第七讲 算法伦理:人工智能的风险与规制

华东师范大学哲学系 潘斌

为何算法需要伦理?

算法决策的社会影响已超越技术范畴,成为紧迫的伦理议题。当算法开始决定谁能获得贷款、谁会被判刑、谁能得到工作机会时,我们必须审视其背后的伦理考量。

AI刑事定罪

算法评估累犯风险,影响量刑决策

自动驾驶困境

面对"电车难题"的道德抉择

算法招聘歧视

AI筛选简历时的系统性偏见

本节的核心关切

我们不(主要)关心算法的效率。我们关心的是算法如何重塑人类社会的基本价值。

01

正义问题

算法如何分配社会资源与机会?

02

自主性问题

算法如何重塑人类认知与自主性?

03

治理问题

我们应如何规制这种新型权力?

哲思之"算法":算法是什么?

技术定义

解决问题的步骤序列

数据处理的自动化流程

伦理学视角

编码化的权力(Codified Power)

自动化的判断(Automated Judgment)

将复杂的伦理决策简化为量化模型

哲思之"风险":算法风险的规范性维度

超越技术风险(如系统崩溃、精度不足),伦理学视角的风险指对人类基本规范(Norms)的威胁。

对公平的侵蚀

系统性地制造不平等的机会分配

对正义的威胁

延续和放大历史性的社会不公

对尊严的剥夺

将人简化为数据点和风险评分

对自主性的削弱

人类判断力的外包与去技能化

算法伦理的学科定位

作为应用伦理学(Applied Ethics)的一个分支,算法伦理的核心任务是识别、分析和回应算法系统在设计、部署和使用中引发的伦理挑战。

不是"软约束"

而是构建可信赖AI的"硬前提"

