



## 审视“智能”的边界：

## 面部识别和 AI 招聘中的伦理困境与治理路径批判

邓江波

3230102347

信息安全 2301 darstib@zju.edu.cn

### 摘要

人工智能（AI）技术的加速渗透正重塑社会结构与个体生活，引发日益严峻的伦理挑战。尽管全球范围内已形成一系列强调公平、透明、问责的 AI 伦理原则共识，但这些宏观准则在面部识别、AI 招聘等高风险应用场景下，从“原则到实践”的转化面临显著鸿沟。本文聚焦这两个直接关涉公民基本权利与社会公平的领域，进行批判性伦理分析。通过剖析系统性偏见放大、规模化隐私侵蚀、决策不透明、问责机制缺失及“数字相面术”等具体问题，本文旨在揭示通用伦理原则在面对具体技术、商业模式与社会情境时的实践困境与内在张力。在此基础上，本文审视了现有治理框架的不足，探讨了更具针对性、可操作性与适应性的治理策略，如强制性伦理影响评估、偏见审计、强化数据主体权利及多方参与治理等，旨在为推动 AI 技术朝向更负责任、公平和尊重人权的方向发展提供批判性思考与实践路径参考。

关键词：人工智能伦理；应用伦理；算法偏见；面部识别；AI 招聘；透明度；问责制；伦理治理；批判性分析；原则实践鸿沟

### Abstract

The accelerating integration of Artificial Intelligence (AI) technologies is reshaping societal structures and individual lives, posing increasingly severe ethical challenges. Despite a global consensus on AI ethical principles emphasizing fairness, transparency, and accountability, a significant gap exists between these macro-level guidelines and their practical implementation in high-risk scenarios like facial recognition and AI hiring. This paper critically analyzes these two domains, which directly impact fundamental rights and social equity. By dissecting issues such as systemic bias amplification, large-scale privacy erosion, decision opacity, accountability deficits, and “digital physiognomy,” it reveals the practical dilemmas and inherent tensions when applying general ethical principles to specific technologies, business models, and social contexts. Building on this critique, the paper examines the shortcomings of existing governance frameworks and explores more targeted, actionable, and adaptive strategies, including mandatory ethical impact assessments, bias audits, strengthened data subject rights, and multi-stakeholder governance. It aims to provide critical insights and practical pathways for guiding AI towards a more responsible, equitable, and human rights-respecting trajectory, bridging the principle-practice divide.

**Keywords:** Artificial Intelligence Ethics; Applied Ethics; Algorithmic Bias; Facial Recognition; AI Hiring; Transparency; Accountability; Ethical Governance; Critical Analysis; Principle-Practice Gap

# 引言

人工智能（AI）作为关键驱动力，其应用已广泛渗透至社会经济各层面，展现巨大潜力。然而，技术飞速发展伴随着深刻的伦理考量。AI 系统日益增强的自主决策能力和拓展的应用边界，持续引发新的伦理争议。例如，个性化推荐可能固化“信息茧房”并涉嫌操纵用户选择<sup>[1]</sup>，而生成式 AI（AIGC）则在知识产权、信息真实性、学术诚信乃至人类创造力本质上带来前所未有的伦理困境<sup>[2]</sup>。这些广泛关切共同指向核心议题：如何审慎引导和规制 AI 发展，确保其符合人类福祉与核心价值。

面对 AI 伦理挑战，国际社会、各国政府、学界及产业界已积极响应，发布一系列 AI 伦理原则与治理框架，如欧盟的《人工智能法案》（草案）<sup>1</sup>、中国的《新一代人工智能治理原则——发展负责任的人工智能》<sup>2</sup>、《人工智能伦理治理标准化指南（2023 版）》<sup>3</sup>、OECD 的《人工智能原则》<sup>4</sup> 等。这些框架普遍强调公平、透明、可问责、隐私保护、安全可靠等核心价值<sup>[3]</sup>。然而，将高阶原则有效转化为具体应用场景下的技术约束、操作规范和监管措施，仍面临巨大挑战。原则在实践中常遭遇现实复杂性的消解：价值冲突（如公共安全与隐私）、商业利益扭曲伦理考量、新技术使既有框架滞后等。

