# 人工智能伦理问题与治理对策综述研究*

孙羽佳　郭雪俐　苏　淞　唐红红

（北京师范大学经济与工商管理学院，北京　100875）

**内容提要：**人工智能（AI）作为培育和发展新质生产力的重要引擎，其快速发展也引发了诸多伦理问题。本文对这些伦理问题进行了系统分析，并提出了相应的治理对策。首先，本文系统梳理了现有文献，明确了AI伦理的内涵，揭示了其研究演进路径与趋势。其次，依据不同AI功能与人机关系，深入剖析了数据获取与利用、算法分类与预测、生产代理与学习顾问、社交与互动四个领域的伦理问题，阐明了其具体表现形式及其后果的产生机制。最后，构建了一个涵盖人机非交互与交互过程，整合用户、企业与社会多元主体的AI伦理问题理论框架，并结合中国文化背景，提出了契合本土实际的治理框架。本文的主要创新点包括：第一，系统揭示了AI伦理问题及其后果；第二，针对中国语境，提出了具有可操作性的本土治理路径；第三，回应了“十五五”期间国家在社会治理创新与可持续发展中的理论需求。通过这些创新，本文不仅为AI发展过程中的伦理挑战提供了新的分析视角，也为我国科技创新与社会治理提供了理论支持与实践指导。

**关键词：**人工智能　伦理问题　道德　治理对策

**中图分类号：**F270-05　**文献标志码：**A　**文章编号：**1002—5766（2025）05—0025—19

## 一、引　言

人工智能（artificial intelligence，AI）广义上可被定义为任何能够表现出类人智能行为的计算系统（Martınez-Miranda和Aldea，2005）[1]。AI的进展主要体现在两方面：一方面，AI正逐步从单纯工具转变为具备自主决策能力的代理以及社会互动团队的重要成员（Kim和Mcgill，2024）[2]；另一方面，企业的数字化转型战略持续推动各行业对AI的应用与投资。据Forbes（2022）[3]预测，2020—2028年，全球AI投资增长将超过1250%，累计投资额预计达到6410亿美元。以上两大进展表明，随着AI角色的转变及其应用规模的迅速扩张，AI正更深层次、更广范围地融入人类生活的各个领域。

AI在不同应用情境下——无论是与人类的非交互过程还是交互过程——均可能引发伦理问题。发展AI的目标在于构建能够模拟人类执行任务的算法或机器（Huang和Rust，2021）[4]。发展至今，AI的主要功能可归纳为四类：数据获取与利用、算法分类与预测、生产代理与学习顾问、社交与互动（Puntoni等，2021）[5]。这四种功能反映了AI在非交互与交互过程中的不同作用，并涉及多元主体的参与。非交互与交互过程的区分依据在于AI是否与人类直接产生互动与反馈（Ebrahimi

**收稿日期：**2024-08-13

* **基金项目：**国家自然科学基金面上项目“基于认知神经机制的消费者财务决策规避行为研究”（71872016）；中央高校基本科研业务费专项资金项目（1233200014）。

**作者简介：**孙羽佳，女，博士研究生；研究领域是数字营销和消费者行为，电子邮箱：202231030021@mail.bnu.edu.cn；郭雪俐，女，博士研究生；研究领域是数字营销和消费者行为，电子邮箱：202131030024@mail.bnu.edu.cn；苏淞，男，教授，博士生导师，管理学博士，研究领域是决策心理和数字伦理，电子邮箱：sus@bnu.edu.cn；唐红红，女，副教授，研究生导师，认知神经学博士，研究领域是管理和营销中的人机交互决策过程以及消费者心理与行为，电子邮箱：tang.h@bnu.edu.cn。通讯作者：唐红红。

等，2022）[6]。非交互过程主要涉及AI的内部运行与数据处理。科技企业作为AI的开发者与应用推动者，主要支持实现前两类功能。然而，该过程可能引发数据隐私泄露等伦理问题。交互过程则主要涉及用户的实际参与和行为反馈，对应于AI的后两类功能。该过程可能引发责权分布冲突等伦理问题，并可能进一步扩展至社会整体层面的规范与治理挑战。

总体而言，上述四种功能的实现过程凸显了AI在非交互与交互过程中引发的伦理问题，以及由此带来的对社会治理的迫切需求。基于此，本文以“AI功能—非交互与交互过程的伦理问题—后果产生机制—社会治理”为研究脉络，聚焦以下核心问题：AI在实现功能的过程中带来了哪些伦理问题？潜在的后果如何体现？社会各方应如何有效应对与治理？

实践背景方面，在AI既具备生产力赋能作用又带来治理挑战的背景下，明确其伦理问题至关重要。首先，AI具备典型通用技术特征，是培育和发展新质生产力的重要引擎。其次，根据中国日报网对“十五五”（2026—2030年）规划的分析，创新驱动和绿色发展将成为推动中国式现代化的重要支撑①，这一战略导向对AI发展提出了更高的伦理规范要求。在此实践背景下，AI伦理规范的意义体现在以下四个方面：首先，数据隐私泄露和算法偏见等伦理风险愈发凸显，构建透明、公正的伦理框架成为实现高效治理的基础。其次，推动经济高质量发展要求加强AI应用的监管。然后，公众参与是实现“以人为本”发展理念的重要支撑，深入探究AI对用户的影响及其作用机制，有助于从根源上保护人类利益。最后，深化对AI伦理问题的认识，有助于重构公众对技术的信任，从而促进社会与技术的良性互动。综上所述，明确AI伦理问题并健全相关治理对策，不仅是应对技术发展挑战的必要措施，更是确保“十五五”目标顺利实现的重要保障。

鉴于AI伦理规范在实践层面的重要性，亟需研究从理论层面系统梳理AI伦理问题，并探讨其潜在机制，从根源上厘清问题的本质。然而，本文通过梳理现有文献发现，与AI技术的快速发展相比，AI伦理问题及其潜在后果的研究仍处于初步阶段，缺乏系统性和深度。具体而言，现有研究主要存在以下四点局限：第一，早期由于AI技术尚未广泛应用，研究主要聚焦于自动化机器的自主性（Wiener，1960）[7]，较少探讨伦理问题和治理对策。第二，随着AI技术的发展和广泛应用，研究开始关注特定场景的伦理问题，如自动驾驶危机（Gill，2020）[8]、算法歧视（Akter等，2021）[9]、威胁就业（李舒沁等，2021）[10]等。然而，这些研究主题较为分散，且较少探讨AI对社会整体层面的影响。第三，现有综述型研究多集中于静态伦理考量，即AI在后台运行而不直接与用户互动的情境（赵志耘等，2021）[11]，人机动态交互过程中涉及的伦理问题尚未得到充分总结和机制分析（例如，与AI交互后人们的认知和行为会发生怎样的变化）。第四，AI伦理问题的基础原则和整合性治理对策缺乏讨论。鉴于上述研究局限，本文旨在系统解析AI引发的伦理挑战，并基于中国社会背景与全球经验，构建系统性的伦理框架与治理对策。

在上述实践背景和研究基础上，首先，本文将梳理AI伦理的内涵和研究现状，进而明确本文的理论价值。其次，以AI的不同功能和人机关系为导向，沿着问题的产生原因、作用机制和后果的逻辑顺序，对人机非交互和交互过程中涉及的伦理问题进行系统分析。此外，归纳总结AI在非交互和交互过程中伦理问题的共性与差异，进而构建AI伦理问题的理论框架。最后，结合“十五五”时期我国社会治理创新与可持续发展目标，从战略角度提出治理对策与研究议题，为我国科技创新与高质量发展提供理论支持和实践指导。

本文的研究意义体现在理论和实践两个方面。理论上，第一，通过构建AI伦理问题的理论框架，为不同应用场景下AI伦理问题的产生与解释机制提供了理论依据，丰富了AI伦理的理论体

① 中国日报网.“十五五”规划的核心是发展新质生产力、全面推进中国式现代化[EB/OL].https://bj.chinadaily.com.cn/a/202409/24/WS66f250bfa310b59111d9ac68.html，2024-09-24。

系。第二,从社会各方(用户、企业、政府)的角度揭示了AI伦理问题的后果,全面分析了其对社会的深远影响。第三,结合研究现状与实践需求,提出了一系列具备学术前沿性和指导意义的未来研究议题。实践上,本文结合宏观与微观治理对策,针对AI伦理问题提出具体建议,为AI的健康发展和社会治理创新提供了参考。同时,本文回应了"十五五"期间我国在推进社会治理创新与实现可持续发展目标方面对理论研究的现实需求。总体而言,本文通过对AI伦理问题的全面分析及治理对策的构建,旨在为应对AI技术发展中的挑战提供理论支持,并为实现社会和谐与高质量发展贡献智慧。

## 二、AI伦理研究现状分析

### 1.AI伦理的内涵界定

AI伦理是伴随AI技术发展而兴起的研究领域,既体现了国家在AI驱动生产力发展过程中的引导,也反映了公众对AI影响力的广泛关注。最初,学者们主要探讨了AI的道德地位,即是否应将AI视为具有道德判断的独立主体(Thompson,1999)[12]。随着AI在情感理解等能力上的进步,人们可以与AI建立实时交互关系和情感纽带,从而引发了更加复杂的伦理问题。然而,现有研究仅指出AI在某些细分领域所引发的不良后果,如自动驾驶导致交通事故(Gill,2020)[8]等,缺乏对AI伦理的内涵界定。因此,在全面分析AI伦理问题之前,有必要廓清AI伦理在数智化时代的内涵。

AI伦理的原则可以借鉴并延续人类伦理的基础,从而以人类伦理为框架来解释AI伦理的内涵。随着技术的发展,AI不仅是完成任务的工具,还逐渐成为了社会成员。根据计算机为社会行动者(the computers are social actors paradigm,CASA)理论,人们在与计算机进行互动时倾向于遵循类同于人际交往的社会规范(Nass和Moon,2000)[13]。随着AI在外形设计和交互方式上更加拟人化,这一倾向更加显著(Song和Kim,2022)[14]。因此,AI在能力、外观以及交互方式上的演变,使其与人类之间的界限更加模糊(Kim和Mcgill,2024)[2]。鉴于AI作为社会行动者的特殊性与其拟人化特征,AI伦理的内涵界定可借鉴人类伦理的基本原则。

