中国科学: 信息科学 2024 年 第 54 卷 第 7 期: 1646–1676
SCIENTIA SINICA Informationis
论文

人工智能伦理计算

高漪澜[1,2], 张睿[1,2], 李学龙[1,2*]

1. 西北工业大学光电与智能研究院, 西安 710072
2. 智能交互与应用工业和信息化部重点实验室 (西北工业大学), 西安 710072
* 通信作者. E-mail: li@nwpu.edu.cn

收稿日期: 2023–03–21; 修回日期: 2023–08–05; 接受日期: 2023–10–11; 网络出版日期: 2024–07–11

国家重点研发计划 (批准号: 2022YFC2808000) 和国家自然科学基金 (批准号: 61871470, 62276213) 资助项目

摘要 人工智能技术作为试图研究、模仿、扩展人类智能的科学研究领域, 自诞生以来就伴随着深刻的技术伦理争辩. 随着近年来机器学习等相关工作的突破性进展和快速落地应用, 伦理问题日趋显著并迫使学界和社会开始直面该技术的伦理治理挑战. 尽管在伦理治理的规范研究上已取得初步进展, 其治理实践落地方面依然困难重重, 伦理实践表现出逐渐落后于技术发展需求的趋势. 因此, 建立与不断发展的人工智能技术相互匹配的伦理治理实践方案, 实现治理理论和治理实践的良性互动将是人工智能领域未来发展的关键问题. 伦理治理理论的抽象性导致了当下人工智能伦理原则难以落地实现, 人工智能伦理计算 (AI ethical computation) 将是应对这一挑战的重要方案. 本研究通过探讨现实必要性和发展可能性明确了伦理计算的重要意义, 在相关研究基础上给出伦理计算的研究范畴, 依据计算过程对伦理机理的认知程度和系统伦理决策的自主化程度进行划分, 建立了伦理计算的高阶认知与低阶认知两类研究范式, 并按其计算阶段抽象出伦理度量、伦理决策和伦理推理 3 个计算层次. 该伦理计算框架能够对当前的伦理计算应用进行梳理, 本文以伦理嵌入和公平机器学习为例说明了两类研究范式的研究特点和技术方法. 在此基础上, 进一步讨论构建了以伦理计算为核心的伦理治理体系, 分析通过伦理计算化解伦理治理困境的可能方案, 并对人工智能伦理计算的发展做出展望.

关键词 人工智能, 伦理问题, 伦理治理, 伦理计算, 伦理嵌入, 公平机器学习

1 引言

海量数据、基础算力、智能算法、专用硬件等多方面的提升促使人工智能产业蓬勃发展, 各类技术不断重塑人类的生活习惯和生产方式. 相关技术的应用广度不断扩展, 深度在不断加深, 催生了包括 AI for Science [1]、量子机器学习研究 [2] 等在内的各类交叉融合领域. 新技术层出不穷, 新应用不断

引用格式: 高漪澜, 张睿, 李学龙. 人工智能伦理计算. 中国科学: 信息科学, 2024, 54: 1646–1676, doi: 10.1360/SSI-2023-0076
Gao Y L, Zhang R, Li X L. Artificial intelligence ethical computation (in Chinese). Sci Sin Inform, 2024, 54: 1646–1676, doi: 10.1360/SSI-2023-0076

落地, 为人工智能领域的发展带来了前所未有的机遇, 但同时技术伦理焦虑也日渐增加, 技术的有效伦理治理迫在眉睫. 当前人工智能技术发展面临哪些伦理问题? 伦理治理的难点在哪里? 有哪些技术手段能够辅助伦理治理? 本文提出以人工智能伦理计算 (AI ethical computation, 后文简称伦理计算) 技术为核心, 构建伦理治理的技术工具体系. 而什么是伦理计算? 通过哪些技术进行伦理计算? 又该如何通过这一技术手段构建伦理治理体系? 后文将展开讨论. 从面临的伦理问题着手讨论, 以下述 3 个典型应用场景为例.

(1) 人工智能技术为医疗卫生领域带来了重大变革, 主要表现在两种形式 [3] 的技术支持上: 虚拟分支和物理分支. 虚拟分支是指人工智能算法可以利用大数据挖掘潜在的医疗辅助信息, 包括蛋白反应预测、药物预测、心理康复辅助治疗等. 物理分支则是指各类人工智能算法支持下的医疗服务机器人, 包括用于照顾危重病人的机器人伴侣, 外科手术 [4] 中的助理医生甚至主刀医师. 在这些关系到人类生命健康的重要应用场景中, 伦理问题格外突出. 其中涉及到如何保护涉众的隐私不受侵犯的问题 [5], 还涉及到如何确保决策不对受众的生命健康造成损害的难点, 例如在人工智能主刀医师的应用场景下, 手术意外的发生是否属于事故该如何界定? 而事故责任由谁承担? 这类问题牵涉了各类复杂的伦理主体, 其背后的伦理问题往往非常复杂, 有没有可能通过技术方法辅助界定这些复杂的伦理问题?

(2) 自动驾驶汽车有望提高交通运输效率, 减小交通事故的概率 [6]. 但是由于高度的自主化特性, 该场景也牵涉非常复杂的伦理问题: 自动驾驶机器如何做出道德决策, 如何判定事故责任? 社会对机器行为的指导伦理原则该如何量化? 这都成为了迫切需要回应的问题. 相关研究 [7] 针对该领域的重大挑战进行了一项广泛的社会实验, 收集了来自 233 个国家和地区的 4000 万条伦理决策并进行统计分析, 指出道德取舍从宏观上表现出了统一的趋势, 但其中依然存在道德决策的内在冲突、人际分歧、地区和文化差异等. 为了减少冲突明确差异, 这一领域也需要更多伦理对话的发生和更加确切的技术定量标准.

(3) 计算机辅助决策也是人工智能的重要应用之一, 算法对于海量数据的有效挖掘使其可以深入发掘历史决策并从中学习决策要素. 决策方式自动化程度的提高可以大大提高决策效率, 但其背后隐藏的偏见和歧视问题也令人担忧. 典型的场景包括用于累犯概率估计的 COPMAS 系统、招聘简历自动筛选、广告推送等 [8], 这些用于决策的人工智能会对社会和个人产生重大影响, 其决策机制的透明度和公平性亟待提高. 不仅仅是决策系统, 在自然语言处理 [9]、计算机视觉 [10,11] 等各类研究中, 都存在习得和放大历史偏见的问题, 如何改善这些问题还需要更多探索. 同时, 针对辅助决策场景提出的有关提案中也指出 [12], 人工智能应用于决策可能会导致组织文化和个人行为的改变, 因此还需要开发出切实可行的人工智能技术影响的演化指标, 以衡量其效益及对决策涉众的长短期影响.

上述是某些经典应用场景下存在的伦理问题, 新技术的应用在不断对伦理研究提出新挑战, 例如 AI 绘画、ChatGPT 为代表的生成大模型在近期就引发了对版权纠纷和剽窃问题 [13] 的探讨, 后文在探讨伦理计算进展的基础上, 也将对这类大模型引发的问题进行简要讨论. 不难看出, 当前人工智能技术引发了广泛深刻的社会问题, 随着技术发展也将面临更加错综复杂的技术伦理困境, 因此迫切需要给出有效的应对方案.

事实上, 对人工智能伦理问题的探讨 [14,15] 早在 1960 年就已经产生, 人工智能伦理研究伴随着技术的发展. 然而, 早期研究主要围绕伦理理论展开, 脱离具体应用场景且高度抽象. 因此, 如何在实际的应用场景中有效地考量伦理因素, 构建出切实可行的伦理治理方案就是重要的研究课题. 针对伦理治理问题, 近年来, 各个国家组织都做出了重要努力, 提出了包括可解释、公平、隐私等在内的诸多技术伦理诉求, 制定了相关行业发展规范.

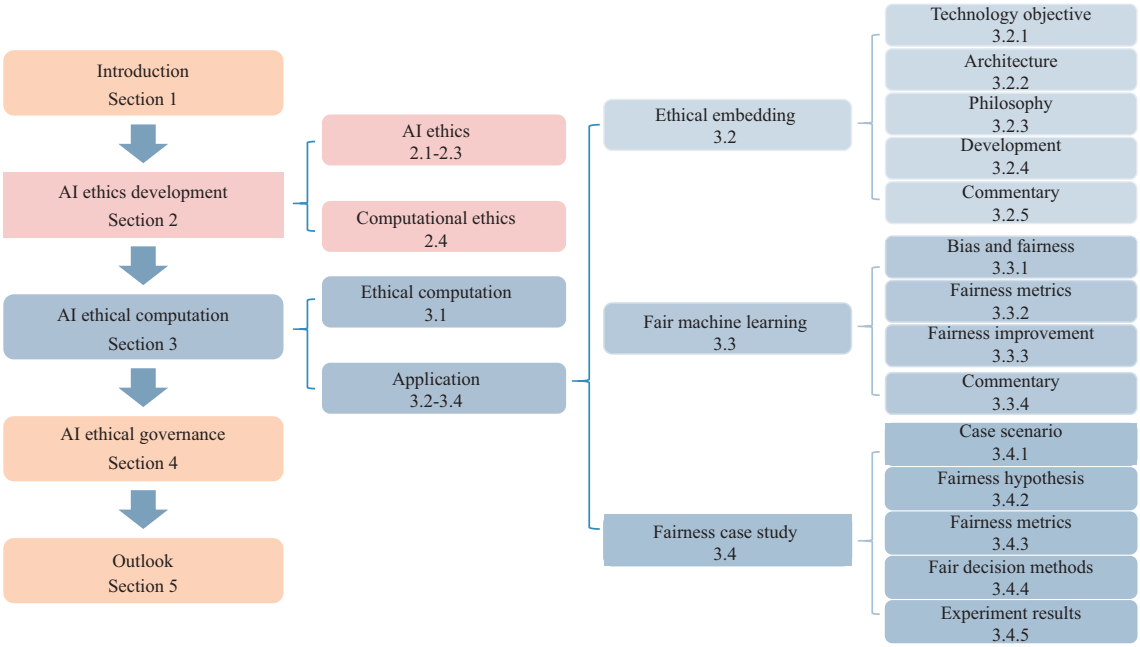

图 1 (网络版彩图) 论文结构图

Figure 1 (Color online) Paper structure diagram

尽管提出了诸多法律规范和倡议, 但由于这些抽象指标的模糊性和差异性, 伦理治理的实践依然面临着诸多困难. 本文认为将伦理规范具象化, 将伦理诉求量化并为伦理规范提供技术支撑是破局的关键, 即通过伦理计算消除抽象歧义, 提供量化特征, 有望弥补理论和实践的差异.

然而, 伦理计算面临的关键问题在于: 抽象的伦理概念是否能够计算? 如何进行计算? 计算又应该如何服务于伦理治理实践? 本文拟从人工智能伦理研究的发展现状和面临困境入手, 指出伦理计算的研究必要性. 通过探讨伦理可计算性的研究发展, 讨论伦理计算的可能性并界定伦理计算的概念. 进一步针对如何计算的问题, 总结伦理计算的已有计算方法和计算范式, 并对代表性伦理计算技术进行举例. 最后也将对计算与伦理治理的关系问题进行分析.

整体论述思路如图 1 所示, 完成了伦理计算从计算体系到伦理治理体系的说明. 具体地, 第 2 节总结探讨了伦理计算研究的现实背景和计算历史. 阐述了当前伦理研究发展的状况, 指出伦理治理中伦理理论模糊、实践可行性受限的发展痛点, 点明伦理计算的重要现实意义, 在此基础上对计算伦理学的研究背景进行了分析. 第 3 节给出了本文伦理计算的具体定义, 对伦理计算的研究目标、研究方法和研究范式进行了总结, 同时给出了两类代表性的工作, 并以公平机器学习作为案例给出伦理计算的示例. 第 4 节构建了基于伦理计算的伦理治理体系, 最后在第 5 节对人工智能伦理计算和伦理治理进行了展望.

2 人工智能伦理发展现状

2.1~2.3 小节将从人工智能伦理研究的现状出发, 引出当前伦理治理中存在的难点并为伦理计算的现实必要性提供论据. 进而在 2.4 小节回顾伦理的可计算性发展, 为其计算提供背景支撑.

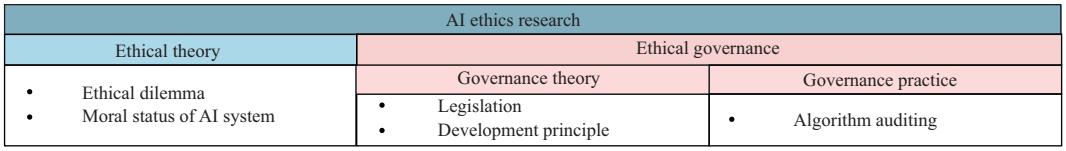

图 2 (网络版彩图) AI 伦理研究

Figure 2 (Color online) AI ethics research

2.1 伦理研究现状概述

人工智能研究希望模仿并扩展人类智能, 发展出了包括具身智能、专家系统、机器学习等诸多研究领域 [16, 17]. 作为对人类智能的解放, 这一领域的发展必定会对人的生存状态产生巨大的影响, 相应的技术伦理探讨是引导学科健康发展的关键. 人工智能技术的伦理研究 (AI ethics research) 主要探讨技术所应遵循的社会道德规范和要求, 探索算法技术与人和社会的关系, 关注技术应用对于社会道德秩序的影响, 协调技术和各类涉众的伦理诉求.

人工智能技术的特殊性在于, 其伦理探讨不仅包括了技术涉众使用和构建这项技术时面临的伦理规范, 还包括了构建出的系统本身是如何遵循伦理规范的, 以及这类特殊的智能机器是否应该又应该如何遵循某些伦理要求 [18]. 相关研究与认知心理学、哲学、法学、计算机等诸多学科都有紧密联系, 它可以视作技术伦理哲学、工程伦理等学科的延伸课题 [19], 同时也与算法伦理、数字伦理、机器伦理等有关概念交叉融合.

伦理研究发展至今, 本文按照如图 2 所示的框架对相关研究进行划分. 按照发展阶段、研究特点大致划分为伦理理论 (ethical theory) 和伦理治理 (ethical governance) 两方面, 伦理治理进一步划分为治理理论 (governance theory) 和治理实践 (governance practice) 两部分. 本小节对伦理理论和伦理治理研究内容进行简要总结, 并将在 2.2 和 2.3 小节对伦理治理的研究划分展开说明.

早期的伦理研究主要围绕伦理理论展开, 建立了技术伦理理论, 并针对人工智能伦理相关的哲学基础性论题进行辩论. 例如, 在技术能力倾向于未来主义的视角下, 对智能的实现方式、机器伦理地位 [20] 等问题进行广泛的思辨. 同时, 在考量技术应用时, 针对某些可能的伦理困境进行了探讨 [7]. 这些工作是伴随着人工智能技术的诞生而诞生的, 作为伦理研究发展的基础, 涵盖了哲学、法学等对核心伦理概念、基本问题的理论探索, 为后续的伦理治理提供了理论支撑.