跨学科融合

哲学、法学、计算机科学的交汇

风险的源起 1:数据的伦理困境

数据不是"客观"的,而是"历史的沉淀"。每一个数据集都承载着过去社会偏见和不平等的印记。

历史偏见

Historical Bias

数据反映了过去的歧视性实践

样本偏见

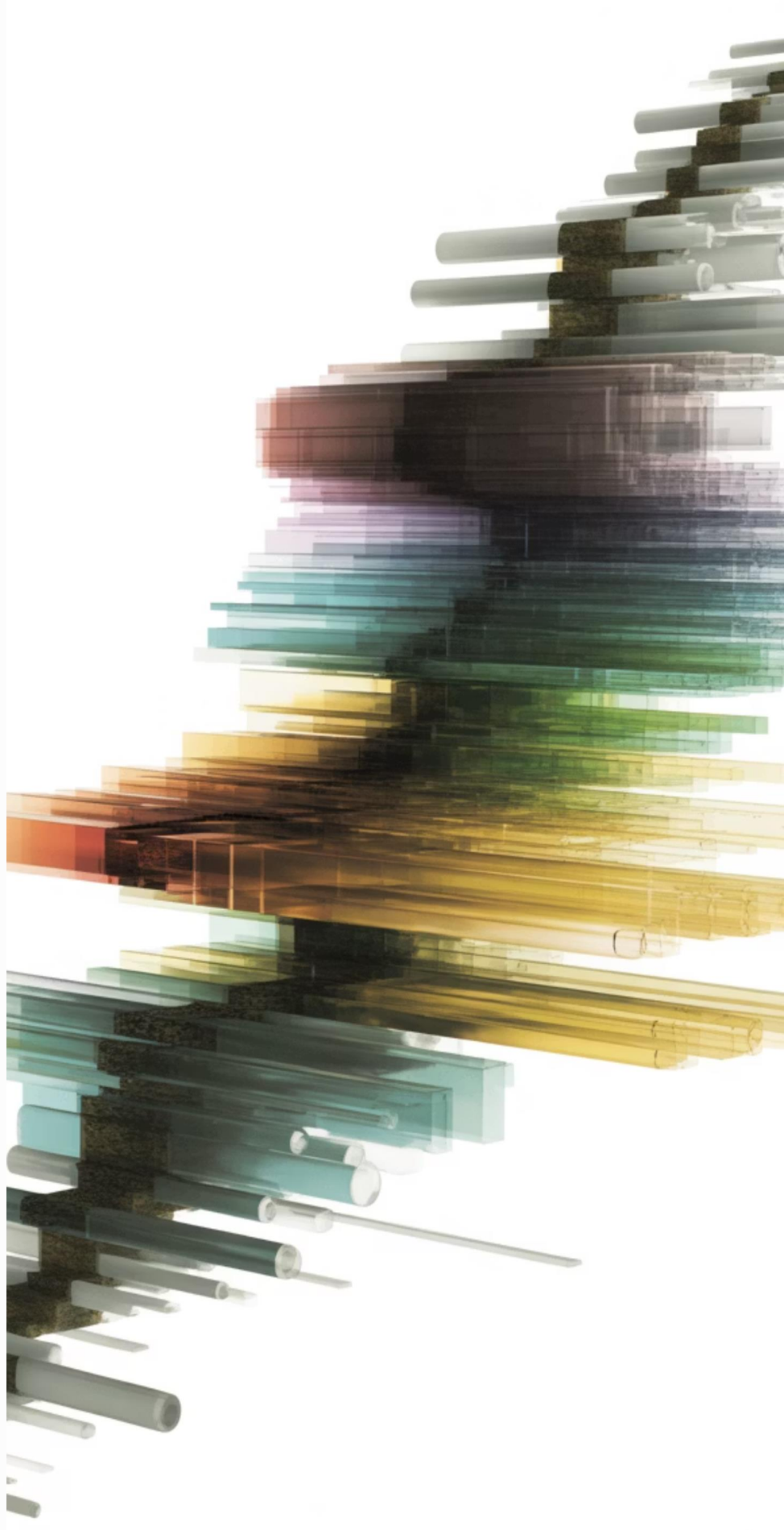
Sampling Bias

数据采集过程中的系统性遗漏

GIGO原则

垃圾进,垃圾出

有偏见的数据产生有偏见的结果



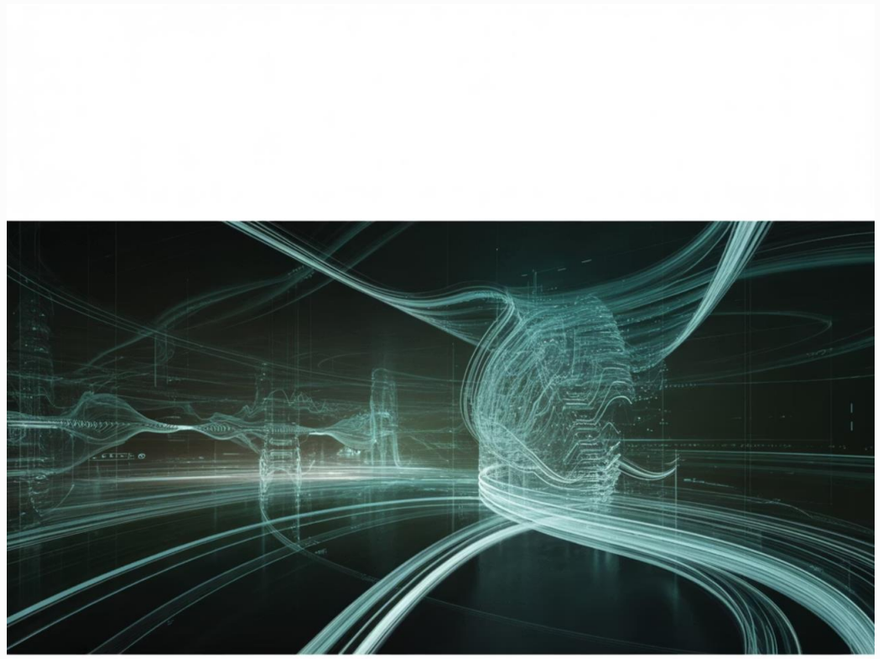
风险的源起 II:模型的伦理困境

模型是"带有偏见的世界观"。设计者的选择——目标函数(Objective Function)的设定本身就是一种价值

代理变量的风险

- 用"邮编"代理"信用"
- 用"教育背景"代理"能力"
- 用"历史行为"代理"未来风险"

这些替代指标往往隐含着对特定群体的系统性歧视。



从"偏见"到"歧视"

区分概念的关键

偏见 (Bias)

统计上的不准确或不平衡

技术层面的问题

可能是无意识的

歧视 (Discrimination)