人类伦理是指在人类社会中指导个体和群体行为的道德准则和价值观体系,通常涉及对正确与错误、美德与恶行的判断。其理论基础可能源自神圣命令论(divine command)、道德现实主义(moral realism)、功利主义(utilitarianism)、实证主义(positivism)和自由主义(libertarianism)等不同的伦理学思想体系(Ayala,2010)[15]。Jones(1991)[16]将不道德的决策定义为在法律上不被允许,或在道德上不被大多数人接受的决策。在学术语境中,道德(moral)和伦理(ethics)通常互换使用(Ayala,2010)[15]。因此,本文以"AI伦理"作为所有与AI相关伦理与道德议题的统称。

AI伦理是一个高度跨学科的研究领域,涵盖计算机科学、哲学、法律、社会学、心理学等多学科内容。为了界定AI伦理,Kazim和Koshiyama(2021)[17]从影响端的角度,将AI伦理定义为"AI对心理、社会和政治的影响"。更为全面的是,赵志耘等(2021)[11]从伦理主体关系、伦理存在形式和伦理价值判断标准三个维度,将AI伦理定义为"为实现人与机器、人与人以及人与自然的和谐发展而形成的有关AI活动的一系列价值观、原则或规范"。该定义强调了AI伦理规范的目标是实现人、机器和自然之间的和谐发展。在此基础上,本文囊括了个体微观层面的伦理基础以及AI应用层面的技术伦理,同时考虑到社会层面伦理问题的后果,进一步细化了AI伦理的内涵。具体而言,本文将AI伦理定义为"AI与人类进行非交互与交互过程中所涉及的道德准则与价值观",将AI伦理问题定义为"AI在应用过程中因违反道德准则与价值观而引发的风险和道德困境",将AI伦理问题的后果定义为"AI伦理问题对人类(个体心理与行为)、企业以及政府(政策)的潜在影响"。

### 2.国内外AI伦理问题研究关注热点

国内外对AI伦理问题进行了广泛的讨论。本文主要遵循以下步骤对中英文期刊文献进行了

收集与筛选。首先，在Web of Science核心数据库（涵盖中国科学引文数据库）中，以artificial intelligence、AI、robot、human-robot interaction、human-agent interaction、autonomous agency和ethics、moral、morality为关键词进行检索。检索时间截至2024年1月10日，文献语言选择英文和中文，文献主题选择Computer Science、Philosophy、Psychology和Behavioral Sciences。其次，剔除社论、新闻、会议论文、工作论文、评论等类型文献后，初步获得相关文献。然后，阅读上一个步骤中获得的文献摘要后，剔除与本文研究主题不符的文献，最终筛选出1502篇符合本文研究主题的文献。

获得有效文献后，本文对文献关键词进行了统计分析。表1展示了AI伦理相关研究的主要关键词占比。首先，负责任的AI（responsible AI）作为研究的核心议题，涉及如何确保AI系统的公平性、透明性和问责制（Alkire等，2024[18]；肖红军，2022[19]）。其次，结果显示，人机交互成为核心关注点，研究内容包括社交机器人（Broekens等，2009[20]；邓士昌等，2023[21]）、人机协作（Luo等，2021[22]；Jia等，2024[23]）及聊天助手（Zhou等，2022）[24]等。此外，AI伦理问题的后果也受到了广泛关注，重点议题包括老年人护理、教育、儿童保护、医疗应用等特定领域的潜在后果（邓士昌等，2023[21]；Karunanithi，2007[25]；Cadario等，2021[26]）。最后，在治理对策的研究中，学者们主要关注伦理准则（赵志耘等，2021[11]；Haidt，2007[27]）、法规制定（Hagendorff，2020）[28]及监管机构的设立（Floridi，2018）[29]等。

表1　　AI伦理问题研究主要关键词占比

| 类别 | 关键词 | 频次（提及次数） | 关注比率（%） |
|---|---|---|---|
| AI应用 | 人工智能 | 709 | 28.77 |
| | 负责任人工智能 | 242 | 9.82 |
| | 人机交互 | 122 | 4.95 |
| | 社交机器人 | 67 | 2.72 |
| | 系统 | 65 | 2.64 |
| | 机器人 | 55 | 2.23 |
| | 自动化 | 32 | 1.30 |
| | 机器学习 | 32 | 1.30 |
| | 数字化转型 | 21 | 0.85 |
| 社会影响 | 老年人护理 | 123 | 4.99 |
| | 政策/法律监管 | 104 | 4.22 |
| | 信任 | 66 | 2.68 |
| | 行为 | 34 | 1.38 |
| | 决策 | 26 | 1.06 |
| | 教育 | 24 | 0.97 |
| | 儿童 | 20 | 0.81 |
| | 医疗 | 14 | 0.57 |
| 伦理道德 | 道德 | 581 | 23.58 |
| | 隐私 | 53 | 2.15 |
| | 公平 | 29 | 1.18 |
| | 道德代理 | 20 | 0.81 |
| | 道德责任 | 12 | 0.49 |
| | 商业道德 | 10 | 0.41 |
| | 道德归因 | 3 | 0.12 |

**3.AI伦理演进路径和趋势**

为了明晰AI伦理在时间维度上的演进路径，本文采用CiteSpace软件（Chen，2006）[30]对AI伦理主题进行了时区图可视化分析（谢洪明等，2019[31]；黄敏学和吕林祥，2022[32]）。具体而言，本文将上述步骤获得的文献数据导入CiteSpace，随后采用“Time Zone”时区网络视图方法对关键词进行了纵向时间序列分析（如图1所示）。根据时区图谱，AI伦理演进路径呈现出循序渐进且不断深化的发展过程，其主要脉络可归纳为“AI伦理问题认知—细分领域应用—治理标准”三个阶段。

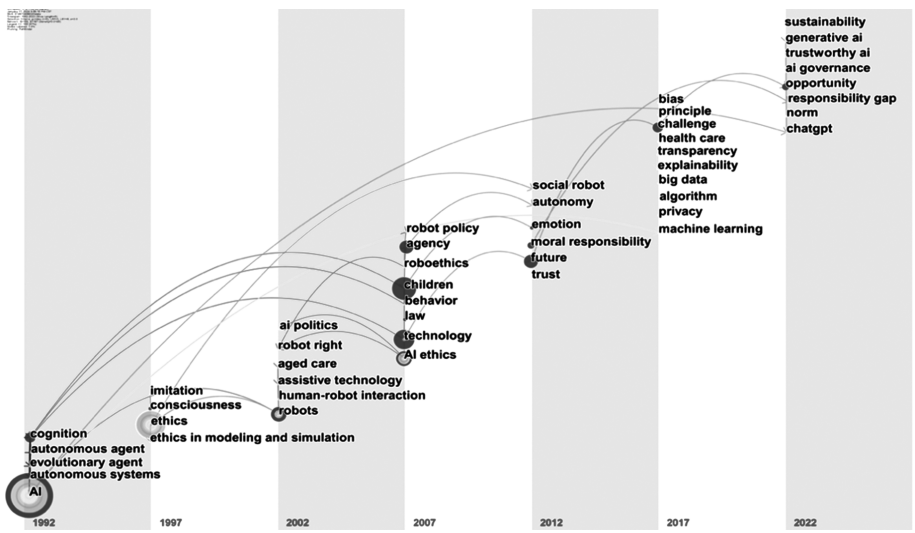


**图1 AI伦理研究关键词时区图谱**

（1）AI技术发展初期的非交互式伦理挑战。在AI发展初期阶段（1992—2002年），研究主要聚焦于技术创新和改进，对AI伦理的讨论局限于非交互过程中的技术与用户认知层面。20世纪中期，随着AI概念的提出，计算机科学开启了这一新兴领域。然而，这一初期阶段的AI技术主要依赖预设规则和算法，不涉及人机交互。因此，AI伦理问题主要体现在隐私保护等静态的伦理担忧（Smith等，1996）[33]。在技术层面，研究者们关注AI设计中的伦理风险，如隐私保护、准确性及可访问性（Burgoon等，2000）[34]，并尝试通过模型优化和自动化AI代理设计来减少潜在的伦理问题（Berdichevsky和Neuenschwander，1999）[35]。这些研究揭示了AI在非交互过程中引发的伦理反思，为后续的AI伦理研究奠定了基础。

（2）人机交互初期的伦理问题探索。2002—2012年，数字技术的快速发展推动了AI从执行预设任务向机器学习和数据驱动方法迈进。这一阶段，AI在医疗、教育、服务等行业实现了初步的人机交互，其社会影响逐步显现。学者们开始探索AI在不同情境中的应用，如老年护理机器人（Karunanithi，2007）[25]、教育机器人（Tanaka和Kimura，2010）[36]及社交机器人（Broekens等，2009）[20]等。

同一时期，随着人们在交互中与AI深入接触，AI逐渐从单纯的任务执行工具演变为具备社会角色的主体（Kim和Mcgill，2024）[2]，这一转变引发了诸多伦理问题。例如，AI在执行任务过程中出现错误或造成损害时，责任归属难以界定，尤其是在复杂的交互场景中，决策通常涉及多个参与主体。此外，人们对AI的信任和依赖增加，这可能削弱人类的自主决策能力。然而，该阶段的相关研究仍集中于AI技术的可行性和交互方式，上述伦理问题尚未得到充分探讨。尽管伦理问题尚未成为研究焦点，伦理准则研究为后续AI伦理规范的法律和政策制定提供了理论基础（Haidt，2007）[27]。