本文聚焦于两个尤为直接关联公民基本权利（隐私权、平等就业权）、社会公平正义及个体尊严，敏感且争议显著的应用领域：面部识别技术（FRT）的广泛部署与人工智能在招聘流程中的应用（AI Hiring Tools）。本文旨在通过对这两个领域的深入剖析，采用批判性视角，揭示通用 AI 伦理原则在具体实践中的局限，阐明伦理问题的具体表现、技术与社会根源，审视现有治理手段的不足，并探寻更具针对性、操作性与前瞻性的伦理治理改进路径。最终目标是推动 AI 在这些高风险领域实现真正负责任的应用提供理论支撑与实践建议。

## 一、面部识别技术的滥用风险与伦理边界的侵蚀

面部识别技术（FRT）凭借其便捷高效特性，在安防、身份验证、营销等领域迅速普及。然而，其规模化应用，尤其在缺乏充分伦理评估、严格规制和有效监督下，正对个人隐私、社会公平和公民自由构成严峻威胁，持续侵蚀技术应用的可接受伦理边界。

### 1.1 公共安全应用中的监控泛化与权利侵蚀

在公共安全领域，FRT 应用最广，也最具争议。许多城市部署大规模监控网络结合 FRT 进行实时追踪与身份比对，意图提升治安效率。例如，伦敦的“Ring of Steel”<sup>5</sup>，及中国部分城市的“天网”工程<sup>6</sup>。然而，这种全域实时身份识别能力将公共空间转变为潜在的永久监控场域，极大削弱了个体在公共场合的匿名权——这被认为是现代自由社会的重要基石<sup>[4]</sup>。公民参与合法公共活动（如和平集会）的意愿可能因担心被追踪记录而受压制，形成“寒蝉效应”<sup>[5]</sup>。

更具侵略性的案例是 Clearview AI 通过网络爬虫，未经同意从公开来源抓取数十亿人脸图像，构建庞大商业化面部数据库，服务于执法机构乃至私人客户<sup>7</sup>。此行为不仅在数据来源合法性上存疑，更严重挑战个人信息自决权、数据使用目的限制及正当程序要求，引发全球法律诉讼、监管调查和强烈伦理声讨<sup>8</sup>。

### 1.2 商业应用中的隐私侵犯与数据商品化

商业场景中，FRT 应用同样引人关切。零售商利用其分析顾客身份、习惯、情绪，进行精准营销或优化布局<sup>9</sup>。部分企业用其监测员工出勤、专注度甚至情绪<sup>10</sup>。这些操作常在用户不知情、未明确同意下进行，将生物特征数据视为商业资产，违背数据主体权利和隐私期待，引发对“监控资本主义”的担忧<sup>[6]</sup>。

<sup>1</sup><https://artificialintelligenceact.eu/>

<sup>2</sup>[https://www.most.gov.cn/kjbgz/201906/t20190617\\_147107.html](https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html)

<sup>3</sup><http://www.ircip.cn/bbx/1044770-1044770.html?id=26645&newsid=4806011>

<sup>4</sup><https://www.oecd.org/digital/artificial-intelligence/>

<sup>5</sup><https://www.bbc.com/news/uk-england-london-38418877>

<sup>6</sup>[http://paper.people.com.cn/rmzk/html/2017-11/20/content\\_1825998.htm](http://paper.people.com.cn/rmzk/html/2017-11/20/content_1825998.htm)

<sup>7</sup><https://www.forbes.com/sites/roberthart/2024/09/03/clearview-ai-controversial-facial-recognition-firm-fined-33-million-for-illegal-database/>

<sup>8</sup><https://www.aclu.org/cases/aclu-v-clearview-ai>

<sup>9</sup><https://www.forbes.com/councils/forbestechcouncil/2024/10/21/the-rise-of-facial-recognition-in-retail-what-shoppers-should-know/>