基于偏见所导致的不公正对待

规范性伤害 (Normative Harm)

对特定群体的系统性伤害

> 一个技术上"中性"的偏见,在社会语境中可能转化为道德上不可接受的歧视。

案例分析:COMPAS算法

美国司法系统中的累犯风险评估工具(Correctional Offender Management Profiling for Alternative Sanctions)揭示了算法如何延续和放大社会不公。

45%

非裔美国人

被错误标记为高风险的比例

23%

白人群体

被错误标记为高风险的比例

对不同族裔的"假阳性率"差异巨大,构成系统性歧视。算法不仅反映了历史偏见,更将其制度化、自动化,使歧视变得更加隐蔽和难以挑战。

课程结构：四大模块概览

01

导论

算法、权力与风险

02

风险谱系

算法的核心伦理挑战

03

原则确立

审视算法的哲学框架

04

伦理规制

从原则到实践的路径

第一节小结

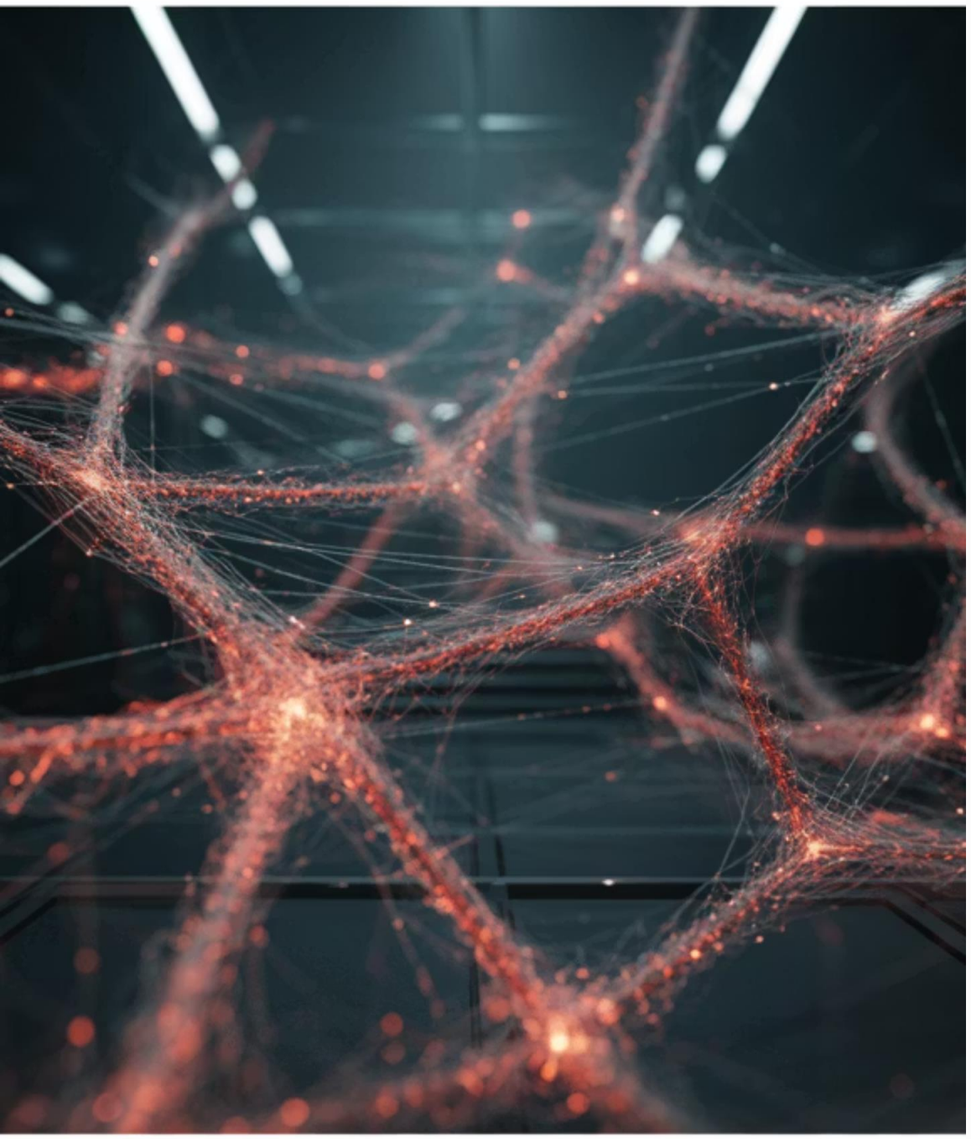
核心观点

算法是一种权力

算法风险是规范性伤害

承上启下

接下来,我们将深入探讨风险的具体表现



风险谱系

算法的核心伦理挑战

风险 I:不透明性与"黑箱"问题

Opacity

1

技术性

深度学习的复杂性

2

商业性

商业秘密保护

3

知识性

用户/监管者的理解障碍

"黑箱"的三重根源使审查、问责和救济成为不可能。当我们无法理解算法如何做出决策时,我们也就无法质疑其公正性。

风险 II:可问责性真空

Accountability Vacuum

"算法替罪羊"

> "是算法决定的,不是我"

核心问题

- 设计者?
- 数据提供者?
- 使用者?
- 算法本身?

传统的责任链条在自动化决策中被打破。

案例分析:优步的"幽灵"派单

算法对司机的精细化管理与控制揭示了新型的"数字利维坦"和劳动异化。

信息不对称

司机无法了解派单逻辑和定价机制

算法控制

通过"动态定价"和"派单策略"实现精准管理

劳动异化

工人服从于算法的效率目标,失去自主性

风险 III:算法的规训与认知塑造