（3）AI伦理问题与治理：涵盖交互与非交互过程。2012年至今，AI在深度学习、虚拟现实和语

言模型等领域取得了显著突破,应用范围从预测型AI延伸至生成式AI(Hermann和Puntoni,2024)[37]。与此同时,AI的进展也带来了复杂的伦理问题和治理挑战。因此,学者们更加重视AI的伦理问题,并将治理纳入讨论,推动AI伦理研究不断深化。具体而言,在非交互场景中,首先,AI的复杂性会导致操作透明度与可解释性下降(Castelvecchi,2016)[38]。其次,大规模数据的使用加剧了隐私担忧(Schmidt等,2020)[39]。最后,算法偏见和歧视问题会威胁社会公平性(Akter等,2021)[9]。在交互场景中,AI自主决策的发展使责任分配成为研究重点(Gill,2020)[8]。研究还致力于构建包括透明性、公正性、平等原则和隐私保护的伦理框架(Hagendorff,2021)[40],并提出多层次的治理方案(Alkire等,2024)[18]。

**4.AI伦理研究总结**

综上所述,过去30多年间,AI伦理逐渐成为研究焦点,其发展呈现出以下演变趋势与局限:首先,关于伦理标准的讨论经历了从关注个体认知与行为,向系统性规范AI研发与应用转变,这一转变不仅反映了学术界对AI伦理的关注,也体现了业界与政策制定者在应对AI伦理风险方面的现实需求。其次,尽管现有研究涵盖了人机非交互与交互过程中的伦理议题,但多集中于特定场景或单一问题,缺乏系统性的归纳与机制层面的解释。这一现状导致不同应用情境下伦理问题的共性与差异尚未得到揭示。

## 三、AI伦理问题及其后果

**1.人类与AI的非交互与交互过程**

AI的核心功能可归纳为四类:数据获取与利用、算法分类与预测、生产代理与学习顾问、社交与互动(Puntoni等,2021)[5]。这四类功能涉及AI的各级使用者(包括企业、用户和政府)以及潜在的被影响者和治理者。为了厘清伦理问题的后果所产生的路径,本文将区分AI在实现不同功能过程中所涉及的人机关系类型,并对应分析其引发的伦理问题及潜在后果。

非交互过程主要涉及的AI功能为数据获取与利用、算法分类与预测。这一过程主要以企业为参与主体,尽管不涉及直接的人机互动,但依然带来了伦理和治理方面的挑战。根据归纳分析,非交互过程中的AI伦理问题主要体现为:在数据获取与利用方面,AI可能引发数据安全与隐私泄露、个体自主控制等问题;在算法分类与预测方面,AI可能引发算法错误分类、算法歧视及同质信息陷阱等问题。

交互过程主要涉及的AI功能为生产代理与学习顾问、社交与互动。由于交互过程涉及人类的直接参与,其伦理问题相较于非交互过程更具动态性,进而衍生出更为复杂和广泛的伦理问题。Köbis等(2021)[41]将导致AI伦理问题的人机关系划分为四种交互模式:AI为人类服务/工作(delegate)、建议与指导(advisor)、观察学习(role model)、人机协作(partner)。基于上述四种交互模式,交互过程中的AI伦理问题主要体现为:在生产代理与学习顾问方面,AI可能引发代理决策伤害、责权分布冲突及威胁人类等问题;在社交与互动方面,AI可能引发独特性忽视、负面道德榜样、虚假联结及道德感缺失等问题。

非交互过程和交互过程呈现出不同的伦理问题。为了全面理解这些问题的根源及其深远影响,本文还深入探讨了这些伦理问题如何通过作用于个体而进一步引发后果。表2列示了这些伦理问题在社会层面引发的后果及其产生机制。

**2.非交互过程中的AI伦理问题及其后果**

(1)数据获取与使用:服务与利用。在信息化社会,AI可能需依赖对包含个人隐私的数据进行处理与分析,以实现其高效功能。因此,AI在提供便利服务的同时,用户的个人隐私数据也有可能被不当利用。这凸显了两大伦理问题:一是数据安全与隐私;二是个体自主控制。

表 2　　非交互与交互过程中的 AI 伦理问题及其后果

| 人机关系 | AI 功能 | 前因 | 伦理问题 | 解释机制 | 伦理问题的后果 |
|---|---|---|---|---|---|
| 非交互过程 | 数据获取与利用 | 敏感数据收集与处理<br>缺乏透明性与可解释性 | 数据安全与隐私 | 隐私担忧 | 社会不平等<br>用户信任危机 |
| | | AI 自主性 | 个体自主控制 | 控制感缺失 | 自由剥夺<br>用户亲社会行为减少 |
| | 算法分类与预测 | 算法预测结果偏差 | 算法错误分类 | 自我认同 | 社会不平等<br>用户身份冲突 |
| | | 歧视特殊群体<br>差异化定价 | 算法歧视 | 愤怒情绪 | 社会不平等<br>社会系统合理化降低 |
| | | AI 输出信息单一<br>算法基于用户偏好筛选信息 | 同质信息陷阱 | 认知趋同 | 信息茧房效应<br>社会极化<br>社会多样性削弱 |
| 交互过程 | 生产代理与学习顾问 | AI 自主性<br>透明性与可解释性<br>人类责任归咎倾向 | 权责分布冲突 | 道德推脱 | 责任归属不明<br>伦理责任弱化 |
| | | AI 自主性 | 代理决策伤害 | 控制感缺失 | 损害社会安全<br>用户健康问题 |
| | | AI 自主性<br>AI 外观和能力上的“类人化” | 威胁人类 | 现实威胁<br>身份威胁 | 消费者不良的补偿性行为<br>员工心理健康和绩效受损<br>职场人际不和谐 |
| | 社交与互动 | AI 缺乏人类情感 | 独特性忽视 | 情绪体验缺失 | 用户健康问题<br>社会系统合理化降低 |
| | | 人类缺乏道德判断 | 负面道德榜样 | 模仿倾向 | 不道德行为增加<br>儿童道德认知发展受阻 |
| | | 互动真实性 | 虚假联结 | 情感依赖 | 用户社交能力下降<br>社会关系网络被破坏 |
| | | 交互过程缺乏情感深度 | 道德感缺失 | 情绪体验缺失 | 用户亲社会行为减少<br>社会伦理约束弱化 |

数据安全与隐私是指企业对 AI 不当监管可能造成用户的数据泄露和隐私风险。例如，在医疗领域，AI 需要处理大量敏感的健康数据，引发了人们对数据泄露、未经授权访问和滥用个人信息的担忧（Price 和 Cohen，2019）[42]。此外，面部识别、位置追踪等技术的广泛应用，显著提高了用户隐私被侵犯的风险。从后果来看，首先，数据信息的不对称可能导致企业与用户之间的不公正交易，如电商平台的信息以商家利益为导向（Xiao 和 Benbasat，2011）[43]，加剧了社会不平等。其次，用户的隐私担忧、个人数据的滥用或未经同意的共享，会使用户对数字生态系统产生不信任（Ayaburi 和 Treku，2020）[44]，从而引发用户信任危机。

个体自主控制是指在非交互过程中，用户对 AI 决策机制缺乏了解与干预能力，从而导致个人数据所有权和决策主导性的丧失。具体而言，研究表明，AI 获取数据的方式具有侵入性和不透明性，因此，用户无法了解或访问背后的推理过程和原理，被称为“黑箱”效应（Castelvecchi，2016）[38]。AI 的“黑箱”运作方式会使用户感到控制感缺失（Puntoni 等，2021）[5]。控制感是指个体对自己行为及其结果的掌控感知（许丽颖等，2024）[45]。由于 AI 在非交互过程中无需用户输入即可通过学习和调整来适应环境并完成任务（De Freitas 等，2023）[46]，其高度自主性可能削弱个体对自身数据和决策的控制感。从后果来看，首先，缺乏控制感会导致自由剥夺，损害伦理准则中的自由主义（Haidt，2007）[27]。其次，在行为层面，研究发现，当算法基于用户数据来提供个性

化的慈善广告时，用户认为AI取代了他们的自主决策过程，从而导致了其捐赠意愿的下降（Lv和Huang，2024）[47]。

（2）算法分类与预测：理解与误解。算法的分类与预测涉及数据输入、模式识别和结果输出。然而，这一过程中，算法的输出依赖于基础数据和模型假设，可能引入偏见或错误。此外，识别并强化历史信息的预测模式可能会缩小信息范围，形成信息同质化。这凸显了三大伦理问题：一是算法分类错误；二是算法歧视；三是同质信息陷阱。

算法错误分类是指AI利用预测模型将用户归类为某一错误类型（Puntoni等，2021）[5]，可能对特定群体产生不公平影响。具体而言，在招聘、信贷审批或司法判决中，算法可能基于性别、种族或年龄等特征错误地定义申请人（Akter等，2021）[9]。例如，一项基于大数据分析的研究表明，即使谷歌广告在设置中引入了“Rather Not Say”和“Custom”等性别隐私选项，算法仍按照传统性别类别划分用户并进行推荐（Shekhawat等，2019）[48]。从后果来看，首先，算法错误分类会直接阻碍某些用户在就业选择等方面的个体权益，造成社会不平等。其次，基于“以人为本”的伦理准则，被算法错误分类的用户还可能因无法获得身份认同而感到身份冲突。

算法歧视是指在不考虑个体特征的情况下，AI基于某个类别或多数群体的成员身份提供不公平的待遇（Kordzadeh和Ghasemaghaei，2022）[49]。传统意义上的歧视通常由掌握权力的人类决策者所为。然而，随着AI在社会各领域的广泛应用，算法歧视逐渐显现。例如，亚马逊曾开发了一款基于机器学习的招聘算法，该算法通过识别与女性相关的关键词来筛选求职简历。这种组织内性别歧视引发了愤怒情绪（Bigman等，2023）[50]。在消费应用领域，住宿平台和网约车平台利用算法对不同标签群体消费者进行差异化定价（李丹，2021）[51]，也会引发愤怒情绪。从后果来看，首先，算法歧视加剧了系统性不平等，使得某些群体持续处于不利地位。其次，受到歧视的用户可能因愤怒情绪而降低对社会系统的合理化评价。

