第 39 卷 第 2 期 湖 南 大 学 学 报(社 会 科 学 版) Vol. 39,No. 2
2 0 2 5 年 3 月 Journal of Hunan University(Social Sciences) Mar. 2 0 2 5

DOI:10.16339/j.cnki.hdxbskb.2025.02.012

当代 AI 技术的伦理现象透视与实践智慧探索*

徐亚州

[中共山东省委党校(山东行政学院) 哲学教研部,山东 济南 250103]

[摘 要] 人工智能(AI)革新了"人类本质"与"机器属性"的传统认知,持续引发了"谁应负责""何为正确"以及"如何规范"的积极探索,深刻介入了道德主体、道德评价及道德规范的理论审视。AI 伦理不仅是关于伦理原则与道德规范的知识积累,更重要的是如何在复杂多变的道德情境中寻求最佳的行动方案。AI 时代的伦理问题揭示了规范伦理学在面对新兴技术挑战时的理论局限,凸显了"实践智慧"的伦理路向在当代技术伦理中的价值意义。儒家伦理学跳出了规范伦理学(义务论、后果论)非此即彼的线性思维,强调对具体情况的敏锐洞察、对行动后果的合理预见以及对人类福祉的深刻关乎。面对新技术、新问题、新挑战,儒家伦理学的实践智慧依然可以发挥持久性贡献。技术革新拓展了伦理思考的空间,而伦理文明则为技术发展提供了价值参照,两者在冲突与融合中推动了人类社会向更加文明、可持续的方向发展。

[关键词] 人工智能;人类本质;机器属性;伦理文明;实践智慧

[中图分类号] B82-057 [文献标识码] A [文章编号] 1008-1763(2025)02-0104-10

The Perspective on Ethical Phenomena of Contemporary AI Technology and Exploration of Practical Wisdom

XU Yazhou

[Philosophy Teaching and Research Department, CPC Shandong Provincial Party School(Shangdong Administrative College), Jinan 250103, China]

Abstract: Artificial Intelligence (AI) has revolutionized the traditional understanding of "human nature" and "machine attributes", continuously triggering an active exploration of "who should be responsible""what is correct" and "how to regulate", deeply intervening in the theoretical examination of moral subjects, moral evaluation, and moral norms. AI ethics are not only about accumulating knowledge of ethical principles and moral norms, but more importantly, how to seek the best course of action in complex and ever-changing moral contexts. The ethical issues in the AI era have revealed the theoretical limitations of normative ethics in the face of emerging technological challenges, highlighting the value and significance of the ethical approach of "practical wisdom" in contemporary technological ethics. Confucian ethics break away from the linear thinking of normative ethics (deontological and consequentialist), emphasizing keen insighty into specific situations, reasonable foresight of the consequences of action, and profound concern for human well-being. In the face of new technologies, new problems, and new challenges, the practical wisdom of Confucian ethics can still make lasting contributions. Technological innovation has expanded the space for ethical thinking, while ethical civilization has provided value references for technological development. The conflict and integration of the two have promoted the development of human society towards a

* [收稿日期] 2024-07-02
[基金项目] 山东省社会科学规划研究项目:跨文化视域下儒家伦理学"公私之辨"的时代价值研究(24CRWJ32)
[作者简介] 徐亚州(1990—),男,河南商丘人,中共山东省委党校(山东行政学院)哲学教研部讲师,山东省政德与廉洁文化研究中心研究员,博士,研究方向:应用伦理学。

more civilized and sustainable direction.

Key words: Artificial Intelligence; human nature; machine attributes; ethical civilization; practical wisdom

“人工智能”(Artificial Intelligence)简称“AI”，通常以计算机软件或包含软件的系统形式存在，能够执行通常需要人类智能才能完成的任务①。AI作为21世纪最具革命性的技术之一，正以前所未有的速度改变着整个人类世界。GPT 系列的发布不仅刷新了人们对 AI 创造力的认知边界，也激发了公众对 AI 未来可能性的无限遐想。近年来，AI 领域在大模型、深度学习等技术上取得了显著进展，它们在理解复杂指令、生成自然语言、创造艺术内容等方面展现出超乎预期的能力，极大地引发了人类社会对 AI 技术的广泛关注。AI 技术的主要目标在于理解和创建具有模仿人类思维模式、学习能力、问题解决技巧及自主决策功能的高级智能系统，使计算机系统能够掌握语言处理、图像识别、数据分析、逻辑推理乃至自我进步等多方面的复杂技能。当代 AI 技术以前所未有的速度发展，深刻地影响着现代社会的各个层面，从日常生活的便捷化到工业生产的智能化，乃至医疗、教育等关键领域的现代化。然而，这一技术的飞速进步也伴随着一系列复杂的伦理问题，这要求我们在追求技术革新的同时，深入思考社会实践活动中的道德责任，以确保技术发展和人类文明保持良性互动。在此进程中，我们必须面对的一个时代问题是：何为“善”的技术？这不仅仅是对技术本身性质的探讨，更是对人类价值体系和技术伦理框架构建的挑战。