超越"歧视"的深层风险:算法不仅"反映"现实,更在"塑造"现实。

信息茧房

Filter Bubbles

回音室

Echo Chambers

认识论权威

Epistemic Authority

算法成为"认识论的权威",定义了我们能看到什么、相信什么、思考什么。

案例:社交媒体与政治极化

推荐算法为追求"参与度"(Engagement)如何系统性地放大了极端和两极化的内容。

算法逻辑

- 最大化用户停留时间
- 优先推荐引发强烈情绪的内容
- 强化用户既有观点

社会后果

- 公共话语空间的撕裂
- 民主基础的侵蚀
- 社会共识的瓦解

风险 IV:人类自主性的让渡

Autonomy

辅助决策

算法提供建议,人类保留最终决定权

影响决策

算法建议显著影响人类判断

替代决策

人类完全依赖算法,丧失判断能力

人类的"去技能化"(Deskilling)和"判断力外包)。我们是否正在失去做出重要道德判断的能力?

风险 V:算法的"非人化"

Dehumanization

算法将复杂的个体简化为"数据点"和"风险评分",剥夺了情境(Context)、同情(Empathy)和尊严(Dignity)在决策中的地位。

算法福利发放系统

将复杂的个人困境简化为资格评分

医疗资源分配

忽视患者的具体情况和特殊需求

风险 VI:致命性自主武器

Lethal Autonomous Weapons Systems (LAWS)

算法风险的极端形态:算法自主决定"生与死"。

核心伦理争议

"有意义的人类控制"(Meaningful Human Control)的必要性

将道德责任赋予机器触及伦理底线

当机器拥有杀戮的权力时,人类尊严和战争伦理的基础将被彻底颠覆。

风险 VII:算法的"异化"

Alienation

从马克思主义哲学视角看算法对人类社会关系的重构。

劳动异化

劳动者与劳动工具(算法)的异化

工人失去对劳动过程的控制

关系异化

人类社会关系被算法逻辑所主导

社交互动变成数据交换

目的颠倒

人类反过来服务于算法的效率目标

而非算法服务于人类

什么是"坏"的算法?

技术上的"坏"

- 不准确
- 效率低
- 不稳定
- 易受攻击

伦理上的"坏"