同质信息陷阱是指用户的认知多样性受到局限。认知多样性指的是人们在思考和处理问题时方式的差异（Messeri和Crockett，2024）[52]。Messeri和Crockett（2024）[52]指出，AI的广泛应用可能导致单一化的研究方法和观点占据主导地位，进而降低科学领域的认知多样性。类似地，生成式AI依据自动化数据分析功能生成的内容类似，使得用户接触到的信息缺乏多样性和创造性（Park，2024）[53]。在社交媒体和个性化推荐系统应用中，算法根据用户的历史行为和偏好来筛选和推送信息，这种缺乏自主选择而直接接受输出的过程也可能导致用户仅接触与自己观点一致的信息（Cinelli等，2021）[54]。从后果来看，首先，由于多样化信息本是认知多样性的重要来源，单一信息接受可能会带来“信息茧房”效应（Kitchens等，2020）[55]。其次，当算法不断强化用户现有的信念和偏见时，可能导致其对不同或相反观点的接受能力下降，进而加剧了社会极化（Kitchens等，2020）[55]。最后，社交网络中的算法推荐使得用户主要与持有相似观点的人交流，进一步加剧群体内部的同质性（Cinelli等，2021）[54]，对社会的整体多样性构成潜在威胁。

**3.交互过程中的AI伦理问题及后果**

（1）生产代理与学习顾问：授权、被操控与被替代。首先，AI的授权问题涉及决策权和责任的分配，当AI被赋予一定的决策权时，可能导致责权分布冲突。其次，AI操控系统的行为可能因算法偏差或数据缺陷对人类造成伤害。最后，AI的类人表现可能引发身份威胁和现实威胁。这凸显了三大伦理问题：一是责权分布冲突；二是代理决策伤害；三是威胁人类。

权责分布冲突是指在AI参与的决策过程中，如果AI出现错误或造成不良后果，往往难以追踪溯源、明确责任。例如，自动驾驶汽车发生交通事故时，责任应归于制造商、车辆所有者还是AI本身，成为一大争议（Gill，2020[8]；Bonnefon等，2016[56]）。生成式AI创作的虚假新闻和不良视频也引发了类似的责任归属问题（Garimella和Chauchard，2024）[57]。解决责权分布冲突需使得AI在决策

过程中具备较高的透明度和可解释性。然而,延续了非交互过程的“黑箱”效应,AI在人机交互中的透明度和可解释性不足,且人们对AI决策机制的了解往往是有限的(Cadario等,2021)[26],从而难以追踪和解释某项决策为何被做出。从后果来看,首先,AI决策过程缺乏可追踪性和可解释性,可能在法律层面上导致权责归属的不明确。其次,权责分布冲突可能导致用户进行责任推脱,即将不当行为归咎于AI(Gill,2020)[8]。且研究指出,随着AI参与度的提高,这种责任推脱的倾向也愈发明显(Bigman等,2019)[58]。

代理决策伤害是指AI做出的不道德决策可能给人类带来身体或心理上的伤害。具体而言,当AI作为生产代理时,其执行的任务可能包括需要做出道德判断的复杂决策(Köbis等,2021)[41],如在无人驾驶汽车、机器人外科医生、儿童看管及老人护理等领域的道德决策。面对复杂、新颖的决策情境时,AI的不道德决策可能带来严重后果,如交通事故(Gill,2020)[8]、对弱势群体的不当护理(Broekens等,2009)[20],或战争中的致命伤害(Formosa和Ryan,2021)[59]。从后果来看,一方面,AI带来的代理决策伤害损害了社会安全;另一方面,由于AI代理在人机交互中仍具有一定的自主性,人类作为决策的接受者,无法干预或改变其过程(Formosa和Ryan,2021)[59],这种被动接受削弱了人们的控制感,进而引发心理不安全感。

威胁人类是指人们在面对AI时所产生的现实威胁或身份威胁感。具体而言,随着技术的发展,AI在外观上更加接近人类,在多项能力上甚至超越了人类(Hermann和Puntoni,2024)[37],虽然这些进步为人类带来了便利,但也引发了人类对其的威胁感,主要体现在现实威胁,如对资源竞争的威胁,和身份威胁,如对人类自身独特性的认同威胁(谢才凤等,2023)[60]。首先,当AI被视为竞争性的劳动替代品时,人们会感受到现实威胁(Jia等,2024[23];Leong等,2025[61])。早期研究提出,当自动化系统的建议与人类专家的判断相冲突时,专家可能会因为感到威胁而防御性地忽视系统建议(Elkins等,2013)[62]。其次,威胁还来源于对人类自身独特性的认同威胁。具体来说,人类倾向于认为自己是与其他群体截然不同的特殊存在,但类人化的AI外形和能力可能模糊机器与人类之间的界限。例如,实证研究表明,与人形机器人交互会引发参与者的身份威胁感(Yogeeswaran等,2016)[63]。

感知威胁还可能带来一系列后果。首先,在消费场景中,消费者与服务机器人交互后可能会感到身份威胁,进而采取补偿性消费行为以缓解心理落差(Mende等,2019)[64]。例如,消费者会进行炫耀性消费,这不仅会导致财务状况恶化,还可能助长物质主义价值观。Mende等(2019)[64]还发现,消费者感知到身份威胁后会摄入更多食物,这会带来健康风险。其次,在组织中,AI的应用会增加员工的心理不安全感,并进一步引发职业倦怠和工作中的不文明行为(Yam等,2023)[65]。最后,无论是现实威胁还是身份威胁,均可能导致员工在职场中将他人视为工具的物化倾向,从而以对待物的方式对待他人(许丽颖等,2024)[45],这一现象可能会导致职场中人际关系的不和谐。

(2)社交与互动:疏远与联结。当AI参与社交与互动时,一方面,疏远的人机关系可能使得人们感到自己的独特性被忽视;另一方面,人们与AI建立的深度心理联结也可能造成负面影响,如用户模仿AI的不道德行为,人际间的真实接触减少,用户无法体验到道德情绪。这凸显了四大伦理问题:独特性忽视、负面道德榜样、虚假联结以及道德感缺失。

独特性忽视是指AI在提供服务时,未能充分考虑用户的个体差异,包括其独特的生物学、心理学和行为学特征(Longoni等,2019)[66]。尽管AI在人机交互中仍具有一定的自主性,但需要指出的是,AI并不具有主观意识,自主性是AI基于算法实现的任务独立执行能力,而非涉及自我认知和真实情感反应的人类特质。因此,用户与AI交互时无法感受到人类特有的情感和理解。从后果来看,情绪体验的缺失一方面会带来用户心理健康问题(De Freitas和Puntoni,2023)[67];另一方面还

会降低社会系统合理化。例如,在医疗领域,使用AI医生诊断会降低患者对医疗决策的满意度和信任度(Longoni等,2019)[66],从而降低其对社会系统合理化的感知。

负面道德榜样是指在人机交互过程中,人们观察并学习AI表现出的不道德或不当行为,从而形成负面道德榜样效应。儿童正处于道德认知的发展阶段,对社会关系和道德准则的理解尚未完全成熟,因此AI的负面道德榜样容易对儿童产生影响。例如,机器人展示出过度顺从的行为,可能导致儿童模仿这种顺从的仆人角色(邓士昌等,2023)[21]。从后果来看,首先,人们的不道德行为可能增加。其次,人们还可能会将不道德行为视为可接受或常态化的行为,进而阻碍其道德认知发展。

虚假联结是指人们与AI交互时可能与之建立起虚假的联结关系,进而减少人际间的真实接触。例如,研究发现,儿童可以与机器人建立亲密关系并对之产生情感依赖。此外,人们与社交机器人、聊天助手、虚拟化身(avatar)等互动来满足其社交需求,这种虚假联结可使人们逐渐依赖于虚拟系统。从后果来看,首先,长期的情感依赖会削弱儿童在真实社交场合中与同龄人互动的能力(Kahn Jr等,2013)[68]。其次,对于成年人而言,虚假联结可能会破坏健康的社会关系网络。例如,研究表明,AI在交互中展示出强大的社会情感能力会增加人们对AI的"人类化"感知,但同时因将机器与人类同化,减少了其对人类的"人类化"认知,从而引发对他人的不公平或不道德对待(Kim和Mcgill,2024)[2]。

道德感缺失是指人机交互过程缺乏对人类内疚、羞愧等道德情绪的有效激发,从而削弱个体的道德感知。虽然在上述分析中,AI的模拟交互在一定程度上模糊了人类与机器之间的界限,然而,从道德情绪的角度来看,人机互动依然与人际间的交互存在根本上的差异。一项功能性磁共振成像(fMRI)神经研究发现,尽管医疗AI使用个性化对话的模式与人们进行交互,人们大脑中与冷漠情绪相关的区域仍会被激活,而与人类医生互动时则会激活与亲社会相关的脑区(Yun等,2021)[69]。从后果来看,首先,这种道德相关情绪体验的缺失可能会降低亲社会行为,进而泛化社会冷漠(Zhou等,2022)[24]。例如,Chen(2023)[70]的研究证实了道德情绪对推动亲社会行为的重要性:当参与者观看机器人在危险中实施救助行为(如在泥石流中救援受害者)时,其勇气感知较低,进而降低了其在后续任务中的亲社会行为。其次,内疚、羞愧等道德情绪的缺失,还会削弱社会伦理规则和道德标准的约束力(Novak,2020)[71]。例如,研究表明,当这些情绪减弱时,人们更容易表现出不诚实和成果剽窃等不道德行为(Kim等,2023)[72]。

## 四、AI伦理问题的理论框架构建与治理对策

### 1.AI伦理问题的共性与差异

在第三部分对非交互过程与交互过程中AI伦理问题进行归纳分析的基础上,本部分将进一步探讨两者在伦理问题上的共性与差异,旨在厘清不同应用情境下AI伦理问题的核心特征与机制,从而为后续AI伦理规范的制定与实践提供理论支持。

