福特提出"电车难题"

场景一

高速"鬼探头"，刹车不及

直行撞行人 vs 变道撞邻车

场景二

山路迎面失控货车

直行对撞 vs 右转坠崖

场景三

红灯人群闯红灯

撞人群 vs 转向撞墙伤乘客

核心洞察：自动驾驶系统的设计者在"代行社会的道德选择职能"
这不是技术问题，而是价值问题

生活中的"电车难题"

- 医疗资源分配：80岁老人 vs 5岁儿童，唯一呼吸机给谁？
- 内容审核："宁可误删正常内容"还是"宁可放过有害内容"？
- 自动驾驶每天在真实道路环境中做数百万次类似权衡决策

【课堂辩论】自动驾驶伦理困境

辩题：不可避免的事故中，应当优先保护车内乘客还是车外行人？

A组：乘客优先

乘客是系统使用者，已承担信任；保护"自己人"是人之常情

B组：行人优先

行人更"无辜"；乘客自愿使用自动驾驶，已承担风险

C组：数量优先

救更多人是唯一标准，不区分乘客/行人

D组：规则优先

谁遵守交通规则就保护谁，违规者自担风险



MIT"道德机器"实验（2018）

- 233个国家/地区，**4000万份**道德判断
- 四种观点恰好对应四种文化差异
- 结论：**没有统一答案**

→ 自动驾驶"伦理设置"可能需要"本地化"

7.1.2 "最小伤害"伦理框架

Minimize Harm Framework：无法避免时，选择总体伤害最小的方案

MIT"道德机器"实验发现

数量优先
多数人选择救更多人

年龄优先
多数人选择救年轻人

文化差异
集体主义→保护乘客

规则优先
保护遵守交通规则者

德国伦理委员会20条准则（核心）

- 保护人的生命优先于保护财产
- 不应以牺牲无关第三方为代价

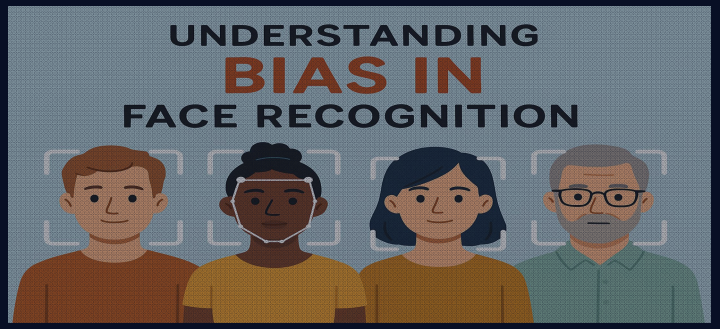
关键追问

"最小伤害"的计算权在谁手中？

工程师？政府？还是车主自选"道德模式"？

7.1.3 算法偏见：AI"学会"了人类的偏见

悖论：AI本身不偏见，但从有偏见的数据中学习



面部识别：深色皮肤女性错误率34.7% vs 浅色男性0.8%

亚马逊招聘：系统性给女性简历更低分 → 已终止使用

司法算法：黑人被告"误报率"远高于白人（历史逮捕率被人为抬高）

自我强化循环

偏见数据 → 偏见算法 → 偏见决策 → 产生新的偏见数据 → ...

解决方向：多样化数据 | 偏见审计 | 公平性约束 | 可解释AI

7.1.4 责任归属：出了事故谁负责？

观点	责任主体	支持理由	反对理由
产品责任论	制造商	最有能力确保安全	AI缺陷是"系统性"的
操作者责任论	使用者	防止过度依赖AI	L4/L5"操作者"概念已模糊
分布式责任论★	多方共担	最公平	责任追踪机制复杂
AI法律主体论	AI本身	形式上"解决"归属	AI无意识，无法真正担责

各国立法进展

欧盟AI法案
全生命周期风险管理

德国《自动驾驶法》'21
L4级由车辆所有人担责

中国《人工智能法》草案
算法备案 + 安全评估

7.2 机器换人：焦虑与真相

WEF预测（2030年）：取代9200万岗 | 创造1.7亿岗 | 净增7800万岗

高风险 >70%

- 流水线装配工
- 仓库分拣员
- 数据录入员、基础翻译

中风险 30%-70%

- 法律助理
- 初级医学诊断
- 基础设计

低风险 <30%

- 艺术家、医生、教师
- 战略咨询师
- 伦理委员会成员

核心概念：技术革命不是"消灭"工作，而是"转换"工作

具身智能特殊冲击：全球约3000万职业司机面临转型 | 人形机器人进入工厂

7.2 历史视角与"不可替代性"

革命	消失的岗位	诞生的岗位
第一次工业革命	手工纺织工	铁路工人、机械师
第二次工业革命	体力劳动者	电工、电器工程师
信息革命	打字员、文员	程序员、UI设计师、自媒体
智能革命	数据录入、基础翻译	AI训练师、提示词工程师、伦理顾问