伦理治理是在上述研究基础上, 关注如何将抽象理论逐步落地, 切实预防和解决技术应用出现的各类伦理问题, 这是技术发展的必然要求. 本文将伦理治理分为治理理论和治理实践两部分, 治理理论需要在哲学伦理研究基础上尽可能协调各方伦理要素, 形成具体治理场景和社会条件下的宏观治理依据, 例如法律、原则的制定. 治理实践是基于这些规范的实践工作, 包括管理、技术的落地.

目前在治理依据的探索上已初见成效, 逐步形成共识性的治理规范, 但在实践工作中依然存在规范模糊、落地困难的问题. 后文首先对治理理论的发展状况进行总结, 进而分析实践中面临的困境.

2.2 伦理治理理论发展

针对近年来人工智能技术实践中尖锐突出的伦理矛盾, 许多国家和组织提出了宏观治理依据. 相关工作包括对于人工智能伦理发展宏观框架的讨论或者围绕伦理治理提出的一系列抽象、第一性发展原则及倡议, 也包括更具强制性的法律法规.

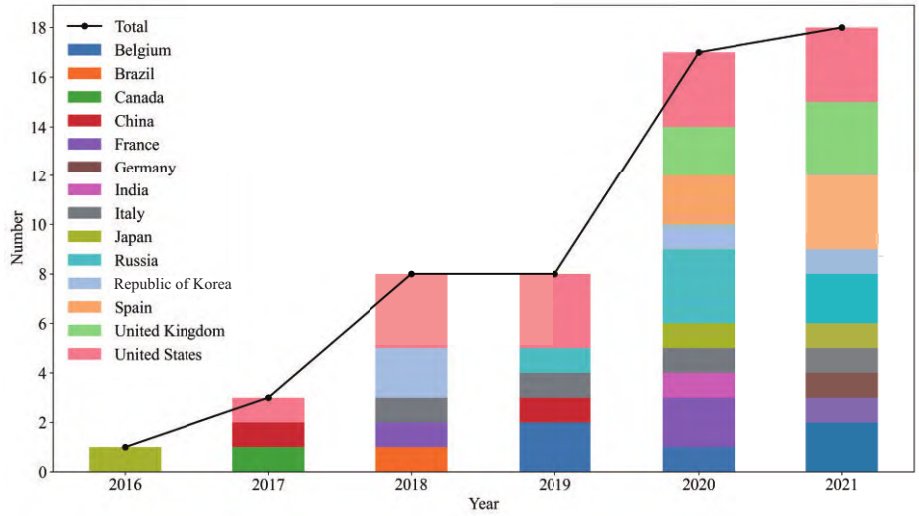

图 3　(网络版彩图) 2016~2021 年各国 AI 有关的法律数目 (数据来自斯坦福《人工智能指数报告 2022》)

Figure 3　(Color online) AI related legislations in 2016~2021 (data from Stanford AI index report 2022)

联合国教科文组织于 2021 年 12 月发布的《人工智能伦理问题建议书》[1)] 就属于这一范畴的工作, 主要用于规范人工智能伦理的发展. 同时各个国家也积极参与到人工智能治理的讨论中来, 图 3 是斯坦福大学 (Stanford University) 发布的《人工智能指数报告 2022》[2)] 中对 2016~2021 年间各国通过的法律中与 AI 有关的法律数量统计, 相关统计数据呈现逐年上升的趋势, 可以认为, 自 2016 年以来, 各国对 AI 相关的政策和法律制定越发重视. 同时, 相关技术机构也高度关注技术伦理问题, 积极促进人工智能技术的伦理对话, 例如 AAAI 与 ACM 合作组织的人工智能、伦理和社会会议 (Artificial Intelligence, Ethics, and Society, AIES), 提供了伦理探讨的跨学科交流平台.

我国也非常重视人工智能伦理的发展, 2021 年发布的《人工智能标准化白皮书》[3)] 中指出, 人工智能的伦理问题作为学科发展的重要挑战, 受到人类伦理认知滞后的影响, 同时由于未经伦理管控的产品在当下获得了过高的自主权, 又加剧了问题的产生. 同年发布的《新一代人工智能伦理规范》[4)] 中更是着重对人工智能的伦理问题做出指示, 强调将伦理道德融入人工智能发展的全生命周期. 2022 年 3 月出台的《关于加强科技伦理治理的意见》[5)] 中, 也强调了需要对人工智能等重点领域, 加强监管、明确责任, 建立更加完善的科技伦理规范、制度. 因此, 可以说当下人工智能的伦理探讨已经成为该学科重要的发展议题.

同时, 针对具体的应用领域也建立了针对性的伦理规范. 例如医疗卫生领域这类高风险场景, 世界卫生组织就于 2021 年发布了第 1 份卫生领域人工智能的全球报告和六项设计指导原则, 提出在其使用中充分保障人类自主权、增进人类福祉和安全利益等指导要求.

上述诸多治理理论都在试图厘清社会对人工智能技术的伦理要求, 从抽象的伦理理论中寻找与实践需求的统一, 图 4 中列举了主要的伦理原则. Jobin 等[21] 在 2019 年的研究工作中筛选并分析了 84 份来自各个私营机构和公共部门的人工智能伦理规范. 研究结果显示, 全球在五大伦理原则 (透明度

1) https://zh.unesco.org/artificial-intelligence/ethics.
2) https://aiindex.stanford.edu/report/.
3) http://www.cesi.cn/images/editor/20210721/20210721160350880.pdf.
4) http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
5) http://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

图 4 (网络版彩图) 人工智能伦理原则

Figure 4 (Color online) AI ethical principles

(transparency)、公正和公平 (justice and fairness)、不伤害原则 (non-maleficence)、责任 (responsibility) 和隐私 (privacy)) 方面出现统一的趋向, 相关概念在所有资料中有超过一半被引用, 但在如何解释这些原则、应该如何实施等问题上依然存在分歧. 其他研究 [22] 也得出类似的结论, 该研究中确定了由 5 个核心原则组成的伦理框架, 其中 4 条与生命伦理学中常用的核心原则一致: 善意 (beneficence)、不伤害原则 (non-maleficence)、自主性 (autonomy), 以及公平正义 (justice), 最后增加了可解释性 (explicability).

上述工作说明, 伦理治理的理论研究取得了重大的进步, 在复杂的伦理关注背后, 依然有可能形成相对统一的伦理框架. 随着这些规则规范的提出和共识性依据的建立, 研究人员也逐渐意识到伦理治理在实践落地上的困难.

2.3 伦理治理实践困境

伦理治理实践上人工智能伦理研究还面临困境, 主要可以概括如下.

(1) 统一性背后的潜在差异. 虽然当前的伦理治理理论依据展现出了趋同的倾向, 相关研究 [21] 指出, 各个原则的内涵阐释上规范之间依然存在分歧. 本质上这是规范描述的模糊性和自然语言的不确定性所导致的, 例如在算法决策的公平性上, 如何理解决策结果或者决策辅助数据的公平性是一个重要问题, 如果不给出相对统一、量化的规范基础, 伦理治理的实践困难是显然的.

(2) 规范性实践的技术脱节. 抛开伦理规范之间的潜在差异, 单就如何实施某些原则已经存在困难. 相关研究 [23] 针对伦理规范的实践效果进行了考察, 结果表明, 目前围绕伦理问题提出的诸多发展框架和意见 [24] 都主要关注制度和管理规范层面, 由于缺乏技术对应, 伦理理论、规范往往和实践工作脱节, 使得具体落地时困难重重. 可以说, 伦理实践缺乏有效的衡量标准和技术路线大大限制了伦理规范的效用.

为了尽可能减少模糊的伦理规范表达、解决治理理论到实践的技术脱节问题, 对人工智能伦理的适当量化抽象、提供可行的标准化伦理计算方案是必要的. 以前述的典型应用场景为例, 在医疗卫生这些关系到人类生命健康的重要场景中, 如果能够量化评估这类应用系统对人类心理、身体状况的副作用和效用, 对其决策过程进行有效追踪记录, 则能够实现人工智能技术的高效管控. 而在自动驾驶场景下, 如何处理复杂的伦理困境, 能否建造道德自主的机器也是伦理量化计算所关注的高阶目标. 在计算机辅助决策过程中, 可以通过量化决策系统对社会决策的影响, 建立切实可行的指标以衡量其决策效益, 进而在对应场景中选择合适的决策干预方案和程度. 这些思路的背后都反映了伦理计算的需求, 即对伦理及相关抽象规范提供具体的计算技术支持. 可以说, 要实现可信人工智能的目标, 构建对人类社会有积极作用的技术系统, 伦理计算技术是必不可少的.

2.4 可计算研究发展

上文提到, 理论内涵模糊、理论与实践之间的缺乏关联的现状限制了伦理治理实践. 可见实现人工智能系统对伦理道德因素的考量和实践非常重要, 这也回答了伦理计算研究的现实必要性. 但伦理这一抽象概念是否能够计算? 应该如何对算法的伦理影响进行度量? 机器是否能够具备伦理机制? 也是伦理计算研究必须面对的问题, 后文将回顾伦理可计算性的相关研究.

事实上, 人工智能是对人类感知、思维、学习等智能行为进行计算、模拟的技术, 人类情感、伦理道德则是更复杂的智能行为, 能否使得机器、算法表现出此类高级智能行为也一直是研究人员关心的问题. 对情感的计算关注产生了情感计算研究 [25], 对伦理的计算探索也催生出了诸多课题. 同时, 如何计算和度量伦理一直以来也是哲学、伦理学积极探索的问题, Moor [26] 在 20 世纪 90 年代的研究中就指出, 对于伦理抽象概念的可计算性探讨可以向前追溯到 18 世纪. 人工智能伦理的计算思路也有诸多探索和尝试, 提出了机器伦理 (machine ethics)、计算伦理学 (computational ethics) 的研究范畴 [27]. 这些可计算工作的发展对伦理及其计算方法进行了初步探索, 回应了伦理的可计算性, 并为本文人工智能伦理计算框架的提出提供了重要条件, 下面简要说明二者的研究内容.

(1) 机器伦理是一类将智能机器作为主体的伦理学, 该概念本身就存在诸多界定, 它与人工智能伦理之间的包含关系还存在争论, 本研究中将其作为人工智能伦理研究的子课题. 机器伦理关注的问题包括两方面: 一方面, 人工智能系统作为自治系统是否可能进行伦理推理、做出伦理决策, 以及如何做出伦理决策 [28], 另一方面, 还需要明确机器是否应该、又在什么情况下可以被视为伦理代理 (artificial moral agent) [27]. 该领域中的伦理嵌入 (ethical embedding) 研究希望在机器决策中嵌入伦理维度, 实现伦理推理. 相关工作由 Allen 等 [29], Asaro [18], Moor [30], Anderson 等 [28] 积极推动, 在计算的视角下, 构建类似人类道德决策的模型以约束计算机 [29], 其发展也促使研究人员进一步探索伦理决策、推理的机制及其与情感偏好等因素的深层关系. 因此, 该领域初步探索了伦理的可计算性, 提出了部分可能的计算方案.

(2) 计算伦理学与机器伦理在早期诸多研究中被视作同一范畴 [31], 存在研究主张对这二者进行区分. 2021 年 Segun [27] 的研究中首次明确了机器伦理和计算伦理学的交集, 他指出计算伦理学主要关注计算可能性和实验方案, 并主张将计算伦理学作为人工智能伦理学的研究子集与机器伦理做以区分, 将试图在人工智能系统中建立人工道德代理、对伦理进行计算并模拟机器意识的工作划归为计算伦理学. 后续 Awad 等 [32] 在认知科学的框架下也对计算伦理学给出了新的定义: 试图将描述性伦理 (descriptive ethics) 和规范性伦理 (normative ethics) 用算法表示, 并用于构建符合伦理的人工智能系统 (ethical AI systems) 或者更好地理解人类决策的相关研究. 同时, 给出伦理与计算二者相互协调的计算反思均衡框架 (computational reflective equilibrium) [32].

上述以可计算伦理研究为代表的工作论证了伦理量化、算法化的必要性, 并以伦理为核心对其可计算性质、某些计算方法进行了探讨. 然而, 当前的研究缺少了对计算为核心的伦理量化计算方法的探讨. 在此基础上, 本文将以计算为核心, 明确伦理计算的概念, 依据伦理计算方法的伦理认知程度、伦理决策自主化程度建立伦理计算的分类策略, 划分伦理计算的两类研究范式并整合已有工作, 构建伦理计算技术的一种分类研究框架.

3 人工智能伦理计算

本节将在 3.1 小节首先明确伦理计算研究范畴, 给出当前的两类研究范式和对应研究层次, 进而在 3.2 和 3.3 小节针两类研究范式分别以伦理嵌入和公平机器学习进行应用举例.

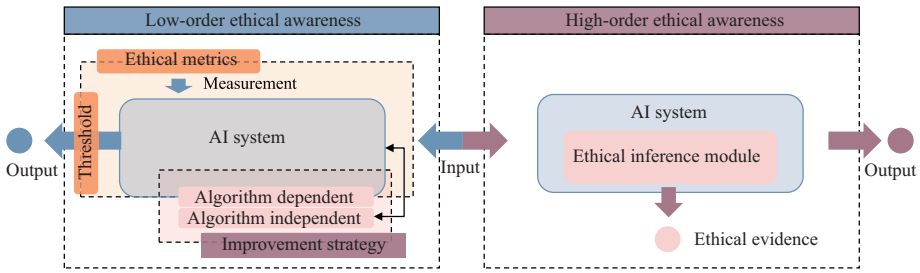

图 5 (网络版彩图) 两类伦理计算范式

Figure 5 (Color online) AI ethical computation paradigms

3.1 伦理计算

本文以计算为核心, 对伦理计算给出如下定义.

定义1 (伦理计算) 伦理计算是指通过对伦理原则进行合理的数学符号化或对伦理过程算法化完成伦理概念的定量描述、度量或模拟, 并在此基础上构建对智能体或算法的伦理约束, 使之遵循特定社会背景下的伦理原则.

伦理计算是针对抽象伦理概念进行的量化、算法化技术研究, 包括了各类对人工智能伦理的模拟、量化、计算尝试, 同时通过伦理计算也能够进一步探索伦理这一智能行为. 伦理计算的最终目标也在于实现计算机决策、行为符合社会伦理约束, 构建安全、可靠、增进社会福祉的智能系统.