- 不公正
- 不透明
- 侵犯尊严
- 削弱自主

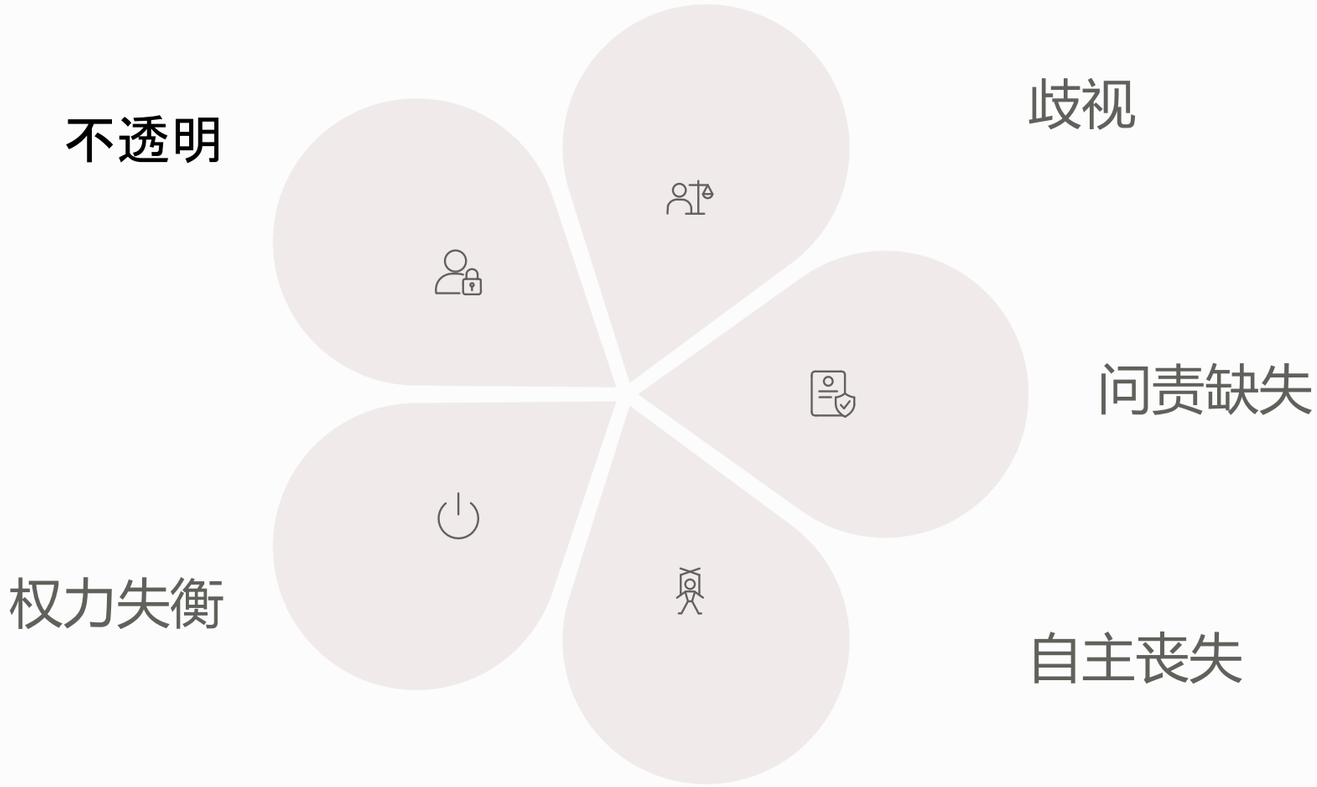
深度思考: 一个技术上"完美"的算法,可能在伦理上是"灾难性"的。

效率与公正、准确与公平,往往存在根本性的张力。

风险的交织性

Intersectionality

偏见、不透明、权力不对等、自主性丧失等风险是相互叠加、相互强化的。



例如:不透明的"黑箱"加剧了歧视,使得问责成为不可能,进而削弱了受影响者的自主性和议价能力。

第二节小结

> 算法风险是多维度、系统性、且具有高度哲学意涵的。它不仅是技术问题,更是关乎人类尊严、自由和正义的根本性挑战。

面对这些风险,我们需要怎样的伦理原则来指引?

原则确立

审视算法的哲学框架

哲学工具 1:后果主义

Consequentialism

核心思想

行为的道德价值取决于其"后果"

代表:功利主义

"最大化总体福祉"

算法应用

追求算法效用的最大化,如最大化社会效率、最低化总体错误率。

面临的挑战

- 如何定义"效用"?
- "大多数人"的利益是否可以牺牲"少数群体"?
- COMPAS案:降低整体犯罪率 vs 对特定族裔的不公

哲学工具 II:义务论

Deontology

核心思想:行为的道德价值取决于其是否遵守了"道德义务"或"规则"。代表:康德的"绝对命令"。

> "人是目的,而非仅仅是手段"
>
> ——伊曼努尔·康德

尊重个体

算法设计是否尊重了个体权利与尊严?