人类的"不可替代性"

想象力
想象不存在的事物

同理心
真正"感受"他人情感

价值判断
模糊情境中伦理选择

跨域创造力
连接不相关领域

"增强"思维：计算器没有替代数学家，印刷机没有替代作家，AI也不会替代人类

7.3 全球AI治理：三条道路

维度	欧盟模式	美国模式	中国模式
核心理念	权利优先	创新优先	发展与安全并重
立法方式	统一立法（AI法案）	分散立法	顶层规划+行业规范
监管强度	严格	宽松	中等
风险定义	基于权利	基于安全	综合（权利+安全+稳定）
企业成本	合规成本高，规则清晰	合规成本低，规则模糊	中等

欧盟AI法案"四色预警"

红灯 禁止
社会评分、实时生物识别

黄灯 严格
自动驾驶、医疗AI

绿灯 告知
聊天机器人需透明

无灯 自愿
垃圾邮件过滤等

7.3 算法问责与具身智能特殊治理

"不可能三角"：性能 ↔ 透明 ↔ 商业机密（三者不可兼得）

解决方案

可解释AI (XAI)
LIME/SHAP足够好的解释

分层透明
监管者/使用者/公众

算法影响评估
类似"环境影响评估"

"算法保险箱"
第三方持有，事故时开箱

具身智能三大特殊挑战

物理交互不可逆性
虚拟错误=数据错误，具身错误=物理伤害

"人在回路"边界
L2需要人，L5不需要人，中间怎么管？

国际标准缺失
同一系统需满足多套法规

"场景化"治理：封闭场景(矿山)→宽松 | 半开放(园区)→中等 | 开放(城市)→严格

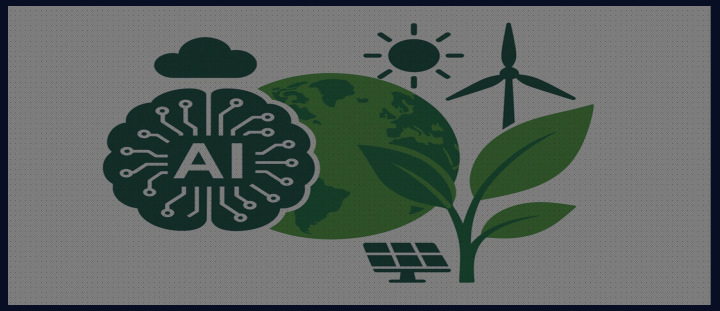
7.4+7.5 公民素养与技术向善

AI素养三层能力

① 理解边界
AI能做什么、不能做什么

② 识别伪真
辨别Deepfake、AI生成内容

③ 参与讨论
AI政策需要公众声音



AI与可持续发展

- 能源优化：DeepMind使风电预测准确率提高40%
- 野生动物保护：TrailGuard AI抓获偷猎者
- "绿色AI"：模型压缩、高效算法、可再生能源

技术人文主义：不在"技术乌托邦"和"技术悲观"之间选边站，承认潜力同时坚守人文价值

【课堂讨论】你是哪种思维？

情境：你是HR经理，公司引入AI招聘系统（自动筛选简历、安排面试、评估"文化匹配度"）

Q1 系统会"替代"你的工作，还是"增强"你的效率？

Q2 发现对女性候选人有系统性偏见，你应该怎么做？

Q3 你希望在多大程度上"依赖"系统的判断？

Q4 作为求职者，希望系统"透明"到什么程度？

引导：技术的角色是"替代"还是"增强"——取决于我们怎么使用它

本章核心框架：技术之上，是人性

技术之上是人性

伦理维度

- 电车难题的具身化
- "最小伤害"伦理框架
- 算法偏见与公平性
- 责任归属四种观点

法律维度

- 产品/操作者/分布式责任
- 欧盟AI法案四色预警
- 德国/中国立法进展
- 算法问责与透明

社会维度

- 机器换人与终身学习
- 人类不可替代性
- AI素养与公民责任
- 技术人文主义

全球AI治理：欧盟(权利优先) | 美国(创新优先) | 中国(发展安全并重)

COURSE COMPLETE

从"被动感知"到"主动想象"——不只是机器的成长

七个章节，七个发展阶段。从AlphaGo的"遗憾"到Atlas的后空翻，
从感知、认知、行动到仿真、现实、伦理——我们走完了具身智能的完整旅程。

"想象力是方向，计算力是动力，但选择什么样的方向，始终是人类自己的责任。"

课后作业

1. 阅读本章专栏7-1、7-2、7-3（欧盟AI法案四色预警、MIT道德机器实验、阿西洛马原则）
2. 思考：如果你要设计一款面向中国市场的自动驾驶系统，你会如何设置"伦理偏好"？写一篇500字短文
3. （选做）给"十年后的自己"写一封信

下节预告：实验课导论——6个实验总览