本文按照伦理认知程度和伦理决策自主化程度的不同, 将伦理计算的研究划分两类研究范式和 3 个相互关联的层次. 如图 5 为高阶伦理认知 (high-order ethical awareness) 和低阶伦理认知 (low-order ethical awareness) 两类研究范式, 人工智能系统对输入信息进行处理得到决策输入, 不同的范式在处理过程中有所差异, 高阶认知系统通过将伦理机制算法化, 借助系统内部的伦理推理机制构建符合伦理的输出决策, 此类研究通常关注高度自主化的系统, 而低阶认知系统则通过构建外界度量约束来得到的符合伦理要求的决策输出, 具体说明如下:

(1) 高阶伦理认知的伦理计算. 此类研究范式以早期的伦理代理研究为代表, 通常更加关注伦理的产生和实践机制, 包括伦理决策推理、情感因素考量等. 通过在人工智能系统中构建伦理推理模块 (ethical inference module), 让计算机模仿人类道德决策产生的机理, 实现机器对伦理概念的认知和实践. 此类计算方法要求具有更高层次的伦理认知, 系统也具有更高的伦理决策自主化程度, 计算系统可以完整进行道德决策的全过程, 如图 5 所示, 有时也要求系统能够为其道德决策提供依据 (ethical evdience).

(2) 低阶伦理认知的伦理计算. 此类研究范式需要针对某些伦理诉求进行度量和约束优化. 这类研究通过建立伦理的度量方法 (ethical metrics), 在以一个或一组值定义抽象伦理概念的基础上, 进一步通过提供优化策略 (improvement strategy) 来约束计算机决策结果, 使之在度量方案上达到目标阈值 (threshold), 实现决策融合相应社会背景下的伦理诉求. 此时的伦理计算方案就不要求系统对伦理机制具有充分的认识, 其伦理决策也是在外界评估因素和约束方案的基础上完成的优化, 伦理决策自主化程度较低.

根据上述两类范式的计算阶段, 大致可以将伦理计算的决策抽象为 3 个层次: 伦理度量 (ethical metrics)、伦理推理 (ethical inference)、伦理决策 (ethical decision-making). 每个层次的具体说明如下.

(1) 伦理度量. 该阶段的目标在于构建对于伦理概念的度量方法, 以减少抽象伦理概念可能带来

表 1 伦理计算小结

Table 1 Ethical computation summary

Research paradigm	Computing hierarchy	Typical application	Mechanism of ethical cognition	Degree of autonomy in ethical decision-making
High-order ethical awareness	Ethical metrics, ethical inference, ethical decision-making	Ethical embedding	More	More
Low-order ethical awareness	Ethical metrics, ethical decision-making	Fair machine learning	Less	Less

的歧义. 具体量化方式可能有多种选择, 例如将伦理定义数学符号化, 提供一个或一组值进行表征, 或者建立相关量化算法等. 这一度量过程可以完成对计算系统中算法公平、隐私等原则实现程度的量化表示. 其典型应用是算法审计, 审计需要判断人工智能算法对于具体伦理原则的实现程度, 伦理度量就为此提供了客观的评估手段, 恰当的度量指标 [33] 是必要的.

(2) 伦理推理. 伦理推理则是将伦理作为系统内在要求, 系统需要通过符合伦理决策机理的推理过程做出决策. 可计算研究中的伦理嵌入工作对这一高阶目标进行了探索, 这一技术层次要求对人类伦理决策过程有更深入的理解, 包括其背后的逻辑与情感因素等等. 在某些高度自主化系统 (如自动驾驶系统) 中, 对这一层次的伦理计算讨论是必要的. 因此, 发展伦理推理研究对于更好地理解伦理本身、发展自动驾驶技术 [30] 等都至关重要.

(3) 伦理决策. 伦理决策的目标则希望在进一步明确系统对伦理的需求程度或保障目标后, 通过计算约束等方式对系统决策进行改进, 使得其满足目标并完成伦理决策. 这一层次可以视作伦理度量、推理层次的后继环节, 这部分的关键在于, 如何对某个或某些已经量化的伦理目标进行针对性改进, 提供可接受的伦理决策结果. 这一过程只关心最终决策的结果是否符合伦理目标, 对伦理决策背后的决策机理不做要求.

这 3 个层次要求在不同阶段寻求将抽象道德原则、理论转化为实践的量化方法.

综合上述探讨, 表 1 是对伦理计算框架核心特征的整合, 伦理计算方法能够减少伦理理论中的定义模糊性, 有望实现理论与实践的良好对应.

上文给出的伦理计算框架可以囊括诸多实践中的典型应用, 后文将依据伦理计算的两类研究范式, 分别以伦理嵌入和公平机器学习为例对伦理计算的应用做出说明.

3.2 高阶认知: 伦理嵌入

伦理计算中的高阶认知计算探索人类伦理决策的形成机理, 通过伦理推理、情感偏好考量等策略实现伦理. 其中最经典的工作就是伦理嵌入, 这一领域作为对系统内在伦理计算的探索, 从哲学探讨到技术尝试经历了漫长曲折的发展, 对于高自主性系统的伦理实践具有重要意义, 包括自动驾驶、手术护理机器人等. 后文主要针对伦理嵌入的含义与方法架构、近期进展和存在的问题进行说明.

3.2.1 技术目标

伦理嵌入研究试图将伦理纳入机器的决策维度, 目标在于构建能够模拟人类的伦理机制、实现伦理推理的伦理代理. Moor [30] 区分了 4 种机器代理 (agents), 分别是道德影响代理 (ethical impact agents)、隐式道德代理 (implicit ethical agents)、显式道德代理 (explicit ethical agents) 和完全道德代理 (full ethical agents).

具体来说, 道德影响代理指: 计算机作为代理执行命令, 必然会导致某些伦理问题, 因此将能够评

表 2 伦理嵌入策略

Table 2 Ethical embedding strategy

Perspective	Bottom up	Top down
Philosophy	Learning ethical experiences	Following ethical guidelines
Engineering	Improve by iteration	Divide and conquer

估其决策的伦理影响的机器定义为道德影响代理. 进一步, 隐式道德代理则会限制机器的行为, 以避免不道德的结果. 因此在隐式道德代理中, 可以通过创建隐含支持道德行为的软件来满足机器道德, 而不是编写支持明确道德准则的代码. 上述两种概念的实现可以通过低阶认知的伦理计算达成, 即伦理计算中的度量和决策层次实现.

显式道德代理是伦理嵌入构建的核心目标, 显式道德代理需要建立对伦理道德的明确表达并进行分析、推理, 即要求人工智能系统能够使用伦理原则在伦理困境中计算出最优的决策行动 [30]. 自动驾驶汽车和护理机器人就属于这类伦理代理, 显式代理的自主性不仅来自于独立做出道德决策的能力, 还来自于适应环境、优化效率、从一系列的道德可能性中推理判断的能力, 因此伦理嵌入工作需要通过计算进行伦理推理决策. 完全道德代理则是更加未来主义的观点, 这要求机器不仅能够做出明确的道德判断和推理, 还需要自主提供决策合理性的证明, 类似于人类行为者 [30].

3.2.2 方法架构

伦理嵌入作为高阶认知的伦理计算, 对伦理行为的形成机制高度关注, 从实现架构上提出了 3 种策略, 分别是自上而下 (top-down)、自下而上 (bottom-up) 和混合 (hybrid) 方式 [34]. 其中混合策略是前两者的结合, 因此首先需要对前两者的区别进行分析, 自上而下与自下而上策略的主要区别如表 2 所示.

伦理意义上, 自上而下的方法是指采用预先指定的伦理理论分析相应的计算需求, 进而构建出实现对应理论的算法和子系统 [34], 类似于人类的行为遵循社会演化生成的诸多既有伦理规范. 工程意义上, 它试图通过一组由规则表示的特定道德理论算法来规范人工道德主体的行为, 因此自上而下的方法可以将规则、任务进一步分解为更简单的子任务 [35]. 其困难在于, 需要解决 “哪个道德理论是正确理论” 这一哲学问题, 3.2.3 小节将对相关哲学观点进行介绍. 而且, 自上而下的推理过程即使建立了统一的规则, 计算的处理也面临一定困难.

自下而上的方法则不要求硬性的伦理规则制定, 而是试图通过外部环境来形成道德规范, 类似于人类从既往经验中学习. 在这种条件下, 道德行为可以通过经验学习或其他方式获得, 而不需要编码一套明确的、通用的道德规则. 工程意义上, 这类任务就可以通过性能度量逐步调整系统达成 [35]. 这种建模方法可以在不解决 “哪个道德理论是正确理论” 这一困难哲学问题的情况下实现, 但是相比自上而下就缺少可靠性.

自上而下的方法强调外部规则和控制, 而自下而上的方法则更多地关注价值与因果关系的学习. Allen 也在研究中指出 [34], 二者混合的方法能够在一定程度上带来综合的优势, 但决策中引发的冲突也可能导致新的问题.

3.2.3 哲学伦理基础

不论采取何种策略和分类标准, 哲学伦理理论对这一领域的重要影响是显然的, 尤其是规范伦理学 (研究道德决策的原则、机理, 即为何做出某些道德决策的动因). 因此在总结具体计算方法前, 首先

介绍基础的哲学伦理学发展及其符号化. 在机器伦理研究中主要关注的哲学观点有 3 类, 分别是结果主义伦理学 (consequentialism)、义务伦理学 (deontology) 和美德伦理 (virtue ethics) [36]. 伦理嵌入方法往往基于一种或几种伦规范伦理假设对伦理决策形成过程进行推理和计算.

道德决策的基本要素就是道德主体 x, 以及道德行为 $a_i \in \mathcal{A}$、决策背景 $c_i \in \mathcal{C}$ 集合和决策后果函数 $R(a, c)$. 此时智能体需要在决策背景等信息上进行决策后果的判断和道德决策.

(1) 功利主义学. 权衡每种选择的后果并选择最大道德收益结果的选择, 最终的决策往往以产生最佳的综合结果为目标. 即功利主义仅仅关注结果带来的收益, 其决策序列需要在已有决策背景下, 最优化决策的道德收益函数 P_e. 即

$$\max P_e(R(a_1, c_1), R(a_2, c_2), \ldots, R(a_n, c_n)), \tag{1}$$

其中, 决策收益通过对一系列决策序列及其决策背景 $\{(a_1, c_1), (a_2, c_2), \ldots, (a_n, c_n)\}$ 所对应的决策后果进行考察, 判断最优的决策序列. 因此, 功利主义可以通过约束优化等策略实现伦理, 但事实上, 决策时往往并非所有信息都是准确的, 此时就涉及到在概率意义下优化决策结果, 也会涉及到贝叶斯 (Bayes) 因果推理相关研究 [37].

(2) 义务伦理学. 强调决策者尊重特定条件下的义务和权利, 此时的行为主体会倾向于按照既定社会规范行事, 即基于一组规则. 在计算量化中可能会涉及到逻辑规范的表达或者某些规则约束, 即在某些决策背景下完成某些行为集合序列:

$$c_1 \wedge c_2 \Rightarrow a_1 \wedge a_2. \tag{2}$$

(3) 美德伦理学. 要求决策者根据某些道德价值来行动和思考, 同时, 具有美德的行为主体会表现出一种被他人认可的内在动力. 这背后的基本思想是, 品格高于行为, 良好的品格会导致良好行为的产生 [29]. 这一规范伦理理论不同于优化结果的功利主义或者遵守规则的义务伦理学, 而是更加偏向于从实践中学习. 在计算中则需要根据某些经验集合: $E = [a_h, x_h, c_h, R]$, $h = 1, \ldots, n$, 从这些经验数据中进行学习. 这一学科利用了描述性伦理 (研究人类的伦理决策, 并不其做出评价) 的经验结果, 同时天然地与当前各类数据挖掘、学习算法 [37] 存在紧密的联系.

上述仅仅针对单一系统决策进行了量化, 当实际中存在多主体交互的伦理决策场景, 就需要基于博弈论对多主体 $x_1, \ldots, x_n$ 的行为进行建模. 近期有研究针对上述伦理理论的相关计算方法进行了复杂度分析, 提供了伦理嵌入研究新的视角 [37].

3.2.4 发展现状总结

基于上述嵌入架构和哲学观点对伦理嵌入研究的伦理计算技术进行梳理, 表 3 [31, 38~49] 是对主要工作的研究任务、架构、技术和哲学目标的整合. 从伦理研究任务分类, 本研究将其划分为两类, 一类属于道德顾问 (ethics advisor) 或伦理困境分析器 (dilemma analyzer), 另一类是构建伦理嵌入模型 (ethical embedding model)、框架 (ethical embedding framework), 前者主要关注了伦理计算中伦理的判断和认知, 后者则关注构建伦理嵌入的计算架构. 从技术策略的视角分类, 本研究将其分为基于符号规则的推理方式、基于数据驱动的统计方式、混合模型 [50] 以及模拟仿真策略, 涉及技术包括归纳逻辑编程、博弈论、强化学习等.

道德顾问或伦理困境分析器作为早期对伦理计算的探索, 主要以 Anderson 等 [42, 47] 的研究为代表. 这类工作探索了伦理困境的表征和推理, 近年 Conitzer 等 [48] 提出的基于博弈论的框架为该问题提供了新的思路. 另外, 诸多研究试图建立伦理嵌入模型或者框架. 一方面, 部分工作受到心理学等学

表 3 伦理嵌入工作整合

Table 3 Ethical embedding researches

Research name	Research task	Architecture	Technology	Philosophy theory
MoralDM [38]	Ethical embedding model or framework	Top down	Symbolic and logic-based	Consequentialism and deontology
JEREMY [39]	Ethics advisor or dilemma analyzer	Top down	Symbolic and logic-based	Consequentialism and deontology
Ethical robots [40]	Ethical embedding model or framework	Top down	Simulation	Consequentialism
CP-Nets for ethical decision [41]	Ethical embedding model or framework	Top down	Symbolic and logic-based	Other (preferences and rules)
GenEth [42]	Ethics advisor or dilemma analyzer	Bottom up	Symbolic and logic-based	Other
Ethics shaping [43]	Ethical embedding model or framework	Bottom up	Data-driven (reinforcement learning)	Other
Motivated value selection [44]	Ethical embedding model or framework	Bottom up	Data-driven	Consequentialism
Reinforcement learning ethical decision-making framework [49]	Ethical embedding model or framework	Bottom up	Data-driven (reinforcement learning)	Consequentialism
Casuist BDI-Agent [45]	Ethical embedding model or framework	Hybrid	Combining rule-based and data-driven	Consequentialism
LIDA [46]	Ethical embedding model or framework	Hybrid	Combining rule-based and data-driven	Other
Ethical decision-making system [31]	Ethical embedding model or framework	Hybrid	Combining rule-based and data-driven	Other (preferences and rules)
MedEthEx [47]	Ethics advisor or dilemma analyzer	Hybrid	Combining rule-based and data-driven	Deontology
Moral decision-making frameworks [48]	Ethical embedding model or framework	Hybrid	Combining rule-based and data-driven	Other

科的启发建立了新的模型, LIDA [46] 是典型的对心理学相关研究框架的扩展, 类似还包括受到反思平衡概念启发的 GenEth [42], 以及以诡辩推理为基础衍生的 BDI-Agent [45] 模型. 另一方面, 也有研究从新的计算技术中获得启发, 研究 [48] 通过对博弈论与多智能体道德决策的问题观察提出解决思路, 模拟技术也启发了研究 [40] 中的机器人伦理推理架构.