道德红线

是否存在不可逾越的"红线"?如禁止基于种族的歧视性决策

哲学工具 III:美德伦理学

Virtue Ethics

核心思想:关注"行动者"(Agent)的品格。代表:亚里士多德。

审慎 (Prudence)

在设计中展现深思熟虑和远见

公正 (Justice)

致力于公平对待所有利益相关者

谦逊 (Humility)

承认模型的局限性和不确定性

算法应用:我们应该培养怎样的算法设计者和使用者?

原则构建 I:公正与公平

Justice & Fairness

罗尔斯的正义原则为算法公平提供了哲学基础。

1

平等的自由权

每个人都应享有最广泛的基本自由

2

差异原则

社会和经济不平等应对最不利者最有利

3

机会公平

职位和地位应向所有人公平开放

算法公平的困境:如何定义"公平"?个体公平 vs 群体公平的张力。

公平的多种定义

多种"公平"定义在数学上不可兼得——这是算法伦理的核心困境之一。

反歧视	群体公平	个体公平
Unawareness 不使用敏感特征(如种族、性别)	Group Fairness 人口均等(Demographic Parity):不同群体的正向结果比例相同	Individual Fairness 相似的个体应获得相似的对待

核心张力:满足群体公平可能违反个体公平,反之亦然。这要求我们在具体情境中进行价值权衡。

原则构建 II:自主与尊严

Autonomy & Dignity

自主性

人类应保留做出重大决策的权利和能力

核心要求

"有意义的人类控制"