<sup>10</sup><https://www.schneier.com/blog/archives/2024/01/facial-recognition-systems-in-the-us.html>

<sup>11</sup><https://ny.eater.com/2024/10/24/24278212/brooklyn-wegmans-facial-recognition>

### 1.3 FRT 伦理困境的深层结构与交织影响

面部识别技术引发的核心伦理争议，并非孤立的技术问题，而是深植于其运行机制与社会效应的复杂交织之中。FRT 的自动化、持续性与规模化身份识别能力，特别是当面部数据被长期存储并与海量个人信息（如位置轨迹、消费习惯、社交网络）关联分析时，极易构建出详尽的个体数字档案。这不仅为数据滥用、歧视性决策乃至社会信用评分系统和政治操纵提供了前所未有的技术基础<sup>[7]</sup>，更从根本上挑战了“公共场所无绝对隐私”这一传统法律观念的边界，使普遍监控成为一种技术上可行且日益常态化的现实。

与隐私侵蚀相伴而生的是算法偏见内嵌及其歧视性后果的放大。大量研究和现实案例（如美国多起非裔公民因 FRT 误识别被错捕事件<sup>[213]</sup>）证实，当前主流 FRT 算法在识别非白人群体、女性、特定年龄段时，准确率显著低于识别成年白人男性<sup>[8]</sup>。偏差主要源于训练数据代表性失衡（数据偏见），也可能与算法设计（模型偏见）相关<sup>[9]</sup>。后果可能是灾难性的：错误刑事指控、金融或就业歧视、社会福利分配不公等，直接损害弱势群体权益。

...

多重困境相互关联、彼此强化，共同描绘出 FRT 技术潜在的巨大伦理风险图景。

### 1.4 高阶伦理原则在 FRT 实践中的张力与不足

面对这些问题，现有高阶 AI 伦理原则（如“公平”、“隐私保护”、“透明度”）在实践中显得力不从心。“公平”如何具体界定和量化？不同公平性度量可能冲突<sup>[10]</sup>。如何有效检测消除深层偏见？“隐私保护”在“公共安全”面前如何划定具体界限？“透明度”应达何种程度？“可解释性”能否满足问责要求？如何构建有效的问责链条？这些问题在 FRT 应用中均缺乏清晰答案和有效机制。

## 二、AI 招聘工具中的隐性歧视与评估有效性质疑

AI 引入招聘流程，旨在提高效率、扩大候选人池、克服人类主观偏见。然而，实践表明，AI 招聘工具（简历筛选、AI 面试、行为评估）不仅未能消除偏见，反而可能以更隐蔽、系统化的方式复制、固化甚至放大现实社会中的结构性歧视<sup>[11,12]</sup>。同时，其评估方法的科学性、公平性和对候选人尊严的影响也备受质疑。

### 2.1 简历筛选系统：历史偏见的自动化复制

简历筛选系统是最常见的 AI 招聘应用。它们通过学习企业过往招聘数据来自动筛选申请。核心风险在于“有偏数据输入，有偏结果输出”<sup>[9]</sup>。若历史招聘决策本身带偏见（如性别、种族、年龄），AI 模型会学习并再现此模式。结果可能是，模型将与受保护特征强相关的代理变量（如毕业院校、社团活动、居住区域）识别为负面信号，系统性地不公筛选合格少数群体候选人<sup>[13]</sup>。亚马逊因其 AI 招聘工具对女性存系统性偏见而弃用的案例<sup>14</sup>便是典型例证。这使得 AI 工具非但未实现“客观性”，反成固化职场不平等的“数字看门人”<sup>[14]</sup>。

### 2.2 AI 面试与评估工具：“数字相面术”的兴起与伦理隐忧

更前沿也更具争议的是 AI 面试与评估工具。一些供应商声称其 AI 能通过分析候选人面部表情、微动作、语音语调、用词等，量化评估“性格特质”、“软技能”、“文化契合度”甚至“工作潜力”<sup>[15]</sup>。这种做法的科学有效性受广泛质疑：