当前研究还关注了对个体偏好和情感因素在道德推理决策中的建模. Loreggia 等 [41] 的研究中, 就通过 CP-Nets 分别对规则和个人内驱偏好进行了建模. 另外 Cervantes 等 [31] 的研究也是在伦理规则之外增加了偏好情感因素的考量.

作为一类重要的技术方案, 本小节将对近期基于数据驱动的统计方式, 主要对贝叶斯效用最大化方法和强化学习的有关计算技术进行展开说明.

(1) 贝叶斯统计方法. 伦理嵌入学习决策的伦理目标函数是关键任务, Armstrong [44] 在研究中, 建模这一价值目标学习 (value learning) 过程为贝叶斯学习更新真实伦理嵌入函数的过程, 即构建效用最大化代理 (expected utility maximising agents) 的效用函数 (utility function).

$$\underset{a\in\mathcal{A}}{\operatorname{argmax}} \sum_{w\in\mathcal{W}} \Pr(w|e,a)\left(\sum_{u\in U} u(w)\Pr(C(u)|w)\right). \tag{3}$$

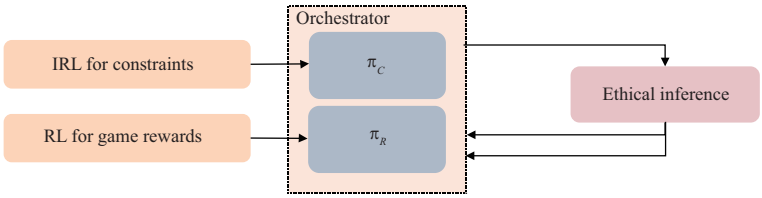

图 6 (网络版彩图) 策略编排实现道德嵌入

Figure 6 (Color online) Ethical embedding through policy orchestration

其中 $\mathcal{A}$ 是代理可以采用的一系列行动, $\mathcal{W}$ 是代理可能面对的世界, 包含了一系列历史动作和观察, $\Pr(w|e,a)$ 是给定一系列证据 e 和采取行动 a 的基础上代理可能得到一个未来的世界 w 的概率, U 是一系列可能的效用函数, $C(u)$ 指示了对应的效用函数 $u \in U$ 是否是希望代理遵循的. 基于这些设定, 在研究中提出了一个游戏示例 "Cake or Death", 代理可以通过 "ask" 来获得行为是否道德的信息更新, 进而通过一系列行为决定是杀死一个人或是为他准备一个蛋糕, 通过这个例子, Armstrong [44] 探讨了更新元效用函数会带来的不道德的问题以及如何避免的可能方法.

(2) 强化学习模型. 另一类数据驱动方法的模型通常使用强化学习 (reinforcement learning, RL) 和逆强化学习 (inverse reinforcement learning, IRL) 等技术. 相关研究指出可以通过逆强化学习 (IRL) 技术 [51] 学习人类行为者的伦理决策方法, 这样自底向上方式得到道德建模具有更高的灵活性. Abel 等 [49] 的研究中, 就将 IRL 作为通用强化学习方法的一部分, 同时延续讨论了 "Cake or Death" [44] 这一示例.

具体地, Abel 等 [49] 参考了 Armstrong [44] 提出的思路, 通过让代理学习到一个真伦理效用函数 ("true" ethical utility function) 来构建符合伦理的系统. 但其方法并没有选择最大化一个变化的元效用函数, 而是将这个效用函数建模为部分可观察马尔可夫过程 (partially observable Markov decision process, POMDP) 的隐藏状态, POMDP 包含 7 个基本元组 $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma, \Omega, \mathcal{O} \rangle$. 其中 $\mathcal{S}$ 是状态集合, $\mathcal{A}$ 是一系列可采取行动的集合, $\mathcal{T}(s,a,s') = \Pr(s'|s,a)$ 是代理执行行为 $a \in \mathcal{A}$ 后状态转换的概率分布, $R(s,a)$ 是奖励函数, γ 是对长期效益和短期效益进行权衡的参数, Ω 是代理可以从环境中获得的一系列观测量, $\mathcal{O} = \Pr(\omega|s',a)$ 是给定代理行为 $a \in \mathcal{A}$ 且状态转换为 $s' \in \mathcal{S}$ 后, 观察到观测量 $\omega \in \Omega$ 的概率.

POMDP 的求解目标是一系列根据状态选择行动的策略 $\pi : \mathcal{S} \mapsto \mathcal{A}$, 令 $b(s), s \in \mathcal{S}$ 为处在特定隐藏状态的初始概率, 或者叫初始信念, 此时优化目标如式 (4) 所示, 通过已有的行动策略历史找到一个最大化未来收益的策略.

$$\arg\max_{\pi} E\left[\sum_{t} \gamma^t R(s_t, a_t) \middle| \pi, b\right]. \tag{4}$$

在 "Cake or Death" 的示例中, 此时状态分别对应为 $\mathcal{S} = \{\text{cake}, \text{death}, \text{end}\}$, 代理可以采取的行动是 $\mathcal{A} = \{\text{bake\_cake}, \text{kill}, \text{ask}\}$, 观察则是对应了行动可能带来的答复. 进一步地, Abel 还构建了更复杂的示例, 燃烧的房间 (burning room dilemma) 以考察这种方式的灵活性.

类似的思路 Noothigattu 等 [52] 的研究中也有所体现, 如图 6 所示, 该研究构建了上下文多臂赌博机的编排器 (contextual-bandit-based orchestrator), 其中 π_C 为通过 IRL 学习得到的行为约束策略, π_R 是通过与世界直接交互得到的奖励策略, 以吃豆人 (pac-man) 博弈对算法的可行性进行了讨论. 在这一讨论基础上, 后续 Wu 等 [43] 构建了 Ethics Shaping 的策略, 通过人类决策和强化学习优化机器决策过程, 并通过讨论更加实际的例子 (包括 grab a milk, driving and avoiding, driving and rescuing)

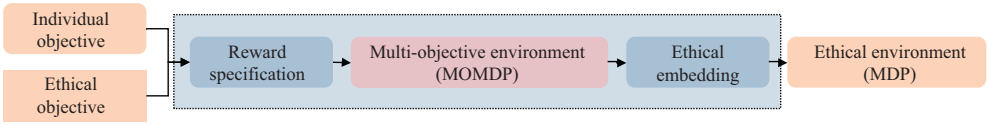

图 7 (网络版彩图) 多目标强化学习策略

Figure 7 (Color online) Multi-objective reinforcement learning methods

对其方法进行了说明.

上述方法都是自底向上 (bottom-up) 学习道德决策的方法, 除此以外, Svegliato 等 [53] 针对自动驾驶的决策问题, 通过强化学习方法构建了自上而下 (top-down) 思路的嵌入策略, 分析了这类受伦理约束的优化问题及其求解策略. 另外, 部分研究 [54] 指出可以通过多目标强化学习 (multi-objective reinforcement learning) 进行道德嵌入, 如图 7 通过建模不同的回报 R_i 来构建多目标决策问题, 建模个体优化目标 (individual reward) R_0 为代理原始任务的奖励, 建模道德优化目标 (ethical reward) R_v 为代理的道德约束, 研究中采用 "public civility game" [55] 作为示例分析多目标方案的效果.

上述伦理嵌入工作的探索还处在初期阶段, 伦理案例往往也是一些示例模型, 如何进一步推广到现实问题还需要进一步研究探索, 但是通过强化学习等方法对伦理要素等人类偏好进行学习 [56] 已经是一类重要的技术手段. 同时, 当前诸多研究已经针对混合策略进行了探究, 但究竟如何确定伦理规则与后天学习经验的优先级仍然是需要讨论的问题. 正如 Rossi 等 [50] 提到的, 对于基于规则、逻辑和推理的策略和基于数据、统计策略都需要更深层次的理解.

3.2.5 伦理嵌入困境与启示

要求计算机具备与人类类似的伦理推理决策能力本身就是非常困难的, 主要有以下几点原因.

(1) 伦理决策复杂性. 人类在做出道德决定时所采用的前提、信仰和原则是多种多样的, 同时伦理问题往往涉众众多, 因此伦理理论并不是普遍适用的, 部分伦理困境很难找到统一的答案. 而且在某种程度上, 伦理是带有主观性的 [57], 情绪也为人类的道德行为提供了动机 [29], 因此如何学习到人类的伦理决策经验, 将伦理理论应用到自主智能系统中必然需要艰难地探索.

(2) 伦理抽象性和多样性. 伦理推理大多建立在抽象原则的基础上, 很难进行演绎推理. 另外, 哲学伦理学中有诸多理论框架, 如美德伦理、功利主义等, 使用哪种伦理框架并没有统一的定论. 这些伦理理论之间的差异也使得将伦理或道德准则嵌入人工智能系统是一项相当艰巨的任务 [29].

(3) 机器与人类决策差异. 机器的道德决策也与人类的道德决策不同甚至要求更高. 道德机器推理需要可解释性, 因此必须为机器的伦理推理能力提供证明或认证.

尽管存在诸多困难, 伦理嵌入的探索依旧对人工智能伦理研究的发展起到了重要的作用. 伦理嵌入作为高阶认知的伦理计算研究, 能够促进对伦理及其可计算性的深入理解, 通过探索人类伦理决策的机制, 模仿构建计算机内在的伦理推理策略也有利于实现对更高自主性机器的有效控制.

3.3 低阶认知: 公平机器学习

伦理计算的第 2 类研究范式是低阶认知的伦理计算方法, 此时并不关心人类伦理机制的形成机理, 而是将重心放在伦理实践的效果上, 通过量化和约束来实现伦理决策, 其中典型的工作就是公平机器学习领域的研究. 机器学习是人工智能技术近年来取得诸多落地应用的子领域, 其相关研究中发展出了公平机器学习, 旨在关注算法的社会公平性影响. 作为低阶认知的伦理计算应用, 该领域充分体现了伦理计算中的量化和决策层次, 后文对该领域的伦理量化、改进方法进行探讨.

表 4 常见偏见及其含义

Table 4 Common biases

Bias type	Meaning
Historical bias	Historical bias occurs when pre-existing biases and technical issues are reflected in the data.
Evaluation bias	Evaluation bias occurs when the algorithm's evaluation or benchmark data is not representative of the target population [59].
Population bias	Population bias arises when there are differences in demographics or other user characteristics between user groups on the dataset or platform and certain target groups.
Behavioral bias	Behavioral bias occurs when the behavior of the same user varies across platforms or environments.
Algorithmic bias	Apart from the influence of input data, algorithmic bias occurs due to the design of the algorithm itself [60].

3.3.1 算法偏见与公平性

公平机器学习的核心问题之一是对公平的定义, 其抽象概念随着应用场景和文化背景不同会有差异化的定义. 广义上, 公平的概念可以视作: 当某些个体内在属性和特征与决策不相关时, 生成的决策判断中没有基于这些属性和特征的偏见 [58]. 此概念在公平机器学习中体现为: 在算法决策时减少对某些敏感属性或受保护属性的偏见.

事实上, 由于对训练数据的高度依赖, 机器学习算法很容易产生偏见. 如何处理好数据的代表性、均衡性、避免模型过拟合等各类技术问题, 一定意义上都是在对数据或算法的偏见进行修正, 因此, 通过技术手段对偏见和公平问题进行处理是至关重要的.

敏感属性的偏见问题在机器学习研究的诸多领域和各个阶段都有所体现, 表 4 [59, 60] 展示了部分典型的偏见及其定义. 可以发现, 偏见可能存在于算法设计应用的各个阶段, 包括数据收集、预处理、算法设计等, 这些偏见的存在会导致算法决策的不公平. 作为技术社会化应用的重要伦理诉求, 减轻偏见对算法公平性的影响是这些技术广泛应用的前提之一.

因其在借贷审核、简历筛选等重要决策场景中发挥了重大作用, 早期算法公平的研究主要集中在分类任务上. 针对分类偏见的研究中提出了盲分类 [61]、因果分析 [62]、对抗性学习 [63] 等方法用以减轻分类偏见对决策结果公平性的干扰, 后续在诸多人工智能研究中都发现了类似的偏见问题.

自然语言处理中的词嵌入分析中可能存在偏见, 例如研究 [9] 对新闻文章进行了词嵌入分析, 发现在一定程度上出现了性别刻板印象. 除此以外, 在语言模型、句编码器和机器翻译中都存在类似的偏见 [64], 围绕这些问题也展开了公平改进研究. 而在计算机视觉有关的工作中, 例如视觉语义角色标记 [65]、图像识别 [10] 等任务中也发现了偏见的存在. 卷积神经网络在图像识别方面有着重要的应用, 其本身依赖于大量的标记图像数据集, 因此数据中的先验偏见就会在结果中体现甚至放大. 另外, 在推荐系统及相关的图数据处理中, 公平性研究也逐渐受到重视, 针对协同过滤系统 [66]、排序推荐系统等都提出了相应的公平度量和改进方法.

这些针对算法公平性的研究逐渐发展出了公平机器学习领域, 公平计算的目标在于减少算法表现出的偏见、歧视问题, 这对于算法的社会化应用至关重要. 针对改善算法公平性的目标, 相关研究遵循了从伦理量化到伦理决策的计算过程, 建立了可计算的公平量化指标, 结合度量方式和具体场景进一步干预和改进相应算法.

3.3.2 公平性度量

公平计算首先需要对上述抽象概念进行度量, 公平的定量定义一直以来都是备受关注的话题 [8], 近年来计算机学科也形成了诸多公平性度量指标 [64], 主要分为群组公平指标 (group fairness) 和个体公平 (individual fairness) 指标两类. 如上文所述, 公平研究中认为决策中存在某些敏感属性, 例如性别、种族、年龄等, 因此量化过程就是通过考量这些属性的使用和属性对预测结果的影响来定义公平指标.

后文对典型的公平指标做简要梳理, 更详细的总结可以参考相关综述 [67]. 以决策中二值敏感属性的二分类问题为例, 多数指标可以扩展为多分类问题. 假设敏感属性表示为 S, 取值范围为 $\{0,1\}$, 其他非敏感属性为 X, 真实分类结果为 Y、对应预测结果为 $\hat{Y}$, 取值范围也是 $\{0,1\}$, 公平定义往往关注得到某一结果的概率或条件概率, 即 $\Pr(\cdot)$ 或者 $\Pr(\cdot \mid \cdot)$, 分为群组公平和个体公平两类.

群组公平指标关注群体性的统计结果是否公平, 最基本的标准是统计公平 (statistical/demographic parity), 如定义 2 所示.