(1)非交互过程和交互过程中伦理问题的共性。非交互过程和交互过程中的伦理问题主要具有三个共性:一是AI自主性为伦理问题的共性来源。自主性使AI能够在特定情境下自主运行并做出决策,但这一能力通常缺乏足够的透明性与可解释性,使人类难以全面理解其决策逻辑与操作机制。这种技术特性导致了伦理问题的扩散。例如,无论是算法歧视、权责分布冲突,还是对人类的威胁,这些问题都指向AI自主性所带来的不可控性与不可预测性。因此,AI自主性不仅是技术进步的重要标志,也成为当前伦理问题普遍存在的结构性来源。二是AI权责分布问题贯穿AI非交互与交互过程。在AI的应用过程中,责任归属的模糊性是一个核心难题。非交互过程中的数据隐私问题,交互过程中的道德推诿现象,实质上都反映了权责分布难以清晰界定的困境。一方

面,AI系统在技术层面缺乏透明性与可解释性,使得其决策过程难以被有效监督,且AI本身无法承担由其行为所引发的后果,从而导致用户对AI的信任缺失;另一方面,在多方共同参与的情境中,责任的重叠与不明确可能诱发推诿行为,加剧后果。例如,当AI系统导致实际损害时,责任究竟应由系统设计者、操作用户,还是AI系统本身承担,往往缺乏明确的判定标准。这种权责不清的局面不仅影响AI的可持续使用,也对现有法律与伦理体系构成了挑战。三是伦理问题的后果具有普遍性。非交互过程和交互过程中的伦理问题不仅直接损害了用户的利益,还可能对整个社会系统造成深远影响。例如,隐私问题可能侵蚀社会信任基础,算法歧视可能加剧社会不平等,责权分布冲突可能导致社会整体的伦理责任弱化,这些后果跨越了具体应用场景,形成了对社会层面伦理规范和法律框架的普遍挑战。

(2)非交互过程和交互过程中伦理问题的差异。非交互过程和交互过程中的伦理问题主要具有四个差异:一是问题焦点和参与方不同。非交互过程中的伦理问题主要集中在AI的应用层面,核心在于数据处理和算法运行所带来的问题,例如数据安全与隐私等。这些问题的主要参与方是企业端。相较之下,交互过程的伦理问题则主要涉及人类行为和社会交往层面,例如威胁人类、负面道德榜样和虚假联结等。这些问题的主要参与方是用户端,即用户在与AI交互的过程中受到AI行为的影响。二是伦理问题前因的侧重点不同。在非交互阶段,伦理问题主要源于技术层面,涉及数据透明性、决策过程的可解释性以及AI的自主性等技术属性。例如,算法的"黑箱"效应作为典型的技术性难题,被视为引发伦理风险的关键因素(Castelvecchi,2016)[38]。相较而言,交互过程中的伦理问题更侧重于社会性与情感因素,如用户的责任推脱倾向等。伦理问题前因的差异反映出企业与用户在不同阶段所承担角色的权重不同:非交互过程强调企业在技术开发与风险控制中的责任,而交互过程则更加强调用户的实际体验与心理反应。这一区分不仅拓展了对AI伦理问题成因的多维理解,也为治理对策提供了差异化的切入路径。三是用户行为反应的动机不同。根据动机社会认知模型(motivated social cognition model),个体的信息加工和决策受认知动机、存在动机和关系动机的驱动(Jost等,2009)[73]。这三种动机为理解人机交互中的用户行为提供了重要理论视角(谢才凤等,2023)[60]。在非交互过程中,由于人机关系的固定特性,用户的行为反应主要受认知动机驱动。例如,AI输出的同质化信息强化了用户的认知趋同倾向,即选择性接触与自身观点一致的信息,同时忽视或回避相悖的信息(Cinelli等,2021)[54]。而在交互过程中,用户的行为反应则主要受存在动机与关系动机驱动。具体而言,存在动机体现了个体对安全感与自我价值的需求(Jost等,2009)[73],而AI外观和能力上的"类人化"威胁了人们对自身独特性的价值判断(Yogeeswaran等,2016)[63]。关系动机则体现了个体对归属感和认同感的需求(Jost等,2009)[73],而虚假联结会增加人们对AI的情感依赖,从而削弱其对人类群体的归属感。四是伦理问题的后果的表现层次不同。非交互过程的后果主要表现为结构性社会问题。例如,数据处理不当可能会加剧用户对技术和企业的信任危机,算法歧视可能导致社会不平等。这些后果通常表现为对社会整体稳定性和公平性的潜在威胁。相较而言,交互过程中的后果则更侧重于人际关系和心理层面,如AI对用户心理健康的影响、社会关系网络的破坏,以及社会伦理约束的削弱等。尽管交互过程中的伦理问题通常表现在微观层面,但其累积效应可能对社会层面产生深远的负面影响。

**2.AI伦理问题的理论框架**

基于上述对AI伦理问题共性与差异的归纳,本文构建了AI伦理问题的理论框架(如图2所示)。该框架的构建逻辑从AI的功能和应用场景出发,沿着作用路径延展至伦理问题及其后果。框架涵盖了企业、用户和政府三个参与主体,体现了技术、个体与治理三者之间的相互作用。

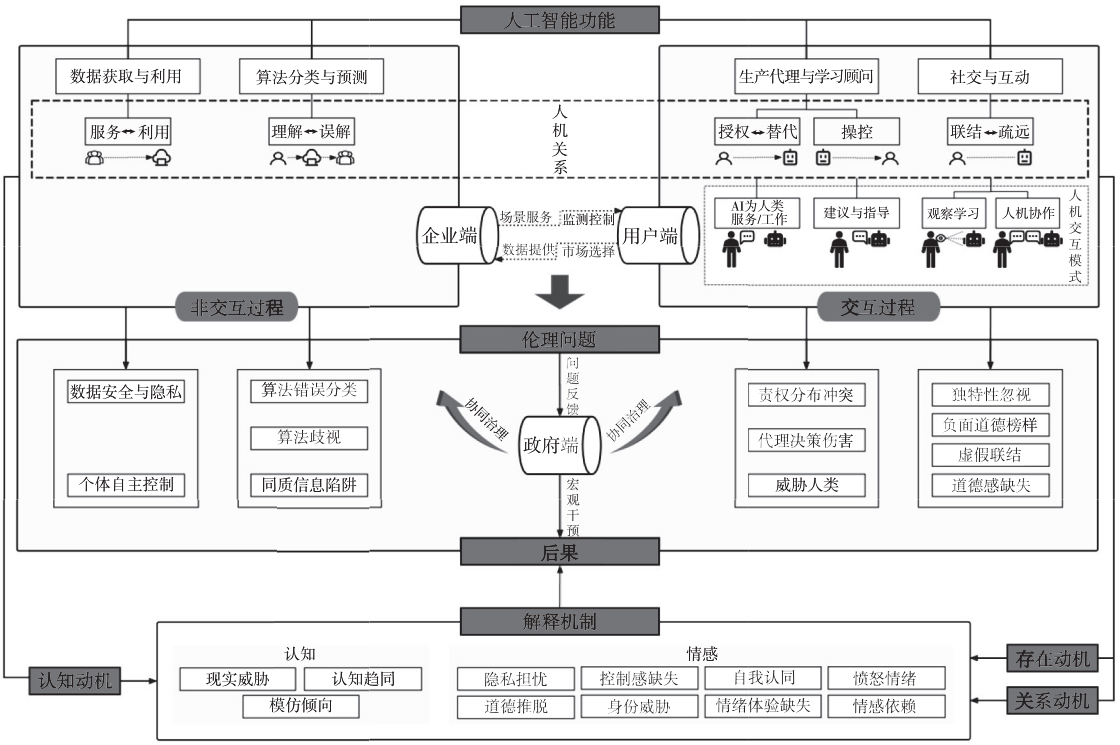


**图2　AI伦理问题的理论框架**

注:箭头方向代表AI应用带来的伦理问题及其后果的作用机制

AI伦理问题的理论框架以“人工智能功能”为核心,将人机关系划分为非交互与交互过程两个模块。这种划分反映了AI在实际应用场景中的功能特性,涵盖了从数据输入到与用户交互的全流程。框架通过企业端、用户端与政府端的动态连接,系统性展现了AI在市场需求与治理机制中的运行逻辑。具体而言,企业通过在具体场景中的技术设计与应用为用户提供AI服务,同时监控并影响用户行为;用户的行为与反馈塑造了技术的实际应用形态;政府则通过监管与政策干预,在协调各方利益的同时,防范潜在的伦理风险。其次,框架进一步强调了伦理问题的后果的多维性,揭示了AI发挥功能的同时对用户与社会整体带来的深远影响。此外,从用户动机和心理机制的角度,框架分析了AI如何塑造用户决策与行为,并通过这些个体影响扩展至社会结构与伦理规范层面。最后,框架突出了政府的反馈接收与干预在治理AI伦理问题中的关键作用。通过对AI伦理问题的宏观干预,政府不仅能够促进技术的规范化发展,还为构建系统性的伦理准则与政策框架提供重要支持,从而确保AI技术在合规的框架下健康发展。

**3.AI伦理问题的治理框架**

为了应对伦理问题,有必要从宏观层面提出治理对策(Awad等,2018)[74]。基于此,本文立足于中国文化语境,提出了AI伦理问题的治理框架(如图3所示)。首先,包含中国特色“权利/责任维度”在内的六个维度的AI道德基础,为该治理体系提供了核心理论支撑。其次,基于“权利/责任”这一道德维度,本文提出应明确人机之间的责权分布。为此,应首先界定AI的道德代理角色,并通过责任评估机制明确责任主体,继而制定相应的监管策略。最后,在道德基础与责权分布明确的前提下,构建以算法社会契约为核心的AI治理框架,强调人机协同与循环机制在优化伦理决策过程中的关键作用。