Meaningful Human Control

- 人类必须理解算法决策
- 人类必须能够干预和推翻
- 人类必须承担最终责任

尊严

算法决策过程应尊重人的内在价值

具体体现

- 提供决策解释
- 允许申诉和纠错
- 保护隐私和数据权利
- 避免将人简化为数据

原则构建 III:透明性与可解释性（XAI）

Transparency & Explainability

1

透明度

系统层面的开放性

公开算法的存在、目的和影响范围

2

可解释性

模型决策层面的可理解性

解释具体决策的原因和逻辑

3

知情权

受影响者的基本权利

了解算法如何影响自己的权利

为何需要可解释性?它是实现公平、问责和自主的前提。不仅仅是技术上的XAI,更是对受算法影响者"知情权"的尊重。

原则构建 IV:稳健性（鲁棒性）与安全性

Robustness & Safety

从伦理学视角看,这是一种"不伤害"原则(Primum non nocere)——医学伦理的首要原则。

功能正确

算法在预期场景下准确运行

抵御攻击

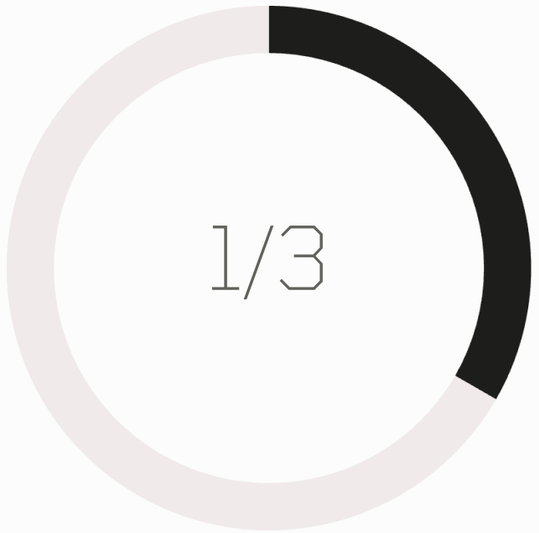
能够抵御对抗性样本和恶意操纵

安全护栏

具有可预见的失效模式和应急机制

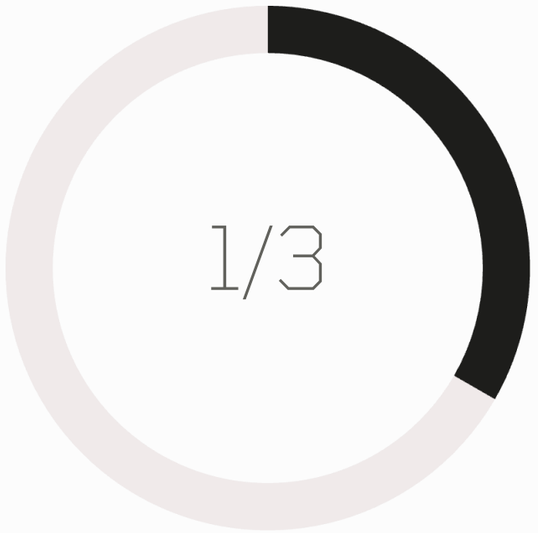
原则的张力:不可回避的难题

算法伦理不是寻找"唯一正确答案",而是在多元价值冲突中进行"审慎权衡"(Prudential Judgment)。



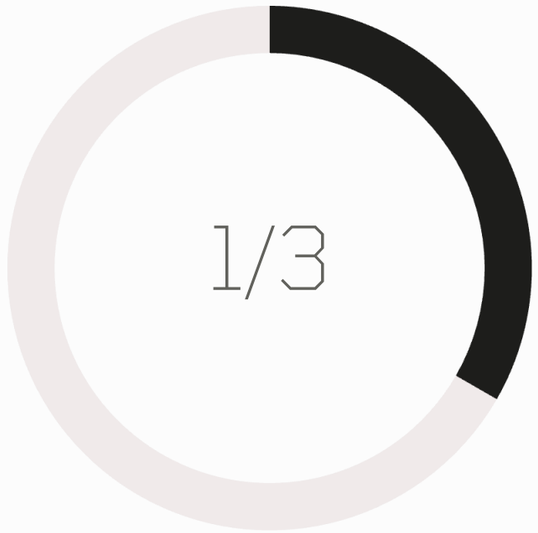
公平性 vs 准确性

提高公平性可能降低整体准确率



透明度 vs 隐私性

公开算法细节可能泄露个人信息或商业秘密



效率化 vs 自主性

自动化提高效率但削弱人类控制

第三节小结

我们从四大哲学流派中推导出了算法伦理的核心原则。

公正与公平

Justice & Fairness

自主与尊严

Autonomy & Dignity

透明与可解释

Transparency & Explainability

稳健与安全

Robustness & Safety

如何将这些"高阶原则"落地为"具体实践"?

伦理规制

从原则到实践的路径

治理的必要性:伦理的"硬化"

Hardening Ethics

问题

仅靠企业"自律"或"原则"是否足够?

答案

否。

历史经验表明,自愿性的伦理承诺往往在商业压力下让位。

解决方案

需要将"软伦理"(Soft Ethics)转化为"硬规制"(Hard Governance)

- 法律法规
- 技术标准
- 审计机制
- 问责制度

治理路径 I:法律与规制

Top-Down Approach

特征:由国家主导,具有强制性和普遍约束力。

欧盟《人工智能法案》

EU AI Act

基于"风险等级"的差异化治理:高风险AI需接受严格监管,包括透明度要求、人类监督和合规评估。

治理路径 II:伦理设计

Ethics by Design

特征:自内而外,在技术开发全周期中嵌入伦理考量。

设计阶段

在需求分析和系统设计时就考虑伦理影响

开发阶段

将伦理原则转化为可操作的技术规范

测试阶段

进行偏见检测和公平性评估

部署阶段

持续监控和迭代改进

口号:"在设计之初就考虑伦理"(Ethics by Design),而非"事后弥补"。

技术工具 1:算法影响评估

Algorithmic Impact Assessment (AIA)

一种前瞻性的治理工具。在算法部署前,系统性地评估其对社会、伦理和基本权利的潜在影响。

01

识别利益相关者

谁会受到算法的影响?

02

评估潜在风险

可能产生哪些伦理问题?

03

制定缓解措施

如何降低或消除风险?