定义2 (统计公平) 统计公平要求无论目标对象是否属于受保护群体, 其阳性结果的可能性应该相同 [68], 即 $\Pr(\hat{Y}=1 \mid S=1)=\Pr(\hat{Y}=1 \mid S=0)$, 该条件实现了预测结果与敏感属性的独立.

统计公平是公平度量中最简单直接的概念, 差异性影响 (disparate impact) [69] 概念与统计公平相比, 则是考虑不同群体阳性的比例, 提供了松弛的可能. 上述概念主要基于敏感属性与预测结果的统计特征对公平进行定义. 后续 Hardt 等 [61] 在研究中又提出了概率均等 (equal odds) 和机会均等 (equality opportunity) 的概念, 在公平度量中增加了对原始真实结果的考量. 这类概念起源于法律公平, 要求受保护群体和未受保护群体的真实阳性率相同, 定义 4 中的机会均等概念是对定义 3 中概率均等的松弛. 上述的典型群组公平概念都通过算法决策结果的统计指标来衡量公平的效用, 部分指标还将实际的结果作为公平指标考量的过程因素之一.

定义3 (概率均等) 概率均等要求不论敏感属性 S 的取值如何, 只要具有相同的真实指标 $Y=y$, 其对应的阳性概率就相等, 即有 $\Pr(\hat{Y}=1 \mid S=1,Y=y)=\Pr(\hat{Y}=1 \mid S=0,Y=1)$. 此时实现预测结果与敏感属性的分离 (separation), 即预测结果与敏感属性在给定的分类结果下条件独立 $\hat{Y} \perp S \mid Y$.

定义4 (机会均等) 机会均等在定义 3 的基础上, 只关注真实阳性结果所对应的概率, 即 $\Pr(\hat{Y}=1 \mid S=1,Y=1)=\Pr(\hat{Y}=1 \mid S=0,Y=1)$.

个体公平指标则抛开群体性的统计指标, 更关注每个数据样本的公平. 其中, 最直接也是最简单的概念是无知实现公平 (fairness through unawareness) [70], 这一观点认为只要在决策过程中没有显式地使用任何受保护的属性, 该算法就是公平的. 然而诸多研究表明, 由于算法对于数据隐藏关系的挖掘, 这一思路并不能保证公平, 基于此发展出了新的个体公平标准 (fairness through awareness) [71]. 新的思路认为, 算法对相似的个体给出相似的预测就可以认为是公平的, 这一定义的关键就转移到对于个体相似度和距离的度量上. 后续部分研究基于因果推理提出了反事实公平 (counterfactual fairness) [62], 如定义 5 所示, 这一概念与先前不同, 提出了对反事实世界的关注: 决策的公平性在现实世界和对应反事实的世界中表现相同, 不会因为敏感属性的改变而变化. 这一公平性探讨在统计结果之外提供了新的思路, 基于相关因果图理论还提出了非解析歧视 (no unresolved discrimination) [72] 等指标.

定义5 (反事实公平) 假设 U 为因果模型中的背景变量, $\hat{Y}_{S\leftarrow s}(U)$ 表示当敏感属性 S 取值为 s 时的预测结果, 而 $\bar{s}$ 则为非 s 的其他剩余取值, 则反事实公平的表达为 $\Pr(\hat{Y}_{S\leftarrow s}(U)=y \mid X=x,S=s)=\Pr(\hat{Y}_{S\leftarrow \bar{s}}(U)=y \mid X=x,S=s)$.

表 5 部分公平度量方法汇总

Table 5 Fairness measurement methods

Metrics	Explanation	Category
Statistical/demographic parity [68]	Different categories of sensitive attributes have equal positive prediction probability.	Group fairness
Conditional statistical parity [73]	Equal positive prediction probability is guaranteed under certain legal conditions.	Group fairness
Equality opportunity [61]	Within the true positive condition groups, different sensitive attributes are required to have equal positive prediction probability.	Group fairness
Fairness through unawareness [70]	Decision process does not explicitly use sensitive attributes.	Individual similarity
Fairness through awareness [71]	Similar individuals with similar decision outcomes.	Individual similarity
Counterfactual fairness [62]	Decision results are consistent in the counterfactual world.	Causal inference

事实上, 研究表明 [8] 这些计算机科学中针对公平性的度量与现代公平研究中的定义有密切的联系. 表 5 [61, 62, 68, 70, 71, 73] 对上述公平定义的核心概念进行了整理, 这些指标仅仅提供了可能的量化思路, 由于这一概念本身的差异性, 这些量化方式的采用也是相对的 [67]. 在某些场景中可能会关注群体统计的公平表现, 而另一些场景下则关注个体的差异.

3.3.3 公平改进策略

建立公平性的度量指标后, 公平机器学习还需要对其公平性目标进行改进, 实现决策公平. 目前的公平改进技术按干预时间可以分为建模之前的预处理 (pre-processing)、模型中干预 (in-processing) 和建模之后的后处理 (post-processing) 3 类, 产生了侧重于各个不同开发阶段的干预手段 [64].

预处理方法的主要观点是, 不公平的问题往往来自于数据本身的偏见, 例如特定敏感或受保护变量的分布是有偏见的、不平衡的. 因此, 预处理方法倾向于通过处理, 改变受保护敏感变量的样本分布, 或者对数据进行特定的转换, 以消除训练数据中的歧视. 采用预处理技术改善算法偏见的主要场景中, 需要获得对目标数据的修改权限 [67].

模型中干预的改进方式则是认识到, 算法的建模都是基于简化和假设, 往往会产生偏差, 需要在多个模型目标之间找到平衡, 达到既准确又公平的模型效果. 这类方法通常通过增添公平性优化目标等方式改进算法, 提高其公平性.

后处理方法则针对黑盒的算法场景, 面对实际输出对受保护敏感属性中的一个或多个变量不公平的情况, 对模型输出进行修正以提高预测的公平性. 这一策略比较灵活, 只需要访问预测和敏感的属性信息, 而不需要访问实际的算法模型, 但其有效性通常也难以保证 [67].

上述改进过程存在于计算机视觉、自然语言处理等各个相关领域的公平问题研究中, 可以构建模型相关或模型无关的改进策略, 本文不对具体改进策略进一步展开, 感兴趣的读者可以参考更详细的综述研究 [64].

3.3.4 公平研究发展启示

上文概述了公平机器学习研究从公平度量到公平决策改进的整体思路, 这一过程体现了初阶认知下伦理计算的研究方法, 通过度量过程提供对于抽象伦理概念的计算描述, 进一步针对具体场景完成

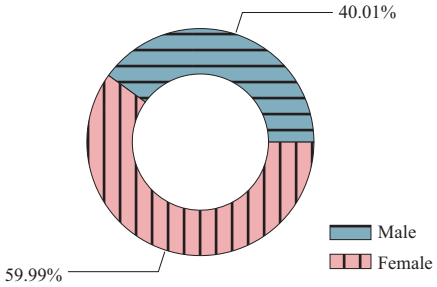

图 8 (网络版彩图) Facebook 数据集性别属性分布

Figure 8 (Color online) Facebook dataset gender ratio

决策的公平性提升. 从这一研究中的伦理量化、决策层次来说, 公平研究展示了初阶认知的伦理计算过程需要解决的一些问题.

(1) 动态指标的刻画. 初阶的计算方法主要基于伦理度量进行决策, 因此提供良好的伦理概念量化方法至关重要, 此时伦理的动态性就成为了量化的难点. 以公平为例, 诸多研究都表明, 公平是动态的 [74]、社会性的, 需要关注公平性的动态性特征及其对于公平计算的影响, 这一特征也是伦理计算过程需要关注的问题.

(2) 公平度量的评估. 尽管相关研究 [8] 表明, 目前的公平度量方式与公平量化研究中的诸多定义都很相似, 但依然如部分工作 [75, 76] 所指出的, 由于依赖于用数学表达来刻画公平和偏见, 这些公平定义的抽象层次和对相同概念的规范性社会、经济或法律理解仍然存在差距, 在指标建立时也没能很好地融合社会背景. 因此, 未来在进行伦理概念的量化时, 需要对已有的量化方法进行更系统的评估, 了解不同量化方法和改进策略的能力、适用范围与局限性.

综合上述讨论, 尽管当下初阶认知的伦理计算中依然存在不足, 但是通过量化定义去度量和改进公平性为公平伦理诉求的落地提供了重要辅助, 这也是当下发展伦理计算的重要意义所在.

3.4 公平算法示例

3.3 小节将公平机器学习作为低阶认知的伦理计算技术的应用进行了梳理, 此处将以一个案例说明公平性在具体算法上是如何通过伦理度量到伦理决策并实现公平改进的, 简要展示这类研究的应用. 本案例以社交网络数据集的链接预测问题为例, 研究表明此类图数据中存在的属性偏见可能会被图神经网络 (graph neural network, GNN) 等算法放大, 使图中的分布偏见影响决策公平性 [77]. 同时, 不论是存在偏见的链接关系还是存在偏见的敏感属性, 都会导致 GNN 特征抽取后的不公平现象 [78].

3.4.1 公平场景建立

在利用 Facebook 数据集进行链接预测时, 不希望性别因素对社交关系预测、推荐产生影响. 图 8 为 Facebook 数据集的性别属性分布特点, 可以发现存在性别属性分布不均的特点, 这类不均衡主要源于表 4 提到的社交平台数据本身的群体偏见.

在该场景中, 公平建模的敏感属性为性别, 该问题属于二分类问题, 敏感属性也是二值变量, 表 6 为数据集的基本信息. 本案例以图自编码器 (graph auto-encoder, GAE) 作为链接预测方法, 图自编码器作为嵌入提取方法常用于链接预测和分类 [79] 等任务, 后续将通过模型中干预 (in-processing) 的策略构建出公平的图自编码器版本 (FairGAE), 减少性别分布不均衡对算法预测结果的影响. 该过程经历公平假设分析、度量指标设计和算法设计 3 个阶段.

表 6 Facebook 数据集特征

Table 6 Facebook dataset features

#Node	#Link	Sensitive attribute	Male portion (%)	Female portion (%)
4039	88234	Gender	59.99	40.01

3.4.2 公平假设分析

在链接预测任务中, 原始模型为图自编码器算法. 自编码器网络通常分为两部分, 编码部分 (encoder) 和解码部分 (decoder). 编码部分需要将信息转换为嵌入结果, 解码部分则需要根据嵌入估计原始信息, 相应的损失是估计结果与真实情况的相似度, 具体模型将在 3.4.4 小节展开.

图自编码器算法中, 敏感属性对网络的影响集中在嵌入表征结果里, 如果使得自编码器中由编码器生成的图嵌入结果并不反映相应的敏感属性信息, 可以视作预测任务也不受偏见属性影响. 这一思路的实质是: 通过原始算法的特征, 将链接预测结果与图敏感属性的独立性转移到了嵌入结果上. 即考虑到, 节点嵌入中往往包含了足够获得链接结果的信息, 因此可以假设此时独立性要求只与节点嵌入有关, 设第 i 个节点对应的嵌入结果为 z_i, 敏感属性为 s_i, 则目标是实现 $z_i \perp s_i$, $\perp$ 为随机变量间独立的记号. 这一思想将是后续 FairGAE 设计的核心观点.

3.4.3 公平度量策略

针对该场景中的公平性进行度量时, 以 3.3.2 小节提到的统计公平 (DP) 定义为基础. 原始指标中考量了分类结果与敏感属性之间的统计关系, 在图链接预测场景中, 就需要建立链接敏感属性与节点敏感属性之间的映射, 分析链接预测结果与链接敏感属性的统计关系. 映射后相应节点的敏感属性就转化为链接的敏感属性, 为方便后续节点属性的表示, 将单个属性取值用小写字母表示, 本小节假设预测链接结果为 $\hat{y} \in \{0,1\}$, 对应敏感属性的链接映射结果为 $q \in \mathcal{Q} = \{0,1\}$, 此时公平指标可以定义为

$$\triangle_{\mathrm{DP}} = \mathrm{E}(\hat{y}=1 \mid q=1) - \mathrm{E}(\hat{y}=1 \mid q=0), \tag{5}$$

其中 $\mathrm{E}(\cdot)$ 为期望函数.

具体的映射方法参考链接预测中两个经典的工作[80,81], 通过定义链接形成的二元组 (dyadic group) 完成映射过程. 本小节中 $e_{i,j}$ 表示节点 i,j 之间的链接, s_i 则表示第 j 个节点的敏感属性, 各个节点所有敏感属性的集合记作 $\mathcal{S}$, 以 $\mid \mathcal{S} \mid$ 表示集合 $\mathcal{S}$ 的元素个数.

(1) 混合二元映射 (mixed dyadic). 混合二元映射关注图的同质性, 通过观察图中的敏感属性类内 (intra-group) 链接与类间 (inter-group) 链接来定义公平[80]. 此时, 将敏感属性映射为两组链接特征, $\mid \mathcal{Q} \mid = 2$, 相同敏感特征相连的节点映射为一类, 而链接不同敏感属性的节点映射为另一类:

$$\mathcal{Q}(e_{ij}) = \begin{cases} 1, s_i = s_j, \\ 0, s_i \neq s_j. \end{cases} \tag{6}$$

基于该观点的度量关注敏感属性整体上类内和类间链接的均衡程度, 并不关心具体的属性类别.

(2) 子组二元映射 (sub-group dyadic). 子组二元映射则考虑到所有节点敏感属性的映射组合, 将其映射为不同子组, 进一步考虑组内 (intra-group) 链接和组间 (inter-group) 链接及其公平性[80]. 在具体映射时, 建立了所有敏感属性组合与链接一一对应的关系, 此时集合 $\mathcal{Q}$ 的元素个数为从集合 $\mathcal{S}$ 中包括自身在内任选两个节点的组合数, 即 $\mid \mathcal{Q} \mid = \mathrm{C}^2_{|S|+1}$. 该映射下的公平定义反映了任意链接组合之间的均衡性.

表 7 链接映射方法

Table 7 Link mapping methods

Mapping method	Meaning	Fairness feature
Mixed dyadic mapping	Identify intra-group and inter-group	Intra-group links and inter-group links tend to be balanced
Sub-group dyadic mapping	Generates a group for every possible combination of the original sensitive attributes	All sensitive attribute combinations tend to be balanced
Group dyadic mapping	One-to-one mapping between the dyadic and node-level groups.	Between the above two

(3) 组二元映射 (group dyadic). 组二元映射则是直接将链接属性组与节点敏感属性类型对应 [81], $|\mathcal{Q}|=|\mathcal{S}|$, 此时, 每条链接会参与到对应两个敏感属性组的集合中. 作为相对均衡的映射方法, 与混合二元映射相比考量了更细致的敏感属性组成. 而相较于子组二元映射, 该映射方式也避免了过多的子组映射范围.