一　AI 技术的伦理问题解析

人工智能作为当今世界科技发展的前沿阵地，其影响已远远超出技术实现的层面，广泛涉及人类社会伦理道德的核心内容②。人工智能的情境处理能力深度触及“道德技术”与“道德准则”的认知问题。有关这一问题的讨论可以追溯到技术哲学的基本命题之一：技术不仅是工具的集合，同时也是人类活动的表达形式，它反映并塑造着我们的社会伦理观。我们将 AI 视为一种“道德技术”，其实是在强调其作为实现特定道德目标的手段[1]。按照这种观点，AI 的设计、开发和使用应当遵循一套明确的道德原则，旨在促进积极的社会影响与行为结果。例如，AI 系统可以在医疗诊断中提供辅助决策，从而提高治疗效果；抑或在自动驾驶汽车中优化出行，减少事故。在这种情况下，AI 更多作为一种手段和工具，承担人类关于幸福、美好以及福祉的相应目标。然而，我们将 AI 看作“道德准则”的体现，则是在探讨 AI 本身所蕴含的价值判断及行为意义。这意味着 AI 不仅仅是被动地遵循外部设定的规则，而且具备了一定程度上的自我判断能力，能够在不同情境下做出符合某种伦理框架的决策。例如，在自动驾驶领域，面对不可预见的情况时，AI 系统需要能够在没有完全预设的情况下作出反应，这要求 AI 具有一定的“道德智能”，能够在复杂环境中进行价值权衡。然而，无论是作为“道德技术”还是“道德准则”，AI 的发展都面临着如何在技术和伦理之间建立合理桥梁的巨大挑战。一方面，技术的进步需要伦理指导，以确保其应用符合人类社会的基本价值观念；另一方面，伦理理论也需要随着技术的发展而不断更新，以应对新的技术环境所带来的伦理问题。它要求我们在讨论 AI 的伦理问题时，不仅要考虑具体的应用场景和技术细节，还需要反思这些规范

① 随着人工智能技术的不断进步和深入应用，如何衡量和评估这些技术的效能与影响变得尤为重要。2023 年出现的新基准，如 AgentBench (Agent－based behavior)、BigToM(Causal reasoning)、HaluEval(Factuality)、MMMU(General reasoning)以及 MoCa(Moral reasoning)，反映了这一领域评估标准的多元化趋势。详见 Perrault R，Clark J，“Artificial Intelligence Index Report 2024”，*Human-Centered Artifical Intelligence*，Stanford University，2024。这种趋势促使开发者和研究者不仅要追求技术上的卓越，还要确保 AI 系统的透明度、公平性和责任感，以实现技术革新与伦理文明之间的和谐共进。

② AI 伦理的关键在于探索工具理性(zweckrational)与价值理性(wertrational)之间的平衡点，它不仅是技术问题，更是深刻的哲学伦理学命题，涉及对人类自身、技术本质及其相互作用的深刻理解。在现代社会的发展进程中，工具理性与价值理性作为两种核心的思维方式，深刻影响着人类社会的基础结构、决策过程及发展方向。马克斯·韦伯最初在社会学理论中引入了工具理性和价值理性这对概念，用以深刻剖析现代社会的行动逻辑与社会结构。按照韦伯的论述，社会化可以典型地尤其建立在通过相互允诺的合乎理性(价值或目的合乎理性)的协议之上。详见韦伯.经济与社会：上卷[M].林荣远，译.北京：商务印书馆，1997：70。韦伯通过区分这两种理性，揭示了现代社会行动逻辑的复杂性。工具理性和价值理性之间的张力反映了现代社会的一个基本矛盾：如何在追求效率、经济增长的同时，保持对人类价值和社会正义的尊重。