04

持续监控

部署后的效果跟踪

类比:"环境影响评估"在工程领域的应用——在项目实施前预判和防范负面影响。

技术工具 II:FATE框架

Fairness, Accountability, Transparency, Ethics

F - 公平 (Fairness)

偏见检测与消除工具,如IBM的AIF 360、Google的What-If Tool

A - 问责 (Accountability)

算法审计(Auditing)与决策日志记录,建立责任追溯机制

T - 透明 (Transparency)

可解释AI技术(XAI),如LIME、SHAP,帮助理解模型决策

强调:技术工具是治理的必要(但非充分)条件。工具本身不能替代人类的伦理判断。

治理路径 III:多方共治

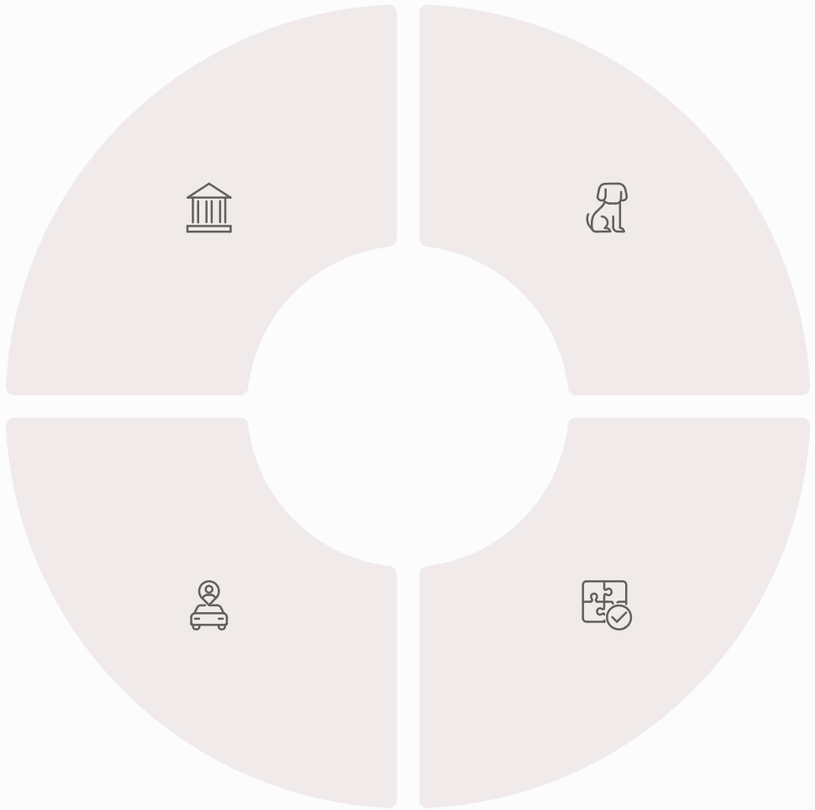
Multi-Stakeholder Governance

政府

制定法律法规,进行监管执法

企业

负责任的研发和部署



公民社会

监督、倡导和参与

学界

提供理论研究和技术支持

挑战:如何建立有效的对话与制衡机制?关键:提升全社会,特别是受影响群体的"算法素养"(Algorithmic Literacy)。

治理的挑战:步调问题

Pacing Problem

核心矛盾

技术创新的速度(指数级)远远快于伦理反思和法律制定的速度(线性)

后果

治理永远"滞后",法律总是在追赶技术

对策

- "敏捷治理"(Agile Governance):快速迭代的监管框架
- "监管沙盒"(Regulatory Sandboxes):在受控环境中测试新技术
- 前瞻性立法:基于原则而非具体技术

治理的挑战:全球化与碎片化

算法和数据是跨国界的,但法律是属地的。

"布鲁塞尔效应"

Brussels Effect

欧盟的高标准能否成为全球标准?GDPR的全球影响力证明了这种可能性。

治理碎片化

缺乏全球统一的算法治理框架,导致"监管套利"和标准冲突

文化差异

不同文化和政治体制对"公平"、"隐私"等核心原则的理解不同

例如:欧洲强调隐私权,美国强调言论自由,中国强调网络主权

迈向"可信赖AI"

Trustworthy AI

"可信赖AI"不仅是技术目标,更是社会契约。它要求算法系统在法律、伦理和技术三个维度上都达到高标准。

合法的 (Lawful)

遵守所有适用的法律法规

合伦理的 (Ethical)

遵守伦理原则和价值观

稳健的 (Robust)

技术上和社会环境上均稳健可靠

本节小结:算法伦理的未来

算法伦理不是技术的"刹车",而是"方向盘"。它要求我们从"技术中心主义"转向"以人为本"(Human-Centric)。

算法应该服务于人类的繁荣,而不是让人类适应算法的逻辑。

伦理不是创新的障碍,而是可持续创新的基础。

这是一场持续的、需要哲学智慧和技术实践相结合的对话。