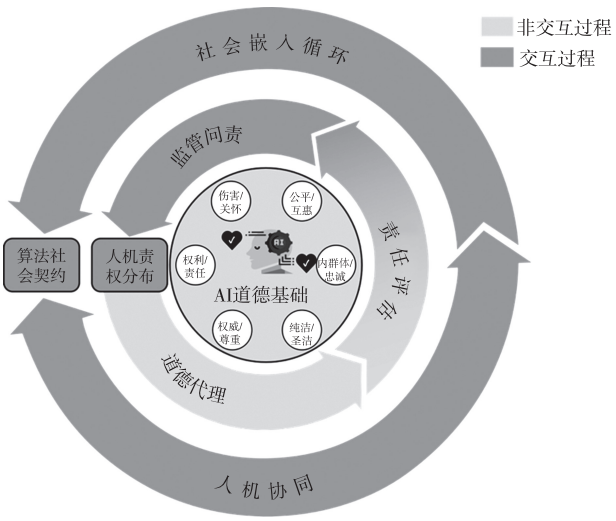


**图 3　AI伦理问题的治理框架**

(1)建立适应中国文化背景的AI道德基础。AI对社会治理和人类福祉的伦理影响根植于其所处社会的道德基础(Bonnefon等,2016[56];Awad等,2018[74])。因此,AI伦理问题的治理对策应以完善的AI道德为基础。本文提出,在构建AI伦理基础框架时,需要充分考虑中国的文化特性和社会治理中的独特价值观。具体而言,Haidt(2007)[27]在其开创性的道德心理学研究中提出,尽管人类普遍具有道德动机,但道德内容因文化而异。这一观点为理解不同文化背景下AI伦理设计提供了理论支持。例如,相比于西方社会的个人主义和社会契约论,中国的伦理观更侧重于群体忠诚、权威尊重与社会责任(王恩界和乐国安,2006)[75]。从AI应用视角来看,现有的偏见检测工具(如微软的Fairlearn Python包)虽然提供了矫正机制,但缺乏文化适应性和具体标准(Fu等,2020)[76]。因此,从非交互层面出发,需要进一步优化和完善AI的应用矫正机制。

为了在中国文化背景下建立AI道德基础,本文认为,可借鉴Haidt(2007)[27]的五维道德基础理论,并引入具有中国特色的权利/责任维度。首先,五维道德基础与中国文化中的一些普世价值观一致。例如,“仁者爱人”与伤害/关怀维度相契合,强调对弱者的保护;公平/互惠维度则体现了“己所不欲,勿施于人”的公正观念;内群体/忠诚与集体主义价值观密切相关,强调对家庭与国家的忠诚;纯洁/圣洁维度则体现了中国文化中的精神净化与道德追求;权威/尊重维度反映了对社会秩序和权威结构的维护。其次,AI也应具备对非合理权威提出质疑的能力,这样能够保证技术创新在伦理和法律的框架内实现。最后,权利/责任维度突出社会责任与道德义务在AI设计中的重要性,这与中国社会治理强调平衡权利与责任的原则一致(肖红军,2022)[19]。

(2)评估并监管人机责权分布。AI道德基础体现了在中国文化背景下,将权利与责任维度融入AI伦理体系尤为重要,为发展负责任的AI提供了社会治理视角。因此,本文从人机责权分布方面提出治理对策。

首先,需要明确的是,从法律角度来看,AI并不具备法律人格,且目前大多研究认为AI不应作为道德代理(Swanepoel,2021)[77]。因此,本文认为,即使AI在非交互过程中能够自主运行,也并不意味着它在道德或法律层面可被视为具有与人类相同的道德权利。

其次,在确认AI不具备道德代理资格后,责任评估的核心在于界定责任主体,从而确保在出现问题时能够有效追责。本文明确了AI自主性是伦理问题产生的重要前因,因此,AI系统的设计应结合AI和用户的自主性。如前文所述,从非交互过程具有完全自主的决策能力和输出能力到交互过程中的主动执行,AI在决策和行动上具有不同程度的自主性。在评估AI自主性时,需要考

虑其技术能力和行为范围。交互过程中,若AI仅作为工具,其责任更多归属于技术开发者或用户。然而,随着AI自主性的增强,其行为在伦理与法律上的责任归属可能需要重新定义。例如,当AI在非交互过程中执行未被明确禁止的指令,且其行为具有自主判断时,责任需部分转移至AI本身。

最后,应建立健全的监督与问责机制,并从用户、企业和政府三个层面进行综合考量。用户在使用AI应用时,应充分认识自身行为对技术生态的潜在影响,自觉遵守道德规范,避免违法或有害行为。企业作为AI技术的研发者与推广者,应承担相应的社会责任,保障技术的安全性、可靠性与合规性。例如,Gartner公司将"AI治理"定义为"为AI技术的应用和使用制定政策、分配决策权并确保企业对风险和投资决策负责的过程",这一观点为企业内部审查机制的构建提供了重要参考。政府则需完善法律法规,明确不同应用领域的AI规范,加强对企业的监管和对违规行为的惩戒,同时加大对基础研究和关键技术攻关的支持,以推动AI技术与社会经济的良性互动。

(3)构建算法社会契约。明确道德基础和人机责权分布后,构建基于社会契约的AI治理框架成为关键。人机协同与循环机制是该框架的核心,强调通过人类与AI的持续互动与反馈优化决策过程。这一治理模式借鉴柯尔伯格(Lawrence Kohlberg)的社会契约理论(Rahwan,2018)[78],即将社会成员的共识转化为技术实践,以确保AI发展符合公众利益与伦理标准。

人机协同(human in the loop)提供了人类与AI合作的基本框架,强调人类作为决策过程中的关键一环,参与到AI的控制和监督中(Rahwan,2018)[78]。在这一过程中,复杂决策需通过人类参与得以优化。一方面,专家嵌入机器学习过程,能够提升决策的伦理性;另一方面,在企业治理中,人机协同不仅能优化生产效率,还能确保技术与人类价值观的契合。社会嵌入循环(society in the loop)是人机协同的扩展和深化,强调将社会层面的价值观和规范纳入到算法治理中,从而形成一个更为全面的协同模式(Rahwan,2018)[78]。社会嵌入循环不仅关注个体的参与,还关注整个社会对算法决策的监督和影响。在社会嵌入循环中,社会作为一个整体参与到算法的监管中,能够确保算法的决策符合社会的价值观和伦理标准。

算法社会契约是连接人机协同和社会嵌入循环的纽带。具体而言,算法社会契约要求将社会的价值观和伦理标准嵌入到算法的设计和运行中,形成一个算法和社会判断相结合的反馈回路。一方面,通过将社会规范编码到AI设计中,循环机制促进了算法社会契约的完善。例如,在供应链管理中,配对(twinning)概念通过特征工程和数据工程,将业务逻辑与专家知识融合到AI技术中,体现了社会契约导向(Hasija和Esper,2022)[79]。另一方面,在非交互过程中,某些社会群体在接受AI服务时可能受到歧视或市场排斥,而算法社会契约能够帮助实现公正权衡(Rahwan,2018)[78]。总体而言,在智能治理的背景下,算法社会契约通过人机协同与循环机制,重构了人类的自主性、权利与责任。

## 五、总结和展望

### 1.研究结论

随着AI技术的快速发展,其伦理问题日益凸显,亟需系统性研究。本文基于文献计量分析与功能分类视角,对AI伦理问题进行了深入探讨。首先,通过文献计量分析揭示,AI在提升生产力的同时,也引发了数据隐私泄露等一系列伦理问题。随后,本文基于AI功能分类,系统分析了人机交互与非交互过程中的伦理问题。其次,本文进一步分析了AI伦理问题在前因、作用机制及后果方面的共性与差异。最后,本文提出了"道德原则-责权评估与监管-社会契约"协同的治理框架。该治理框架以道德原则为核心出发点,经由人机责权分布的界定,延伸至社会契约的系统构建,体现出治理从核心基础到外围动态更新、从非交互到交互过程、从个体参与到群体反馈的递进演化逻

辑,最终构成一个多维协同的系统化框架。

**2.未来研究展望**

鉴于政府、企业和用户在AI发展和应用中的利益诉求各异,除了从宏观治理层面提出缓解AI伦理问题的对策外,未来研究还可从微观层面深入探讨相关议题。

(1)探讨AI设计对降低伦理影响的作用。AI自主性是伦理问题一个共性来源。首先,未来研究可以比较不同自主性水平的AI对人类责任感的影响。通过设计不同程度自主性的AI(如全自动、半自动和辅助型AI),测量参与者在决策错误或不良后果出现时的责任归因。其次,AI权责分布问题贯穿非交互与交互过程,而增加决策过程的透明度和可解释性能够明确责任归因,从而达到减少伦理问题的效果。因此,未来研究可检验提高算法的可解释性,如解释算法的工作原理是否可降低参与者将责任归因于算法的倾向。此外,针对交互过程中"AI威胁人类"问题,未来研究可通过设计实验,操纵AI的外观(如人形程度、面部表情)和行为(如语音语调、互动方式),测量参与者的感知威胁和行为反应。最后,针对道德感缺失问题,在AI主导的交互环境中,研究可设计有效的干预措施以维持或增强人类的道德情绪。例如,可通过引入具有人格化特征的AI反馈来激发内疚、羞愧等道德情绪。

(2)分析伦理问题的后果所产生的心理机制。根据本文分析,人们在与AI进行非交互和交互过程中均会表现出不道德倾向。例如,人们授权委托AI时,有责任推脱倾向;当AI带来威胁感时,人们会采取不良补偿行为;与AI交互后,人们的亲社会行为意愿降低。但目前研究未清晰廓清其中的心理机制。尽管本文通过分析和归纳,将心理机制分为情感和认知两个方面,但缺乏实证验证。为了揭示这些不道德倾向的内在原因,本文建议未来研究深入分析伦理问题的后果所产生的心理机制。首先,未来研究可以分析责任归因、道德推脱和责任感对人们态度或行为反应的解释作用。其次,如前文总结,控制感缺失是解释人们亲社会行为减少的心理机制。未来研究可借助功能性磁共振成像(fMRI)等技术识别个体与AI交互时的大脑活动,比较人机交互和人际交互中个体大脑活动区域的差异,以及在不同心理状态下与AI交互所引发的大脑活动区域差异。