- 将外在表现与内在品质简单关联缺乏坚实科学依据<sup>[16]</sup>。表达方式在不同文化、性格、神经多样性个体间差异巨大，统一标准评判易生误解和歧视<sup>[17]</sup>。
- AI 解读易受无关因素干扰（网络质量、光线、背景噪音、身体状况、残疾、外貌等），导致评估结果信效度存疑。

批评者将此类技术比作现代“数字相面术”<sup>[18]</sup>或“情感计算”的不当应用<sup>[19]</sup>，认为其可能基于不可靠关联和刻板印象对候选人标签化排序，不仅导致错误决策，更侵犯候选人尊严，构成难以察觉、辩驳的新型歧视。

### 2.3 AI 招聘伦理困境的症结重重

AI 在招聘领域引发的核心伦理问题，其根源在于现有技术并不能完全掌握其所需要的技术逻辑与社会现实的复杂互动。

<sup>12</sup><https://www.aclu.org/press-releases/michigan-father-sues-detroit-police-department-wrongful-arrest-based-faulty-facial>

<sup>13</sup><https://www.aclu.org/news/privacy-technology/police-say-a-simple-warning-will-prevent-face-recognition-wrongful-arrests-thats-just-not-true>

<sup>14</sup><https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>

AI 系统隐性偏见的嵌入与系统性放大风险，特别是基于历史数据训练的简历筛选模型，极易学习并自动化过往招聘决策中存在的显性或隐性偏见。更令人担忧的是，AI 能够从看似中立的数据点中挖掘出与受保护特征（如性别、种族、社会经济背景）高度相关的代理变量<sup>[20]</sup>，从而实施一种更难被察觉、证明和纠正的间接歧视。这使得 AI 非但没能成为客观性的保证，反而可能沦为固化职场不平等的“数字看门人”<sup>[14]</sup>。

评估标准的科学有效性质疑与“伪科学”风险是另一大痛点，尤其体现在基于面部表情、语音语调或行为模式分析的 AI 面试与评估工具上。这类工具声称能够量化评估候选人的“软技能”、“性格特质”乃至“潜力”，但其底层逻辑往往缺乏坚实的心理科学或行为科学依据<sup>[16]</sup>，将复杂的人类特质与简单的外在表现进行粗糙关联。评估过程常常如同“黑箱”，其声称测量的构念的科学效度（validity）和信度（reliability）往往未经充分的独立验证<sup>[21]</sup>。这种近乎“数字相面术”<sup>[18]</sup>的做法，不仅可能因方法论缺陷导致错误的筛选决策，更因其对个体尊严的潜在侵犯和标签化风险而备受诟病<sup>[19]</sup>。

透明度的普遍缺失与申诉救济机制的失效、设备差异导致的加剧数字鸿沟与机会不平等……AI 招聘伦理困境的症结重重，距离“完全落地可用”道阻且长。

## 2.4 通用伦理原则的适用性困境

与 FRT 类似，现有的高阶 AI 伦理原则（如“公平”、“非歧视”、“透明度”）在应用于 AI 招聘这一具体场景时，同样面临着严峻的落地难题和适用性困境。例如，“公平性”应如何在招聘语境下被操作化定义和进行有效测试？是追求群体层面的统计均等，还是确保个体层面的无偏见决策？这两种目标在实践中可能相互冲突<sup>[22]</sup>。对于那些声称能评估软技能或性格特质的 AI 工具，如何建立一套可靠的标准来验证其科学效度和信度<sup>[21]</sup>？现有的反歧视法律框架如何才能有效应对算法决策带来的新型歧视形式，尤其是间接歧视和难以证明的歧视<sup>[23]</sup>？这些问题都指向了一个核心难点：通用原则必须转化为针对特定风险、可执行、可监督的具体规则，才能在 AI 招聘领域发挥实际的规范作用。