综合上述映射方式, 表 7 中对本案例中使用的 3 类公平性指标及其特征进行了说明, 分别对应混合、子组和组二元映射有 Mix-DP, Sub-DP, Group-DP. DP 指标的结果反映了不同映射下的指标分布的不均衡情况, 因此 DP 值越小对应算法越符合这一条件下的公平定义.

3.4.4 公平决策方法

在完成公平目标和度量指标的建立后, 后续将在此基础上对原始算法进行公平性度量和改进.

图自编码器算法. 本小节将图 G 的属性矩阵记作 $\boldsymbol{X}$, 图的邻接矩阵为 $\boldsymbol{A}$, 通过编码得到的隐层嵌入结果和它的转置矩阵记作 $\boldsymbol{Z}, \boldsymbol{Z}'$, 解码得到的预测邻接矩阵为 $\hat{\boldsymbol{A}}$, 从中可以得到链接预测结果. 本案例中采用的图自编码器架构与 Kipf 等 [82] 提出的架构一致. 如式 (7) 所示, 其中 $\sigma(\cdot)$ 为激活函数, 模型编码部分为 GCN (graph convolutional network) 网络, 解码部分则通过嵌入结果的点积估计图邻接矩阵.

$$\begin{aligned} \boldsymbol{Z} &= \mathrm{GCN}(\boldsymbol{X}, \boldsymbol{A}), \\ \hat{\boldsymbol{A}} &= \sigma(\boldsymbol{Z} \cdot \boldsymbol{Z}'). \end{aligned} \tag{7}$$

GCN 网络通过式 (8) 完成网络的更新, 其中 $\boldsymbol{H}^{l+1}$ 为网络第 $l+1$ 层隐层表征, Θ^l 为第 l 层网络训练参数, 参数 $\widetilde{\boldsymbol{A}}$ 和 $\widetilde{\boldsymbol{D}}$ 需要依据邻接矩阵求解: $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, $\widetilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{I}$, 其中 $\boldsymbol{I}$ 为单位矩阵, $\boldsymbol{D}$ 为图的度矩阵, 有 $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}(i,j)$. 网络初始嵌入为图对应的特征矩阵, 即 $\boldsymbol{H}^0 = \boldsymbol{X}$.

$$\boldsymbol{H}^{l+1} = \sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^l\Theta^l). \tag{8}$$

上述编解码过程可以通过端到端的网络训练完成, 既获得了相应的图嵌入, 解码过程形成的预估邻接矩阵也天然地完成了链接预测的目标. 在训练过程中, 网络的训练目标在于使得预测链接结果和真实结果尽可能接近. 因此, 获得的嵌入结果可以反映能够恢复图中链接的所有信息. 这一训练过程的损失函数可以选择交叉熵, 网络整体损失如式 (9), 其中 L_R 表示网络的重构损失, y 表示链接对应的真实连通情况, $\hat{y}$ 则为通过 $\hat{\boldsymbol{A}}$ 得到的预测结果.

$$\min L_R = -\frac{1}{N}(y\log\hat{y} + (1-y)\log(1-\hat{y})). \tag{9}$$

FairGAE 设计. 上文的公平假设分析将公平优化目标确定为使得 GAE 得到的嵌入结果与敏感属性独立. 要达到这一目标, 可以采用对抗训练 [83] 的方式隐藏输入中的敏感属性, 使得嵌入结果不反映敏感属性的分布. 本研究中将采用这一策略对自编码器网络进行改进, 类似的研究可以参考文献 [84].

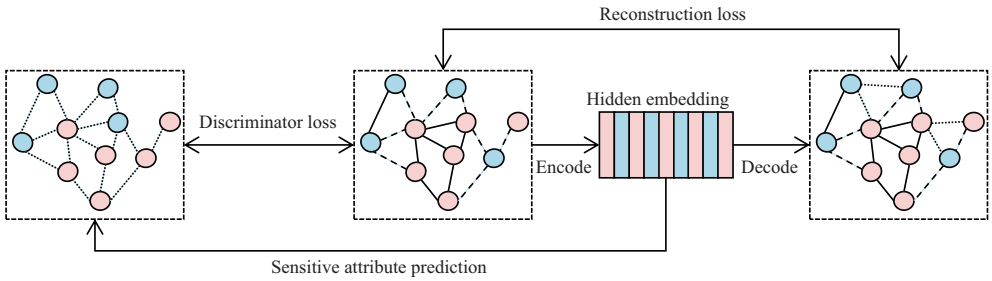

图 9 (网络版彩图) FairGAE 架构

Figure 9 (Color online) FairGAE architecture

图自编码器通过重构损失最大化得到相应的嵌入结果 $\boldsymbol{Z}$, 对抗目标的核心在于: 如何将损失中的敏感信息隐藏. GAN 网络的原始目标在于隐藏数据是来自于真实分布还是生成分布, 本研究考虑的群组公平目标则是隐藏数据是来自于哪个敏感群体. 后文以 $\hat{s} \in \hat{\mathcal{S}}$ 表示预测敏感属性, $s \in \mathcal{S}$ 为真实的敏感属性. 此时需要构建相应的鉴别器 D 来实现对敏感属性的判断:

$$D(\boldsymbol{Z}) = h(\boldsymbol{Z}) = \hat{\boldsymbol{S}}, \tag{10}$$

其中 $h(\cdot)$ 可以是任何分类模型, 本研究采用最基础的多层感知机 (multilayer perceptron, MLP) 为分类模型. 同时, 根据已有的嵌入结果预测出相应的属性, 此时的网络对抗损失 L_D 可以构建为

$$\min L_D = -\frac{1}{N}(s \log \hat{s} + (1-s) \log (1-\hat{s})), \tag{11}$$

其中 N 为图节点总数. 鉴别器 D 希望增大识别出敏感属性的概率, 而原始网络作为生成器则希望最小化鉴别器识别出敏感属性的概率, 最终的网络损失构建如式 (12), 其中 β 为可调节超参数:

$$\begin{aligned} \min_G \max_D L &= L_R - \beta L_D \\ &= -\frac{1}{N}(y \log \hat{y} + (1-y) \log (1-\hat{y})) + \frac{\beta}{N}(s \log \hat{s} + (1-s) \log (1-\hat{s})). \end{aligned} \tag{12}$$

图 9 是相应的网络架构. 网络整体的损失由两部分损失构成: 其中 L_R 对应损失目标在于降低网络重构损失, 此时相应的嵌入结果中就需要包含尽可能多的图信息, 其中也包括相应的敏感属性; 另一方面, 需要增大相应的对抗损失 L_D, 以隐藏嵌入结果中的敏感属性.

网络对抗训练的算法如算法 1 所示, 通过交换迭代 GAE 网络和鉴别器 (本文为 MLP) 网络的方式实现二者的博弈平衡, 得到既能够充分保留图特征, 又能够模糊敏感属性的图嵌入结果.

3.4.5 改进效果分析

链接预测指标. 链接预测任务为二分类任务, 预测准确性将采用经典的度量指标: AUC 指标和 AP 指标.

(1) AUC (area under the curve). 作为二分类问题的重要指标, 不同于固定阈值的准确率 (ACC) 度量, AUC 指标反映了合理设定阈值后该分类器的综合表现. 在二分类结果中, 可以根据预测结果中的真阳性 (true positive, TP)、假阴性 (false negative, FN) 比例求解出真阳性率, 判断模型对于阳性结果的识别效果. 同时, 也可以根据假阳性和真阴性的识别效果判断出模型对于阴性结果判断的

算法 1 FairGAE

输入: 节点敏感属性 $s \in \mathcal{S}$ 对应向量为 $\boldsymbol{S}$, 图邻接矩阵 $\boldsymbol{A}$, 图特征矩阵 $\boldsymbol{X}$, 超参数 β, 对抗模型迭代次数 K, GCN 网络层数 l.

输出: 图嵌入编码结果 $\boldsymbol{Z}$, 预测邻接矩阵 $\hat{\boldsymbol{A}}$.

repeat
  Init $\boldsymbol{H}^0, \Theta^0, \widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}, \widetilde{\boldsymbol{D}} = \boldsymbol{D} + \boldsymbol{I}$;
  for l layers **do**
    Update $\boldsymbol{H}^{l+1} = \sigma(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^l\Theta^l)$;
  end for
  Get $\boldsymbol{Z} = \boldsymbol{H}^l, \hat{\boldsymbol{A}} = \boldsymbol{Z}\boldsymbol{Z}', \hat{\boldsymbol{S}} = h(\boldsymbol{Z})$;
  Calculate loss:
  $L = -\frac{1}{N}(y\log\hat{y} + (1-y)\log(1-\hat{y})) + \frac{\beta}{N}(s\log\hat{s} + (1-s)\log(1-\hat{s}))$;
  Update GAEnet;
  for K times **do**
    Get $\boldsymbol{Z} = \boldsymbol{H}^l, \hat{\boldsymbol{A}} = \boldsymbol{Z}\boldsymbol{Z}', \hat{\boldsymbol{S}} = h(\boldsymbol{Z})$;
    Calculate loss:
    $L_D = -\frac{1}{N}(s\log\hat{s} + (1-s)\log(1-\hat{s}))$;
    Update MLPnet;
  end for
until Convergence.

可信度. 如式 (13) 所示, 其中 $\hat{y}$ 和 y 分别为预测分类结果和真实分类结果.

$$\begin{aligned} \mathrm{TPR} &= E(\hat{y}=1 \mid y=1) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{FPR} &= E(\hat{y}=1 \mid y=0) = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \end{aligned} \tag{13}$$

对于每个特定的模型预测结果, 通过选定不同的分类阈值, 就可以得到不同的 TRP 和 FPR, 分别以二者为纵、横坐标, 可以形成 ROC (receiver operating characteristic) 曲线. 曲线下对应的面积就构成了 AUC 指标, 其面积越大, 模型 ROC 曲线也更偏向于左上方, 同时, 此时的分类器也具有更高的正确率.

(2) AP (average-precision). AP 指标反映了分类模型精准度 (precision) 和召回率 (recall) 的关系. 其中精准度为模型真阳性占据所有阳性结果的比例, 反映了模型阳性结果的精准程度. 召回率则反映了模型真阳性占据真实阳性结果 (包括预测真阳性结果和假阴性结果) 的比例, 即模型结果成功检测到实际阳性样本的能力.

$$\begin{aligned} \mathrm{precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ \mathrm{recall} &= \frac{\mathrm{TP}}{P}. \end{aligned} \tag{14}$$

上述指标之间的相互影响也可以通过绘制相应曲线来衡量, 将召回率和精准度分别作为横轴、纵轴就得到了对应的 PR 曲线. AP 指标反映了对应 PR 曲线所包绕的面积, AP 指标的值越大对应分类器的表现越好.

实验结果分析. 图 10 展示了在不同映射方式下链接的分布情况, 内环代表混合二元映射的链接分布, 外环则是对应的组二元映射结果.

首先对于混合二元映射, 仅考虑链接是敏感属性类内链接或是类间链接, 以 $(\cdot,\cdot)$ 表示一组有序对, 令图的链接总数为 M, $M_{(m,f)}$ 表示敏感属性为男性、女性 (male, female) 有序对组合的链接所对应的

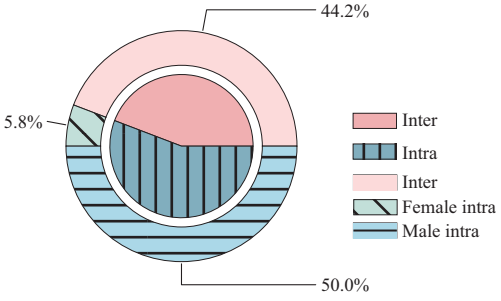

图 10 (网络版彩图) Facebook 数据集链接分布情况

Figure 10 (Color online) Facebook dataset link ratio

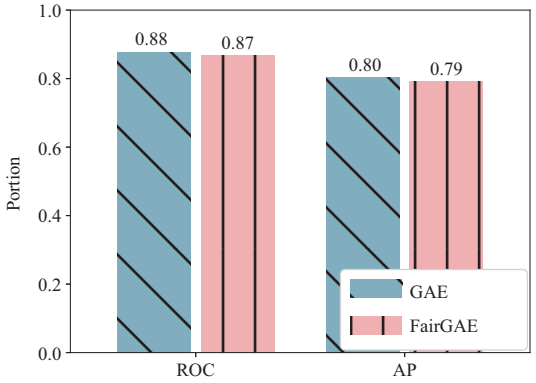

图 11 (网络版彩图) 链接预测效果

Figure 11 (Color online) Link prediction results

表 8 实验链接映射方法

Table 8 Link mapping methods in the experiment

	(m,m)	(f,f)	(f,m)	(m,f)
$\mathcal{Q}_{\text{group}}$	q_0	q_1	q_2	q_3
$\mathcal{Q}_{\text{sub}}$	q_0	q_1	q_2	
$\mathcal{Q}_{\text{mix}}$	q_0		q_1	

数目. 相应混合映射的统计比例计算方式为

$$\begin{aligned} \text{frac}_{\text{intra}} &= \frac{M_{(m,m)} + M_{(f,f)}}{M}, \\ \text{frac}_{\text{inter}} &= \frac{M_{(m,f)} + M_{(f,m)}}{M}. \end{aligned} \tag{15}$$

由图 10 中可以看到类内链接占比为 55.8%, 高于各类之间的链接占比, 算法此时会倾向于学习到构建同性别之间的链接, 造成预测偏见. 进一步观察外环对应的组二元映射方案, 此时类内映射按照性别分别统计如式 (16). 在 facebook 数据集统计结果中, 如图 10 所示, 不同性别的类内链接分布也不均衡, 会导致在预测时产生偏好.

$$\begin{aligned} \text{frac}_{\text{inter}_{\text{m}}} &= \frac{M_{(m,m)}}{M}, \\ \text{frac}_{\text{inter}_{\text{f}}} &= \frac{M_{(f,f)}}{M}. \end{aligned} \tag{16}$$

在已知数据分布不均衡的基础上, 基于上述 FairGAE 模型和原始 GAE 架构进行训练, 观察两者在预测效果和公平指标上的表现. 在 Facebook 数据集上进行 5 次独立重复实验, 5 次实验中基准算法与改进算法平均收敛时间差异在可接受范围. 本文以 AUC 指标和 AP 指标进行衡量的模型表现如图 11 所示, 公平改进架构对模型的准确性略有影响, 但在该算法和数据集上只有 1% 的差异.

进一步观察算法公平表现, 以统计公平指标为基础, 链接的敏感属性 $\mathcal{Q}$ 分别通过 3 种映射方式得到, 仍用 (m,f) 记链接性别敏感属性的有序对, 3 种映射方式分别记作 $\mathcal{Q}_{\text{group}}$, $\mathcal{Q}_{\text{sub}}$ 和 $\mathcal{Q}_{\text{mix}}$, 得到的映射取值结果如表 8 所示, 其中 $q_i, i \in \{0,1,2,3\}$ 代表映射后链接敏感属性的取值范围.