(3)探讨降低AI伦理影响的边界条件。根据前文的文献计量结果,自2012年社交机器人引发广泛关注以来,随着其功能的不断提升,应用领域也持续拓展。然而,目前缺乏研究讨论社交机器人在应对复杂的社会情境和人际关系时可能引发的伦理问题。首先,未来研究可探讨社交准则对人们如何使用社交机器人的影响。研究可通过文献综述、专家访谈和焦点小组讨论以确定人际关系中惯用的社交准则,并通过设计实验来考察持有不同社交准则的人是否在使用社交机器人时有不同表现。其次,针对人机交互的责任推脱倾向,研究还可以探讨如何通过设计提示机制等,有效减少个体的责任推脱倾向。最后,前文总结了AI造成威胁感后引发的人类不道德行为,未来研究可以探讨AI对人类产生威胁感的边界条件。例如,可考虑AI的角色定位及个体的自我效能感或控制欲望等因素,如何调节其威胁感的产生与后续行为反应。

**参考文献**

[1]Martınez-Miranda,J.,and A.Aldea.Emotions in Human and Artificial Intelligence[J].Computers in Human Behavior,2005,21,(2):323-341.

[2]Kim,H.Y.,and A.L.Mcgill.AI-Induced Dehumanization[J/OL].Journal of Consumer Psychology,2024,https://doi.org/10.1002/jcpy.1441.

[3]Forbes.How Technology Will Revolutionize Marketing in 2022 and Beyond [EB/OL].(2022-02-04)[2024-12-10].https://www.forbes.com/councils/forbesagencycouncil/2022/02/04/how-technology-will-revolutionize-marketing-in-2022-and-beyond/.

[4]Huang,M.H.,and R.T.Rust.A Strategic Framework for Artificial Intelligence in Marketing[J].Journal of the Academy of

Marketing Science,2021,49:30-50.

[5]Puntoni,S.,R.W.Reczek,and M.Giesler,et al.Consumers and Artificial Intelligence:An Experiential Perspective[J].Journal of Marketing,2021,85,(1):131-151.

[6]Ebrahimi,S.,M.Ghasemaghaei,and I.Benbasat.The Impact of Trust and Recommendation Quality on Adopting Interactive and Non-Interactive Recommendation Agents:A Meta-Analysis[J].Journal of Management Information Systems,2022,39,(3):733-764.

[7]Wiener,N.Some Moral and Technical Consequences of Automation:As Machines Learn They May Develop Unforeseen Strategies at Rates That Baffle Their Programmers[J].Science,1960,131,(3410):1355-1358.

[8]Gill,T.Blame It on the Self-Driving Car:How Autonomous Vehicles Can Alter Consumer Morality[J].Journal of Consumer Research,2020,47,(2):272-291.

[9]Akter,S.,G.Mccarthy,and S.Sajib,et al.Algorithmic Bias in Data-Driven Innovation in the Age of AI[J].International Journal of Information Management,2021,60,102387.

[10]李舒沁,王灏晨,汪寿阳.人工智能背景下制造业劳动力结构影响研究——以工业机器人发展为例[J].北京:管理评论,2021,(3):307-314.

[11]赵志耘,徐峰,高芳,赵志耘,徐峰,高芳,李芳,侯慧敏,李梦薇.关于人工智能伦理风险的若干认识[J].北京:中国软科学,2021,(6):1-12.

[12]Thompson,H.S.Computational Systems,Responsibility,and Moral Sensibility[J].Technology in Society,1999,21,(4):409-415.

[13]Nass,C.,and Y.Moon.Machines and Mindlessness:Social Responses to Computers[J].Journal of Social Issues,2000,56,(1):81-103.

[14]Song,C.S.,and Y.K.Kim.The Role of the Human-Robot Interaction in Consumers' Acceptance of Humanoid Retail Service Robots[J].Journal of Business Research,2022,146:489-503.

[15]Ayala,F.J.The Difference of Being Human:Morality[J].Proceedings of the National Academy of Sciences,2010,107,(supplement_2):9015-9022.

[16]Jones,T.M.Ethical Decision Making by Individuals in Organizations:An Issue-Contingent Model[J].Academy of Management Review,1991,16,(2):366-395.

[17]Kazim,E.,and A.S.Koshiyama.A High-Level Overview of AI Ethics[J].Patterns,(New York,N.Y.),2021,2,(9):100314.

[18]Alkire,L.,A.Bilgihan,and M.Bui,et al.Raise:Leveraging Responsible AI for Service Excellence[J].Journal of Service Management,2024,35,(4):490-511.

[19]肖红军.算法责任:理论证成、全景画像与治理范式[J].北京:管理世界,2022,(4):200-226.

[20]Broekens,J.,M.Heerink,and H.Rosendal.Assistive Social Robots in Elderly Care:A Review[J].Gerontechnology,2009,8,(2):94-103.

[21]邓士昌,林子涵,陆昱谦,李象千.智能时代的新玩伴:儿童与机器人的互动特征及其对儿童发展的影响[J].北京:心理科学进展,2023,(12):2319-2336.

[22]Luo,X.,M.S.Qin,and Z.Fang,et al.Artificial Intelligence Coaches for Sales Agents:Caveats and Solutions[J].Journal of Marketing,2021,85,(2):14-32.

[23]Jia,N.,X.Luo,and Z.Fang,et al.When and How Artificial Intelligence Augments Employee Creativity[J].Academy of Management Journal,2024,67,(1):5-32.

[24]Zhou,Y.,Z.Fei,and Y.He,et al.How Human-Chatbot Interaction Impairs Charitable Giving:The Role of Moral Judgment[J].Journal of Business Ethics,2022,178,(3):849-865.

[25]Karunanithi,M.Monitoring Technology for the Elderly Patient[J].Expert Review of Medical Devices,2007,4,(2):267-277.

[26]Cadario,R.,C.Longoni and,C.K.Morewedge.Understanding,Explaining,and Utilizing Medical Artificial Intelligence[J].Nature Human Behaviour,2021,5,(12):1636-1642.

[27]Haidt,J.The New Synthesis in Moral Psychology[J].Science,2007,316,(5827):998-1002.

[28]Hagendorff,T.The Ethics of AI Ethics:An Evaluation of Guidelines[J].Minds and Machines,2020,30,(1):99-120.

[29]Floridi,L.Soft Ethics and the Governance of the Digital[J].Philosophy & Technology,2018,31,(1):1-8.

[30]Chen,C.Citespace II:Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature[J].Journal of the American Society for Information Science and Technology,2006,57,(3):359-377.

[31]谢洪明,陈亮,杨英楠.如何认识人工智能的伦理冲突?——研究回顾与展望[J].上海:外国经济与管理,2019,(10):109-124.

[32]黄敏学,吕林祥.心理契合视角下智能产品营销研究的评述与展望[J].北京:经济管理,2022,(7):193-208.

[33]Smith, H.J., S.J.Milberg, and S.J.Burke.Information Privacy: Measuring Individuals' Concerns About Organizational Practices [J].MIS Quarterly,1996,20,(2):167-196.

[34] Burgoon, J. K., J. A. Bonito, andB. Bengtsson, et al. Interactivity in Human-Computer Interaction: A Study of Credibility, Understanding, and Influence[J].Computers in Human Behavior,2000,16,(6):553-574.

[35]Berdichevsky, D., andE.Neuenschwander.Toward an Ethics of Persuasive Technology[J].Communications of the ACM,1999,42, (5):51-58.

[36]Tanaka, F., and T.Kimura.Care-Receiving Robot as a Tool of Teachers in Child Education[J].Interaction Studies,2010,11, (2):263-268.

[37] Hermann, E., and S. Puntoni. Artificial Intelligence and Consumer Behavior: From Predictive to Generative AI[J]. Journal of Business Research,2024,180,114720.

[38]Castelvecchi, D.Can We Open the Black Box of AI?[J].Nature,2016,538,(7623):20-23.

[39]Schmidt, L., R.Bornschein, and E.Maier.The Effect of Privacy Choice in Cookie Notices on Consumers' Perceived Fairness of Frequent Price Changes[J].Psychology & Marketing,2020,37,(9):1263-1276.

[40] Hagendorff, T. Linking Human and Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning[J].Minds and Machines,2021,31,(4):563-593.

[41]Köbis, N., J.F.Bonnefon, and I.Rahwan.Bad Machines Corrupt Good Morals[J].Nature Human Behaviour,2021,5,(6):679-685.

[42]Price, W.N., andI.G.Cohen.Privacy in the Age of Medical Big Data[J].Nature Medicine,2019,25,(1):37-43.

[43]Xiao, B., andI.Benbasat.Product-Related Deception in E-Commerce: A Theoretical Perspective[J].MIS Quarterly,2011,35, (1):169-195.

[44] Ayaburi, E.W., and D.N.Treku.Effect of Penitence on Social Media Trust and Privacy Concerns: The Case of Facebook[J]. International Journal of Information Management,2020,50:171-181.

[45]许丽颖,王学辉,喻丰,彭凯平.感知机器人威胁对职场物化的影响[J].北京:心理学报,2024,(2):210-225.

[46]De Freitas, J., S.Agarwal, and B.Schmitt, et al.Psychological Factors Underlying Attitudes toward AI Tools[J].Nature Human Behaviour,2023,7,(11):1845-1854.

[47] Lv, L., and M. Huang. Can Personalized Recommendations in Charity Advertising Boost Donation? The Role of Perceived Autonomy[J].Journal of Advertising,2024,53,(1):36-53.