## 三、综合讨论与治理路径的反思

AI 在高风险场景的伦理风险具体、严峻且影响深远，仅从上文分析即可发现：

- 数据偏见根深蒂固及其通过算法的系统性放大<sup>[24]</sup>。
- 透明度与可解释性的普遍缺失<sup>[25]</sup>。
- 高阶伦理原则在实践中面临落地困境与价值冲突<sup>[26]</sup>。
- ……

仅依赖企业自律或宏观原则宣言，不足以有效应对。“伦理嵌入设计”（Ethics by Design / VSD）理念须从口号落实到具体工程实践<sup>[27,28]</sup>。这意味着 AI 系统全生命周期——从问题定义、数据收集、模型训练、风险评估，到部署后监控、评估、更新——都须系统性融入对伦理风险（隐私、公平、安全、人权）的识别、评估和管理。但这需开发推广具体方法论（公平性度量与干预技术<sup>[29]</sup>、可解释 AI<sup>[30]</sup>、隐私增强技术<sup>[31]</sup>）、工具箱，及建立有效跨学科协作机制<sup>[32]</sup>。

更重要的是，对 FRT 和 AI 招聘这类影响基本权利、可能造成严重社会后果的高风险 AI 应用，须建立强有力、具法律约束力、且针对特定场景的监管框架<sup>[33]</sup>。超越“软法”走向“硬法”规制，应采取的措施包括但不限于：

- 风险分级管理：借鉴欧盟 AI 法案思路<sup>15</sup>，对不同风险等级 AI 采差异化监管。
- 强制性要求：对高风险 AI 设严格上市（部署）前合格评定要求，包括强制性伦理与人权影响评估<sup>16</sup>、数据质量管理<sup>[34]</sup>、偏见检测与缓解<sup>[8]</sup>、透明度说明（如模型卡<sup>[35]</sup>）、人类监督机制、网络安全保障等。
- 强化个人权利：保障数据主体在 AI 环境下的各项权利，特别是知情、访问、更正、删除（被遗忘）、限制处理权，及对自动化决策提异议并要求人工干预的权利<sup>17[36]</sup>。
- 明确法律责任：建立清晰法律责任分配机制，明确开发者、提供者、部署者和使用者在发生损害时的责任，确保受害者能获有效救济<sup>[37,38]</sup>。
- ……

有效治理还需构建包容性、多主体参与的治理生态<sup>[39]</sup>。政府主导立法监管；开发者部署者承担伦理责任，践行负责任创新<sup>[40]</sup>；研究机构深化伦理风险研究；公民社会与公众积极参与监督讨论；受影响社群需被赋予话语权<sup>[41,42]</sup>。建立开

<sup>15</sup><https://artificialintelligenceact.eu/>

<sup>16</sup><https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

<sup>17</sup><https://gdpr-info.eu/art-22-gdpr/>



放对话平台，促进关于 AI 伦理边界、风险接受度和社会期望的广泛社会辩论，对形成治理共识、确保治理措施合法、合理、有效至关重要。

## 四、结论


鉴于 AI 技术快速迭代，治理体系本身须具动态适应性和前瞻性<sup>[43]</sup>。需建常态化风险监测、评估与预警机制，及时识别新技术新模式带来的新伦理问题（如深度伪造<sup>[44]</sup>、大型语言模型滥用<sup>[45]</sup>），并灵活调整监管策略。对 AI 伦理的前瞻性研究（Anticipatory Ethics）也应受重视，为应对未来更复杂挑战做准备<sup>[46]</sup>。

审视“智能”的边界，本质是审视技术、权力与人的价值关系<sup>[47]</sup>。这要求我们在拥抱 AI 机遇同时，对其在敏感领域应用保持高度警惕和持续批判反思。通过细致案例分析揭示风险，推动从抽象原则到具体规则落地，构建技术、法律、伦理协同的综合治理体系，并促进全社会广泛参与和深度对话，才能更有力引导 AI 技术朝着真正符合人类整体福祉、尊重个体尊严、促进社会公平正义的方向健康、可持续发展。在探索“智能”边界征途上，伦理指南针须始终在握，指引方向。