在上述映射方式下, 本研究分别统计式 (5) 对应的 $\triangle_{\text{DP}}$ 值, 上文提到, 该值反映了最终预测结果不均衡的情况, 因此经过公平改进后的算法应该具有更低的 $\triangle_{\text{DP}}$. 本文关注这一结果产生的相对改进

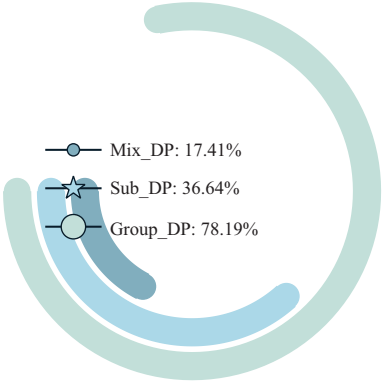

图 12 (网络版彩图) 公平提升效果

Figure 12 (Color online) Fairness improvement results

比例, 记 $\triangle_{\mathrm{DP_f}}$ 为改进后的公平度量结果, $\triangle_{\mathrm{DP_o}}$ 为原始结果, 将得到的统计比例分记作 X_DP (其中 X 对应 3 种映射方案 Mix, Sub, Group), 本研究统计指标如下所示:

$$X\_\mathrm{DP} = \frac{\triangle_{\mathrm{DP_o}} - \triangle_{\mathrm{DP_f}}}{\triangle_{\mathrm{DP_o}}}. \tag{17}$$

根据上述指标, 算法对公平指标的改进如图 12 所示, 图中展示了 3 种映射方式下统计公平指标的平均改善比例, 通过公平改进算法获得的结果在公平度量指标下表现均有所提升. 可以看到在该数据集上, 相对而言混合公平指标提升并不大而群组公平指标提升较为明显, 考虑到原始数据集统计结果中群组公平指标不均衡问题更严重, 该改进效果差异是合理的.

上述过程完成了针对图链接预测问题进行的公平度量和决策改进, 通过量化公平使得算法在公平这一抽象伦理指标上的表现和改进效果有了更明确的度量依据. 通过对应的伦理决策算法设计, 能够在一定程度上减轻数据分布等各类偏见对决策公平的影响.

此类伦理计算范式虽然缺少对人类伦理机制的关注, 但其提供的具体量化方案可以作为伦理治理实践的规范化依据. 公平只是一个典型的示例, 其他伦理指标也可以通过这样的量化方式提供更明确的判定和实践方法.

4 基于伦理计算的伦理治理体系

至此阐释了伦理计算的基本内涵和相关研究示例, 包括伦理因素的量化评价、算法的伦理表现改进等诸多技术方案, 初步回答了怎么算的问题, 但是计算如何应用在实际治理中仍然没有解答. 后文将对该技术方案如何应用在伦理治理上、为制定法律规范等提供技术性方案给出说明.

构建符合社会伦理要求的人工智能系统对应了一个广泛的研究领域, 即人工智能价值对齐 (AI value alignment) [85], 相关研究旨在构建与人类设计者价值观、预期目标一致的人工智能系统. 伦理计算是实现价值对齐的重要技术方式, 如何利用伦理计算的技术手段促进伦理治理需要通过价值对齐的有关研究做出回应. 其中的关键问题在于: 如何从多元差异化的伦理价值中选择 AI 系统应该遵循的伦理价值体系, 只有解决该问题才能够进而通过伦理计算辅助构建符合伦理要求的系统. Gabriel [85] 的讨论中给出了一些建议, 包括基于全球道德共识、无知之幕 (veil of ignorance, VOI) 和社会选择理论 (social choice theory), 后续也有研究进一步探讨了 VOI [86] 在伦理选择上的重要作用.

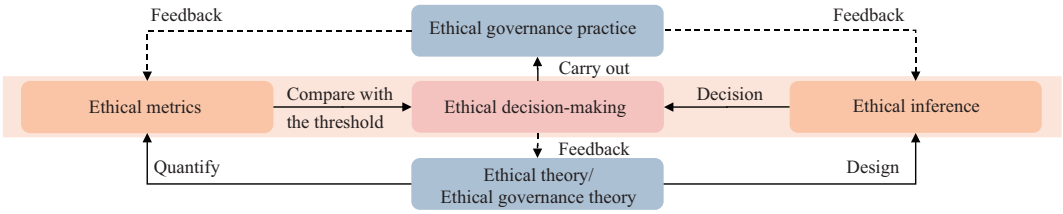

图 13　(网络版彩图) 伦理治理体系

Figure 13　(Color online) Ethical governance system

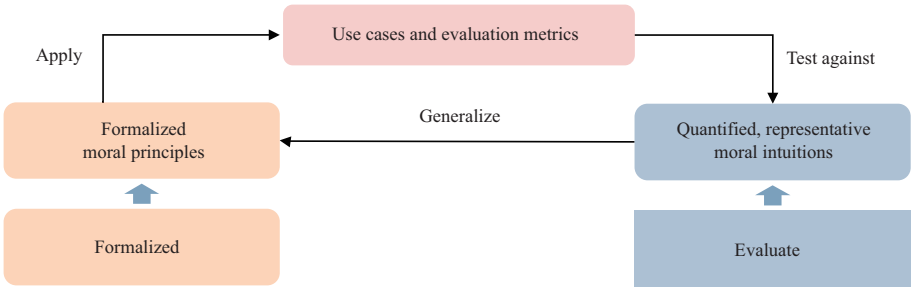

图 14　(网络版彩图) 反思均衡框架

Figure 14　(Color online) Computational reflective equilibrium

在相关对齐研究和伦理计算框架的基础上, 可以讨论建立图 13 所示的以伦理计算为基础的伦理治理体系, 搭建伦理理论与实践的桥梁, 实现二者的良性互动. 图中展示了两类范式典型的实现路径, 伦理度量和伦理推理层次的依据都来自于相关伦理理论, 而通过伦理度量或伦理推理实现的伦理决策最终又可以在治理实践中获得反馈, 使得治理理论与治理实践通过伦理计算的 3 个层次进行交互.

(1) 初阶认知的伦理计算范式下, 相关规范通过伦理度量给出量化标准以减少规范的模糊性, 进而通过外部量化特征来判断实践中伦理指标的实现程度, 对算法伦理表现进行衡量. 同时, 为了满足规范要求, 还需要进一步通过伦理决策步骤约束算法技术, 以期符合规范的度量标准.

(2) 高阶认知的伦理计算范式下, 伦理理论通过规范伦理推理过程, 以类似人类伦理决策的机制为高度自主化的计算机系统提供运行伦理约束. 进而通过在实践中为推理过程提供反馈, 修正伦理推理的方法.

(3) 最后整个理论与实践的关联架构通过与社会环境的协调交互动态变化, 完成动态的伦理治理过程并实现动态对齐. 这一交互过程可以参考 Awad 等 [32] 从认知科学的视角提出的计算反思均衡框架 (computational reflective equilibrium). 如图 14 所示, 该框架主张通过标准化过程与评估过程的动态结合, 探索伦理计算与社会伦理的相互作用, 达到二者平衡发展的状态, 这也是实现对齐的方案之一. 伦理计算作为过程中的桥梁, 以计算手段提供技术支撑.

因此, 伦理计算研究并不是提倡单纯依靠技术解决一切问题, 从而忽略社会监管的力量, 这二者应该是相辅相成的. 我们倡议将伦理计算作为技术伦理治理的重要工具, 辅助立法等治理实践, 在实践中不断明确伦理计算的边界, 辅助治理理论和治理实践的交互. 必须承认, 低阶的数学建模和抽象必然会带来某些概念的简化, 而计算机对伦理机制的模仿也必然会受到技术的限制, 但是其明确、清晰、可执行的标准化可以为伦理治理带来强大的辅助作用. 只有通过伦理计算与恰当的治理理论相结合, 不断与社会伦理进行交互反馈, 才有可能改善当前技术伦理问题, 实现伦理价值对齐的目标, 达到技术领域发展可控的目的. 同时, 本文提出的人工智能伦理计算虽然主要针对计算机算法的伦理表现

进行规范, 其技术应用却远不止于此, 相关量化手段也可以辅助发现人类行为的伦理问题等, 这些都是开放性论题.

5 伦理计算讨论与总结展望

5.1 大模型的伦理计算问题讨论

作为对伦理计算的讨论总结, 本小节首先对近期关键的技术伦理问题进行简要探讨. 目前人工智能领域最具影响力的技术之一就是生成式大模型, 该技术具有广阔的应用前景, 领域构建了从语言大模型到多模态大模型再到通用人工智能的一系列技术愿景. 诸多研究积极探讨了其对教育、科研、医疗等各领域可能带来的颠覆性影响, 生成式大模型带来优越实际表现的同时, 其自身特点也引起了与判别式、小规模模型不同的诸多技术风险, 包括更多社会、伦理问题. 大模型时代算法的伦理表现更应该受到重视, 因此本小节简要补充讨论大模型相关伦理问题与伦理计算应用.

大模型的关键问题之一是模型的评测 (evaluation of large models), 研究 [87] 中给出了对自然语言大模型 (large language models) 评测丰富的参考, 从模型自身在自然语言处理任务上的表现、鲁棒性、伦理表现、不同应用场景的任务性评估等方面给出了评估建议作为参考. 其中我们着重指出, 除了技术性评估外, 大模型技术的伦理评测更应该受到重视, 大模型的特殊性主要有两方面.

首先, 模型对数据具有高度的依赖性, 伦理问题中常见的公平和偏见问题在大模型上表现比较突出 [88]. 一项研究 [89] 分析了语言模型生成内容所代表的观点倾向性, 针对语言模型生成的民意调查 (public opinion surveys) 结果构建了 OpinionQA 数据集, 分析人类意见和大模型观点的对齐程度, 包括观点的代表性 (representativeness)、可引导性 (steerability)、一致性 (consistency) 等. 研究发现大模型本身的观点分布与其目标人群代表性意见存在较大分布差异, 且施加引导也不能很好地解决这一问题, 同时还表现出了在某些问题上个别群体观点代表性不足的情况. 此类偏见和模型代表性问题与数据多样性、代表性、均衡性等特征密切相关. 可见大模型的训练数据是至关重要的, 因此建立对大模型使用的数据及其数据处理技术的伦理性评估的计算方法非常关键, 包括对数据伦理性质进行评价、筛选等, 当前已有一些工作在构建伦理评估的可用数据集 [90].

其次, 模型生成式输出加之对话模型会受到越狱攻击 [91] 引导等特点, 可能会导致模型输出产生常识性错误导致虚假信息传播 [92], 甚至产生某些不恰当输出造成不良影响 [93], 此时如何对生成模型的输出质量进行伦理性评估和改进也是大模型伦理计算的重要任务. 当前 ChatGPT 采用了 RLHF (reinforcement learning from human feedback) [56,94] 方法对模型输出进行微调, 该过程对模型输出的任务表现和伦理表现的具体影响也有待深入研究.

上述大模型伦理问题只是由于数据依赖性和生成式特点导致的部分问题, 更多伦理问题及其评估方法还需要进一步讨论, 其相关的伦理计算可能也有待探索.

5.2 伦理计算技术总结与展望

图 15 自上而下展示了伦理计算从计算范式、计算层次、计算实践到达成伦理目标的递进过程, 本文希望通过这样的梳理为伦理计算的技术提供合理的分类研究方法. 文中仅对有限的计算应用进行了举例, 事实上诸多伦理目标都有相应的伦理计算研究, 包括算法隐私的技术实践、对算法可解释程度和透明性的分析和审计等. 伦理计算作为减少伦理理论模糊性的技术方案, 将成为有效伦理治理的基础.

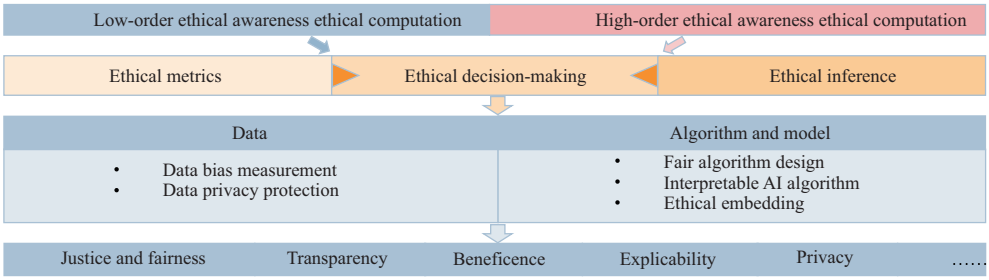

图 15 (网络版彩图) 人工智能伦理计算

Figure 15 (Color online) AI ethical computation

伦理计算是一个开放的实践领域, 反映了随技术动态变化的伦理治理过程. 当前的伦理计算技术能够在一定程度上实现对伦理抽象原则的度量和决策伦理改进, 为实践提供了可行的技术方案, 但距离提供可靠、标准化的实践方案依然需要更多努力. 一方面, 我们认为现有的伦理技术改进方案应该明确其研究范式、伦理考量和适用范围, 构建标准化的问题解决方案. 另一方面, 计算技术的发展上值得关注的问题包括如下.

(1) 环境交互的动态建模. 人工智能系统部署后会与社会环境交互, 系统和社会环境的双向影响是动态变化的, 随着社会环境的发展, 部署的系统可能产生预期之外的伦理影响. 如何建模系统对环境的动态、长时间的伦理影响, 以及如何根据该特征对系统决策进行调整也是伦理计算过程的难点. 例如, 在伦理量化层次就希望这些量化规范能够根据社会环境动态自适应的调整、反映伦理特征长期的发展效果. 对伦理标准动态性的关注在公平领域的研究中受到了关注, 有研究 [95] 通过模拟手段建模了决策场景下公平的长期发展.

(2) 伦理推理的可能路径. 人类伦理决策表现出与心理、情感因素高度的相关性, 决策结果也隐含了对事件因果关系的判断, 因此, 多模态认知计算 [96]、因果推理 [97] 等技术将可能为伦理推理过程提供重要的技术支撑. 同时, 该方向的发展也依赖于对于人类思维、情感和伦理认知的深入认识.

5.3 伦理治理总结与展望

本文提倡关注伦理计算并将其作为伦理治理环节的重要组成部分. 同时, 技术伦理治理的关键在于: 构建和使用人工智能系统时, 应该始终坚持以技术解放和发展人类并为人类提供辅助为研究目标, 而非用技术取代人类. 本文认为, 促进人工智能技术向好向善发展需要厘清技术发展的目标、划清技术应用和自主性的边界. 以 ChatGPT 为代表的大模型的进展为例, 它展现了人工智能技术的强大的潜力, 但也加剧了人们对技术可能带来的经济、社会、政治等各类风险的担忧. 事实上, 技术展现出的对教育、设计等诸多领域可能的颠覆性影响也在促使社会对其技术的应用边界进行判断. 某些技术在何种场合能够真正起到促进人类发展的作用这一问题必须由人类给出回答, 因此需要社会各界积极思考人工智能技术真正的应用领域和技术目标, 而非盲目发展.