[48]Shekhawat, N., A.Chauhan, and S.B.Muthiah, et al.Algorithmic Privacy and Gender Bias Issues in Google Ad Settings[A]. Proceedings of the 10th ACM Conference on Web Science[C].New York:ACM Digital Library,2019.

[49]Kordzadeh, N., and M.Ghasemaghaei.Algorithmic Bias:Review, Synthesis, and Future Research Directions[J].European Journal of Information Systems,2022,31,(3):388-409.

[50] Bigman, Y. E., D. Wilson, and M. N. Arnestad, et al. Algorithmic Discrimination Causes Less Moral Outrage Than Human Discrimination[J].Journal of Experimental Psychology:General,2023,152,(1):4-27.

[51]李丹.算法歧视消费者:行为机制,损益界定与协同规制[J].上海财经大学学报,2021,(2):17-33.

[52]Messeri, L., and M.Crockett.Artificial Intelligence and Illusions of Understanding in Scientific Research[J].Nature,2024,627, (8002):49-58.

[53] Park, H.E.The Double-Edged Sword of Generative Artificial Intelligence in Digitalization: An Affordances and Constraints Perspective[J].Psychology & Marketing,2024,41,(11):2924-2941.

[54]Cinelli, M., G.De Francisci Morales, and A.Galeazzi, et al.The Echo Chamber Effect on Social Media[J].Proceedings of the National Academy of Sciences,2021,118,(9):1-18.

[55] Kitchens, B., S. L. Johnson, and P. Gray. Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption[J].MIS Quarterly,2020,44,(4):1619-1649.

[56]Bonnefon, J.F., A.Shariff, and I.Rahwan.The Social Dilemma of Autonomous Vehicles[J].Science,2016,352,(6293):1573-1576.

[57]Garimella, K., and S.Chauchard.Is AI Misinformation Influencing Elections in India?[J].Nature,2024,630,(8015):32-34.

[58]Bigman, Y.E., A.Waytz, and R.Alterovitz, et al.Holding Robots Responsible: The Elements of Machine Morality[J].Trends in Cognitive Sciences,2019,23,(5):365-368.

[59]Formosa,P.,and M.Ryan.Making Moral Machines:Why We Need Artificial Moral Agents[J].AI & Society,2021,36,(3):839-851.

[60]谢才凤,邬家骅,许丽颖,许丽颖,喻丰,张语嫣,谢莹莹.算法决策趋避的过程动机理论[J].北京:心理科学进展,2023,(1):60-77.

[61]Leong,A.M.W.,J.Y.Bai,and M.I.Rasheed,et al.AI Disruption Threat and Employee Outcomes:Role of Technology Insecurity,Thriving at Work,and Trait Self-Esteem[J].International Journal of Hospitality Management,2025,126,104064.

[62]Elkins,A.C.,N.E.Dunbar,and B.Adame,et al.Are Users Threatened by Credibility Assessment Systems?[J].Journal of Management Information Systems,2013,29,(4):249-262.

[63]Yogeeswaran,K.,J.Złotowski,and M.Livingstone,et al.The Interactive Effects of Robot Anthropomorphism and Robot Ability on Perceived Threat and Support for Robotics Research[J].Journal of Human-Robot Interaction,2016,5,(2):29-47.

[64]Mende,M.,M.L.Scott,and J.Van Doorn,et al.Service Robots Rising:How Humanoid Robots Influence Service Experiences and Elicit Compensatory Consumer Responses[J].Journal of Marketing Research,2019,56,(4):535-556.

[65]Yam,K.C.,P.M.Tang,and J.C.Jackson,et al.The Rise of Robots Increases Job Insecurity and Maladaptive Workplace Behaviors:Multimethod Evidence[J].Journal of Applied Psychology,2023,108,(5):850-870.

[66]Longoni,C.,A.Bonezzi,and C.K.Morewedge.Resistance to Medical Artificial Intelligence[J].Journal of Consumer Research,2019,46,(4):629-650.

[67]De Freitas,J.,and S.Puntoni.Chatbots and Mental Health:Insights into the Safety of Generative AI[J].Journal of Consumer Psychology,2023,34,(3):481-491.

[68]Kahn Jr,P.H.,H.E.Gary,and S.Shen.Children's Social Relationships with Current and near-Future Robots[J].Child Development Perspectives,2013,7,(1):32-37.

[69]Yun,J.H.,E.J.Lee,and D.H.Kim.Behavioral and Neural Evidence on Consumer Responses to Human Doctors and Medical Artificial Intelligence[J].Psychology & Marketing,2021,38,(4):610-625.

[70]Chen,F.Robots or Humans for Disaster Response? Impact on Consumer Prosociality and Possible Explanations[J].Journal of Consumer Psychology,2023,33,(2):432-440.

[71]Novak,T.P.A Generalized Framework for Moral Dilemmas Involving Autonomous Vehicles:A Commentary on Gill[J].Journal of Consumer Research,2020,47,(2):292-300.

[72]Kim,T.,H.Lee,and M.Y.Kim,et al.AI Increases Unethical Consumer Behavior Due to Reduced Anticipatory Guilt[J].Journal of the Academy of Marketing Science,2023,51,(4):785-801.

[73]Jost,J.T.,C.M.Federico,and J.L.Napier.Political Ideology:Its Structure,Functions,and Elective Affinities[J].Annual Review of Psychology,2009,60,(1):307-337.

[74]Awad,E.,S.Dsouza,and R.Kim,et al.The Moral Machine Experiment[J].Nature,2018,563,(7729):59-64.

[75]王恩界,乐国安.东西方文化背景下的道德观差异——来自于文化心理学视角的分析[J].天津:道德与文明,2006,(2):52-56.

[76]Fu,R.,Y.Huang,and P.V.Singh.Artificial Intelligence and Algorithmic Bias:Source,Detection,Mitigation,and Implications[M].Catonsville:Informs,2020.

[77]Swanepoel,D.The Possibility of Deliberate Norm-Adherence in AI[J].Ethics and Information Technology,2021,23,(2):157-163.

[78]Rahwan,I.Society-in-the-Loop:Programming the Algorithmic Social Contract[J].Ethics and Information Technology,2018,20,(1):5-14.

[79]Hasija,A.,and T.L.Esper.In Artificial Intelligence(AI) We Trust:A Qualitative Investigation of AI Technology Acceptance[J].Journal of Business Logistics,2022,43,(3):388-412.

# A Review of Ethical Issues and Governance Strategies in Artificial Intelligence

SUN Yu-jia, GUO Xue-li, SU Song, TANG Hong-hong

(School of Economics and Business Administration, Beijing Normal University, Beijing, 100875, China)

**Abstract**: The rapid integration of artificial intelligence (AI) into various facets of human life has heightened the urgency of addressing its ethical implications and developing effective governance strategies. As AI evolves from simple tools to autonomous agents embedded in social structures, understanding its societal impact is essential for promoting responsible development. This paper responds to this imperative by examining key ethical challenges in the digital era and proposing governance strategies that are both innovative and culturally relevant.

The accelerated development of AI—driven by digital transformation across industries—has significantly deepened its integration into daily life. While early AI systems focused on performing tasks that required human intelligence, contemporary AI functions extend to data acquisition and utilization, algorithmic classification and prediction, learningguidance, and social interaction. These applications influence not only human behavior and cognition but also the broader processes of social governance and policy-making. As a result, it is vital to embed ethical principles throughout the AI lifecycle to mitigate its potential risks to human development and societal stability.

This study offers a systematic review of AI ethics, defining it as the moral principles and values that govern both non-interactive and interactive human-AI processes. It identifies ethical issues that emerge when these principles are violated, leading to substantial consequences for individuals, organizations, governments, and society.

Although AI ethics has received growing academic attention, abibliometric analysis reveals that existing studies often focus on isolated issues or specific application contexts, lacking comprehensive theoretical integration. Inresponse, this paper systematically analyzes AI ethical concerns, comparing shared and distinct challenges across non-interactive and interactive contexts. It also delineates the mechanisms through which these issues arise and maps their distributional characteristics.

In non-interactive scenarios, AI systems raise concerns primarily related to large-scale data collection, privacybreaches, and diminished individual autonomy. Forinstance, Algorithmic misclassification may lead to unfair treatment, and biased predictions can reinforce social inequalities. Moreover, overreliance on algorithmic outputs risks narrowing cognitive diversity and amplifying societal polarization.

In interactive scenarios, AI systems introduce more complex ethical dilemmas, primarily related to responsibility allocation, threats to human agency, and the erosion of social and moral norms. Forinstance, AI-driven decision-making may adversely affect psychological well-being and physical health, while ambiguous responsibility in human-AI collaborations challenges traditional accountability frameworks. Furthermore, advanced AI capabilities may undermine human uniqueness, affectingidentity, agency, and interpersonal behavior. Ethical risks also emerge in social interactions, including reduced authentic human contact, imitation of harmful behaviors, and erosion of emotional and moral norms.

To address these multi-faceted challenges, this paper proposes a multi-dimensional and layered governance framework rooted in six-dimensional ethical theory. This framework emphasizes the need to clarify the moral basis for defining rights and responsibilities between humans and AI. Key proposals include defining AI's role as an agent, assigning accountability through robust evaluation mechanisms, and developing adaptive regulatory strategies. Additionally, the paper advocates for a social contract-based model of governance, which underscores human-AI synergy and iterative feedback loops to improve ethical decision-making.

In conclusion, this study highlights the critical importance of embedding ethical considerations into AI development and governance. By constructing a comprehensive AI ethics framework and proposing culturally adapted governance strategies, particularly suited to China's socio-cultural context, it offers both theoretical insights and practical solutions. These efforts aim to guide the responsible evolution of AI, promote social cohesion, and support the country's strategic goals in technological and social innovation.

**Key Words**: artificial intelligence; ethical issues; morality; governance strategy

**JEL Classification**: M00, M19

**DOI**: 10.19616/j.cnki.bmj.2025.05.002

(责任编辑:刘建丽)