原创性声明

本文系原创，并对此负责。

## 参考文献

- [1] PARISER E. The filter bubble: What the Internet is hiding from you[M]. penguin UK, 2011.
- [2] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  [C/OL]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event, Canada: Association for Computing Machinery, 2021: 610-623. <https://doi.org/10.1145/3442188.3445922>. DOI:10.1145/3442188.3445922.
- [3] JOBIN A, IENCA M, VAYENA E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389-399.
- [4] NISSENBAUM H. Privacy as contextual integrity[J]. Wash. L. Rev., 2004, 79: 119.
- [5] RICHARDS N M. Intellectual privacy[J]. Tex. L. Rev., 2013, 87: 387.
- [6] ZUBOFF S. The age of surveillance capitalism: The fight for a human future at the new frontier of power[M]. Profile books, 2019.
- [7] SOLOVE D J. The digital person: Technology and privacy in the information age[M]. NYU Press, 2006.
- [8] BUOLAMWINI J, GEBRU T. Gender shades: Intersectional accuracy disparities in commercial gender classification[C]//Conference on fairness, accountability and transparency. 2018: 77-91.
- [9] BAROCAS S, SELBST A D. Big Data's Disparate Impact[C]//California Law Review: Vol. 104. California Law Review, Inc., 2016: 671-732.
- [10] CHOULDECHOVA A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments[J]. Big data, 2017, 5(2): 153-163.
- [11] AJUNWA I, GREENE D. Automated inequality: How high-tech hiring tools perpetuate bias[J]. Fordham Urb. LJ, 2017, 44: 1123.
- [12] HUNKENSCHROER A, LUETGE C. Understanding and mitigating unintended consequences of AI systems in hiring[J]. AI & SOCIETY, 2021: 1-14.
- [13] CHEN P C, OTHERS. Investigating Algorithmic Bias in Hiring: Evidence from a Field Experiment[J]. Journal of Labor Economics, 2018, 36(4): 853-888.
- [14] NOBLE S U. Algorithms of oppression: How search engines reinforce racism[M]. nyu Press, 2018.