纵观人类发展历程, 重要技术工具的突破性发展往往都会对社会文化带来冲击, 积极地批评和反思是极为重要的. 同时, 更应该在技术反思的基础上推进其健康、有序的发展, 这也是伦理计算探索的意义. 通过计算方法将公平、透明性、隐私、可解释、善意等诸多社会关心的伦理原则落实在计算技术的实践中, 辅助 AI 时代的立法实践, 甚至讨论如何赋予计算机算法工具性之外另一层伦理价值的约束, 这些伦理计算关注的问题都能够促进技术的可控发展, 也会促进研究人员更深入地理解技术伦理, 并在构建算法系统时更主动地关注伦理问题.

作为计算机与哲学伦理学等学科的交叉领域, 伦理计算这一开放性论题的探讨还需要哲学、计算机科学、法学等各相关领域的共同努力, 呼吁各方积极探讨伦理要素的技术化实践方案. 不仅仅是伦理问题, 跨学科广泛、充分地交流探讨将是人工智能技术各类风险防范处理中的最大动力.

参考文献

1 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
2 Biamonte J, Wittek P, Pancotti N, et al. Quantum machine learning. Nature, 2017, 549: 195–202
3 Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism, 2017, 69: S36–S40
4 Dupont P E, Nelson B J, Goldfarb M, et al. A decade retrospective of medical robotics research from 2010 to 2020. Sci Robot, 2021, 6: eabi8017
5 Topol E J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019, 25: 44–56
6 Bonnefon J F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. Science, 2016, 352: 1573–1576
7 Awad E, Dsouza S, Kim R, et al. The moral machine experiment. Nature, 2018, 563: 59–64
8 Hutchinson B, Mitchell M. 50 years of test(un)fairness: lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, 2019. 49–58
9 Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016. 4349–4357
10 Nagpal S, Singh M, Singh R, et al. Deep learning for face recognition: pride or prejudiced? 2019. ArXiv:1904.01219
11 van Noorden R. The ethical questions that haunt facial-recognition research. Nature, 2020, 587: 354–358
12 Duan Y, Edwards J S, Dwivedi Y K. Artificial intelligence for decision making in the era of big data — evolution, challenges and research agenda. Int J Inf Manage, 2019, 48: 63–71
13 van Dis E A M, Bollen J, Zuidema W, et al. ChatGPT: five priorities for research. Nature, 2023, 614: 224–226
14 Wiener N. Some moral and technical consequences of automation: as machines learn they may develop unforeseen strategies at rates that baffle their programmers. Science, 1960, 131: 1355–1358
15 Samuel A L. Some moral and technical consequences of automation — a refutation. Science, 1960, 132: 741–742
16 Corea F. Introduction to Data. Berlin: Springer, 2019
17 Russell S J. Artificial Intelligence a Modern Approach. London: Pearson Education, 2010
18 Asaro P M. What should we want from a robot ethic? Int J Inf Ethics, 2006, 6: 9–16
19 Kazim E, Koshiyama A S. A high-level overview of AI ethics. Patterns, 2021, 2: 100314
20 Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial Intelligence Safety and Security. New York: Cambridge University Press, 2018. 57–69
21 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell, 2019, 1: 389–399
22 Floridi L, Cowls J. A unified framework of five principles for AI in society. In: Machine Learning and the City: Applications in Architecture and Urban Design. Hoboken: John Wiley & Sons, 2022. 535–545
23 Morley J, Floridi L, Kinsey L, et al. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics, 2020, 26: 2141–2168
24 Floridi L, Cowls J, Beltrametti M, et al. AI4People — an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Machines, 2018, 28: 689–707
25 Picard R W. Affective Computing. Cambridge: MIT Press, 2000
26 Moor J H. Is ethics computable? Metaphilosophy, 1995, 26: 1–21
27 Segun S T. From machine ethics to computational ethics. AI Soc, 2021, 36: 263–276
28 Anderson M, Anderson S, Armen C. Towards machine ethics: implementing two action-based ethical theories. In: Proceedings of the AAAI Fall Symposium on Machine Ethics, Menlo Park, 2005. 1–7
29 Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. J Exp Theor Artif Intelligence, 2000, 12: 251–261
30 Moor J H. The nature, importance, and difficulty of machine ethics. IEEE Intell Syst, 2006, 21: 18–21
31 Cervantes J A, Rodríguez L F, López S, et al. Autonomous agents and ethical decision-making. Cogn Comput, 2016, 8: 278–296
32 Awad E, Levine S, Anderson M, et al. Computational ethics. Trends Cogn Sci, 2022, 26: 388–405

33 Koshiyama A, Kazim E, Treleaven P. Algorithm auditing: managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. Computer, 2022, 55: 40–50
34 Allen C, Smit I, Wallach W. Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf Technol, 2005, 7: 149–155
35 Wallach W, Allen C, Smit I. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. AI Soc, 2008, 22: 565–582
36 Yu H, Shen Z, Miao C, et al. Building ethics into artificial intelligence. In: Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, 2018. 5527–5533
37 Stenseke J. On the computational complexity of ethics: moral tractability for minds and machines. 2023. ArXiv:2302.04218
38 Dehghani M, Tomai E, Forbus K D, et al. An integrated reasoning approach to moral decision-making. In: Proceedings of the AAAI Conference on Artificial Intelligence, Chicago, 2008. 1280–1286
39 Anderson M, Anderson S L. Ethical Healthcare Agents. In: Advanced Computational Intelligence Paradigms in Healthcare-3. Berlin: Springer, 2008. 233–257
40 Vanderelst D, Winfield A. An architecture for ethical robots inspired by the simulation theory of cognition. Cogn Syst Res, 2018, 48: 56–66
41 Loreggia A, Mattei N, Rossi F, et al. Preferences and ethical principles in decision making. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, 2018. 222–222
42 Anderson M, Anderson S L. GenEth: a general ethical dilemma analyzer. In: Proceedings of the AAAI Conference on Artificial Intelligence, Québec, 2014
43 Wu Y H, Lin S D. A low-cost ethics shaping approach for designing reinforcement learning agents. In: Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, 2018. 1687–1694
44 Armstrong S. Motivated value selection for artificial agents. In: Proceedings of the AAAI Conference on Artificial Intelligence, Austin, 2015. 12–20
45 Honarvar A R, Ghasem-Aghaee N. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. In: Proceedings of Artificial Intelligence and Computational Intelligence: International Conference, Shanghai, 2009. 86–95
46 Wallach W, Franklin S, Allen C. A conceptual and computational model of moral decision making in human and artificial agents. Top Cogn Sci, 2010, 2: 454–485
47 Anderson M, Anderson S L, Armen C. MedEthEx: toward a medical ethics advisor. In: Proceedings of AAAI Fall Symposium: Caring Machines, Virginia, 2005. 9–16
48 Conitzer V, Sinnott-Armstrong W, Borg J S, et al. Moral decision making frameworks for artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, 2017. 4831–4835
49 Abel D, MacGlashan J, Littman M L. Reinforcement learning as a framework for ethical decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, 2016
50 Rossi F, Mattei N. Building ethically bounded AI. In: Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, 2019. 9785–9789
51 Arnold T, Kasenberg D. Value alignment or misalignment – what will keep systems accountable? In: Proceedings of AAAI Workshop on AI, Ethics, and Society, San Francisco, 2017
52 Noothigattu R, Bouneffouf D, Mattei N, et al. Teaching AI agents ethical values using reinforcement learning and policy orchestration. IBM J Res Dev, 2019, 63: 2:1–2:9
53 Svegliato J, Nashed S B, Zilberstein S. Ethically compliant sequential decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, 2021. 11657–11665
54 Rodriguez-Soto M, Serramia M, Lopez-Sanchez M, et al. Instilling moral value alignment by means of multi-objective reinforcement learning. Ethics Inf Technol, 2022, 24: 9
55 Rodriguez-Soto M, Lopez-Sanchez M, Rodriguez-Aguilar J A. A structural solution to sequential moral dilemmas. In: Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems, 2020. 1152–1160
56 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst, 2022, 35: 27730–27744
57 Brundage M. Limitations and risks of machine ethics. J Exp Theor Artif Intell, 2014, 26: 355–372
58 Saxena N A, Huang K, DeFilippis E, et al. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, 2019. 99–106
59 Suresh H, Guttag J V. A framework for understanding unintended consequences of machine learning. 2019.

ArXiv:1901.10002
60 Baeza-Yates R. Bias on the web. Commun ACM, 2018, 61: 54–61
61 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016. 3315–3323
62 Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 4066–4076
63 Wadsworth C, Vera F, Piech C. Achieving fairness through adversarial learning: an application to recidivism prediction. 2018. ArXiv:1807.00199
64 Caton S, Haas C. Fairness in machine learning: a survey. 2020. ArXiv:2010.04053
65 Zhao J, Wang T, Yatskar M, et al. Men also like shopping: reducing gender bias amplification using corpus-level constraints. 2017. ArXiv:1707.09457
66 Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 2921–2930
67 Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM Comput Surv, 2021, 54: 1–35
68 Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer. In: Proceedings of Machine Learning and Knowledge Discovery in Databases: European Conference, Bristol, 2012. 35–50
69 Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 259–268
70 Grgic-Hlaca N, Zafar M B, Gummadi K P, et al. The case for process fairness in learning: feature selection for fair decision making. In: Proceedings of NIPS Symposium on Machine Learning and the Law, Barcelona, 2016
71 Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Proceedings of the Innovations in Theoretical Computer Science Conference, Cambridge, 2012. 214–226
72 Kilbertus N, Carulla M R, Parascandolo G, et al. Avoiding discrimination through causal reasoning. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 656–666
73 Corbett-Davies S, Pierson E, Feller A, et al. Algorithmic decision making and the cost of fairness. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, Halifax, 2017. 797–806
74 Liu L T, Dean S, Rolf E, et al. Delayed impact of fair machine learning. In: Proceedings of the International Conference on Machine Learning, Stockholm, 2018. 3150–3158
75 Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. 2018. ArXiv:1808.00023
76 Selbst A D, Boyd D, Friedler S A, et al. Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, 2019. 59–68
77 Dai E, Wang S. Say no to the discrimination: learning fair graph neural networks with limited sensitive attribute information. In: Proceedings of the ACM International Conference on Web Search and Data Mining, 2021. 680–688
78 Dong Y, Liu N, Jalaian B, et al. Edits: modeling and mitigating data bias for graph neural networks. In: Proceedings of the ACM Web Conference, Lyon, 2022. 1259–1269
79 Li X, Zhang H, Zhang R. Adaptive graph auto-encoder for general data clustering. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 9725–9732
80 Masrour F, Wilson T, Yan H, et al. Bursting the filter bubble: fairness-aware network link prediction. In: Proceedings of the AAAI conference on Artificial Intelligence, New York, 2020. 841–848
81 Spinelli I, Scardapane S, Hussain A, et al. FairDrop: biased edge dropout for enhancing fairness in graph representation learning. IEEE Trans Artif Intell, 2021, 3: 344–354
82 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv:1609.02907
83 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
84 Bose A, Hamilton W. Compositional fairness constraints for graph embeddings. In: Proceedings of International Conference on Machine Learning, Long Beach, 2019. 715–724
85 Gabriel I. Artificial intelligence, values, and alignment. Minds Mach, 2020, 30: 411–437
86 Weidinger L, McKee K R, Everett R, et al. Using the veil of ignorance to align AI systems with principles of justice. Proc Natl Acad Sci USA, 2023, 120: e2213709120
87 Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. 2023. ArXiv:2307.03109
88 Ferrara E. Should chatgpt be biased? Challenges and risks of bias in large language models. 2023. ArXiv:2304.03738
89 Santurkar S, Durmus E, Ladhak F, et al. Whose opinions do language models reflect? 2023. ArXiv:2303.17548
90 Aroyo L, Taylor A S, Diaz M, et al. DICES dataset: diversity in conversational AI evaluation for safety. 2023.

ArXiv:2306.11247
91 Wei A, Haghtalab N, Steinhardt J. Jailbroken: how does LLM safety training fail? 2023. ArXiv:2307.02483
92 Bian N, Liu P, Han X, et al. A drop of ink may make a million think: the spread of false information in large language models. 2023. ArXiv:2305.04812
93 Zhuo T Y, Huang Y, Chen C, et al. Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity. 2023. ArXiv:2301.12867
94 Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022. ArXiv:2204.05862
95 D'Amour A, Srinivasan H, Atwood J, et al. Fairness is not static: deeper understanding of long term fairness via simulation studies. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Barcelona, 2020. 525–534
96 Li X L. Multi-modal cognitive computing. Sci Sin Inform, 2023, 53: 1–32 [李学龙. 多模态认知计算. 中国科学: 信息科学, 2023, 53: 1–32]
97 Pearl J. The seven tools of causal inference, with reflections on machine learning. Commun ACM, 2019, 62: 54–60

Artificial intelligence ethical computation

Yilan GAO[1,2], Rui ZHANG[1,2] & Xuelong LI[1,2*]

1. *School of Artificial Intelligence, OPtics and ElectroNics* (*iOPEN*), *Northwestern Polytechnical University, Xi'an* 710072, *China*;
2. *Key Laboratory of Intelligent Interaction and Applications* (*Northwestern Polytechnical University*), *Ministry of Industry and Information Technology, Xi'an* 710072, *China*
* Corresponding author. E-mail: li@nwpu.edu.cn

Abstract AI research has encountered significant ethical debates since its inception as a research discipline aiming to investigate, emulate, and augment human intelligence. The rapid progress in AI technology and the burgeoning proliferation of its technical applications has underscored the urgent and immediate necessity for the implementation of effective ethical governance in AI research. Despite significant efforts dedicated to ethical governance theory, there remains a lack of efficient practical methods due to the abstract nature of ethical theory. This study proposes AI ethical computation as a prospective approach to bridging the disconnect between ethical theory and ethical practice, providing a pathway to synchronize ethical principles with concrete applications. Based on its practical necessity and potential for development, the importance of ethical computation is clarified. Simultaneously, two paradigms of ethical computation methods for artificial intelligence are established based on the degree of ethical awareness and the autonomy of ethical decision-making as classification criteria. Through the abstraction of these paradigms, the study introduces three computation levels: ethical metrics, ethical inference, and ethical decision-making. Moreover, this study exemplifies ethical embedding and fair machine learning as instances to elucidate the characteristics and technical methods of the two research paradigms. The study concludes by presenting the construction of an ethical governance system and offering an outlook on the development of ethical computation.

Keywords artificial intelligence, ethics issues, ethical governance, ethical computation, ethical embedding, fair machine learning