- [15] RAGHAVAN M, BAROCAS S, KLEINBERG J, et al. Mitigating bias in algorithmic hiring: evaluating claims and practices[C/OL]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: Association for Computing Machinery, 2020: 469-481. <https://doi.org/10.1145/3351095.3372828>. DOI:10.1145/3351095.3372828.
- [16] ACQUISTI A, BRANDIMARTE L, LOEWENSTEIN G. Privacy and human behavior in the age of information[J]. *Science*, 2015, 347(6221): 509-514.
- [17] FAZELPOUR S, DANKS D. Algorithmic injustice: a relational ethics approach[J]. *Philosophy & Technology*, 2021, 34: 1153-1182.
- [18] STARK L, HUTSON J. Physiognomic artificial intelligence[J]. *Fordham Intell. Prop. Media & Ent. LJ*, 2021, 32: 922.
- [19] AJUNWA I. The limits of automating fairness[J]. *German LJ*, 2019, 20: 1122.
- [20] FELDMAN M, FRIEDLER S A, MOELLER J, et al. Certifying and removing disparate impact[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 259-268.
- [21] LEE N T, RESNICK P. We build race: How we helped create—and must help end—the racial wealth gap[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019: 1-15.
- [22] HARDT M, PRICE E, SREBRO N. Equality of opportunity in supervised learning[C]//Advances in neural information processing systems: Vol. 29. 2016.
- [23] BAROCAS S, SELBST A D. The problem of discrimination based on proxy variables[J]. *Big Data*, 2017, 5(3): 173-190.
- [24] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. *ACM Computing Surveys (CSUR)*, 2021, 54(6): 1-35.
- [25] MITTELSTADT B D, ALLO P, TADDEO M, et al. The ethics of algorithms: Mapping the debate[J]. *Big Data & Society*, 2016, 3(2): 2053951716679679.
- [26] FLORIDI L. Soft ethics, the governance of the digital and the General Data Protection Regulation[J]. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2018, 376(2133): 20180081.
- [27] FRIEDMAN B, KAHN JR P H, BORNING A. Value sensitive design and information systems[Z]//Human-computer interaction and management information systems: Foundations. ME Sharpe, 2006: 348-372.
- [28] HOVEN J Van den, VERMAAS P E, POEL I Van de. Designing for values in sociotechnical systems[Z]//Handbook of ethics, values, and technological design: Sources, theory, values and application domains. Springer, 2015: 67-96.
- [29] CORBETT-DAVIES S, PIERSON E, FELLER A, et al. Algorithmic decision making and the cost of fairness[C]//Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining. 2017: 797-806.
- [30] ADADI A, BERRADA M. Peeking Inside the Black Box: Explainable AI[J/OL]. *IEEE Access*, 2018, 6: 52138-52160. DOI:10.1109/ACCESS.2018.2870052.
- [31] DWORK C, ROTH A. The algorithmic foundations of differential privacy[M]. Now Publishers Inc, 2014.
- [32] WHITTAKER M, ALPER M, BENNETT C L, et al. AI Now Report 2018[R/OL]. (2018)[2025-04-11]. <https://ainowinstitute.org/publication/ai-now-2018-report>.
- [33] YEUNG K. Algorithmic regulation: A critical interrogation[J]. *Regulation & governance*, 2018, 12(4): 505-523.
- [34] GEBRU T, MORGENSTERN J, VECCHIONE B, et al. Datasheets for datasets[J]. *Communications of the ACM*, 2021, 64(12): 86-92.
- [35] MITCHELL M, WU S, ZALDIVAR A, et al. Model cards for model reporting[C]//Proceedings of the conference on fairness, accountability, and transparency. 2019: 220-229.
- [36] WACHTER S, MITTELSTADT B, FLORIDI L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation[J/OL]. *International Data Privacy Law*, 2017, 7(2): 76-99. <https://doi.org/10.1093/idpl/ix005>. DOI:10.1093/idpl/ix005.
- [37] CERKA P, GRIGIENĖ J, SIRBIKYTĖ G. Is it possible to hold robots liable? Artificial intelligence and the problems of legal personhood[J]. *Computer law & security review*, 2015, 31(3): 374-380.

- [38] Kingston J K. Artificial intelligence and legal liability[M]. Edward Elgar Publishing, 2018.
- [39] FLORIDI L, COWLS J, BELTRAMETTI M, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations[R/OL]. (2018)[2025-04-11]. <https://www.eismd.eu/wp-content/uploads/2018/11/AI4People-Ethical-Framework-for-a-Good-AI-Society.pdf>.
- [40] STAHL B C, TIMMERMANS J, FLICK C. Responsible research and innovation: The role of privacy in data science[J]. *International Journal of Medical Informatics*, 2017, 105: 43-50.
- [41] SLOANE M, MOSS E, SHAPIRO A, et al. Participation is not a design fix for machine learning[C]//*Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020: 11-21.
- [42] ANDERSSON SCHWARZ J. Participation in AI governance: A systematic literature review[J]. *AI & SOCIETY*, 2021: 1-18.
- [43] RAHWAN I, CEBRIAN M, OBRADOVICH N, et al. Machine behaviour[J]. *Nature*, 2019, 568(7753): 477-486.
- [44] WESTERLUND M. The emergence of deepfake technology: A review[J]. *Technology Innovation Management Review*, 2019, 9(11): 39-52.
- [45] WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from Language Models[R/OL]. (2021). <https://arxiv.org/abs/2112.04359>.
- [46] BORENSTEIN J, ARKIN R C. Seeing robots: Ethics of advanced perception[J]. *IEEE Technology and Society Magazine*, 2017, 36(4): 44-50.
- [47] WINNER L. Do artifacts have politics?[J]. *Daedalus*, 1980: 121-136.