背后更深层的哲学预设。从这一角度出发,AI的发展不仅是技术与伦理之间的互动过程,更是对人类自身价值体系的一种检验和重塑。

随着AI系统越来越智能化,它们在决策中的作用日益显著,从而引发了关于责任归属、算法偏见以及公平正义等方面的伦理问题。例如,如果AI系统做出错误决策导致了伤害,应该由谁来承担责任?如何确保AI系统的决策过程公正无私,特别是在教育、医疗、经济等民生领域?传统的责任框架可能难以适用于AI引发的问题,因为错误可能源自算法设计、数据输入、系统维护等多个环节,甚至可能是AI自主学习的意外结果。当AI系统做出错误或有争议的决策时,如何界定责任成为一个AI实践难题。由于AI能够自我适应和自主创造,其决策过程可能超出最初设计者的预期,导致责任可能涉及开发者、使用者、监管机构乃至社会。AI作为技术革命的同时,也是一场深刻的伦理革命。AI的类人智能化趋势,让我们不得不思考:技术进步将最终导向人的异化,还是开辟新的自我实现路径?这要求我们在探索技术潜力的同时,不断回溯并重构传统伦理框架,确保技术发展服务于人类的全面发展,维护人类社会的和谐与进步。这不仅是一个技术问题,更是一个关于人性、价值与未来的伦理议题,需要哲学伦理学的深入研究与实践探索。随着AI技术逐步渗透到社会生活各层面,人类的身份地位、劳动方式、情感交流等也都面临重塑,这不仅仅是生产、生活方式的变化,更是对"何以为人"的哲学追问。

AI革新了我们对"人类本质"与"机器属性"的传统认知,挑战了既有的道德原则和伦理界限,促使我们重新审视和构建伦理思考的框架。AI通过大数据学习与算法优化,能够处理和解析前所未有的技术问题,甚至在某些领域展现出超越人类的能力,这引发了对人类的独特属性与机器的工具属性的反思[1]。我们不得不面对的问题是:当机器开始"创造"和"决策",知识与智慧的边界在哪里?这一进程迫使我们深入审视人性的本质与价值,同时还要对机器智能的潜力与限制进行更为透彻的思考。一方面,AI引发了我们对"人类本质"的重新审视。AI在抽象思维、语言理解、艺术创作及社会互动等方面的能力,拓宽了我们对"创造"和"决策"这些传统上认为是人类独特属性的认知界限。它们打破了以往将某些智力活动视为人类专属领域的观念,激发我们深入探索那些算法难以复刻的能力领域,例如情感表达、价值判断以及深层次的人文精神。另一方面,AI重塑了我们对"机器属性"的认知理解。在传统认知中,机器往往被视为执行预设任务的工具,其功能和行为严格受限于人类编写的程序。然而,随着AI技术的发展,特别是机器学习和深度学习的兴起,机器开始展现出前所未有的灵活性、创造性和智能性。尽管目前的AI尚未达到真正意义上的自我意识水平,但它们在模拟情感反应、进行个性化交流等方面的能力,引发了哲学伦理学家们对"受限人工智能"向"自主人工智能"过渡可能性的深思。在智能家居、智能医疗、智能伙伴等场景中,AI正逐步融入人类生活的各个方面,成为人类的助手、同伴乃至决策顾问①。这一趋势不仅改变了我们的生活方式和工作模式,也促使我们反思人类与技术之间的互动关系及如何在依赖技术的同时保持人类的核心价值与尊严。

AI的发展深刻地改变了我们对"人类本质"与"机器属性"的传统认知,进而在很多方面对传统的伦理原则和道德规范提出了挑战。传统时代的伦理与AI时代的伦理之间存在显著的差异,伦理关切的对象从纯粹的人类社会扩展到人机混合的社会结构。传统时代的伦理通常以人类为中心,强调个体与社会的关系,以及个体在社会中的责任和义务。这种伦理体系往往建立在宗教、文化和历史传统的基础上,其核心是人与人之间的相互尊重、诚信和正义等。在此伦理框架下,道德判断基于人类的情感、经验和直观理解,强调个体的责任与社群的共同福祉。例如,儒家伦理强调的五伦(君臣、父子、夫妇、兄弟、朋友),基督教伦理强调的博爱与宽恕,等等。AI时代的伦理不再仅仅是人与人之间的关系,而且扩展到人与机器、机器与机器之间的关系。AI伦理的核心议题在于如何确保技术发展符合人类的普遍价值,以及如何在尊重人性价值、维护人类尊严的基础上引导技术向善。概而言之,AI时代的伦理变迁

① AI的决策自主性可以从不同层次理解,类似于自动驾驶汽车的SAE(汽车工程师学会)分级,从完全人工无计算机辅助(类似SAE 0级),到人机协作(SAE 1~2级,人类操作者在环或之上监督决策),再到几乎或完全无人干预(SAE 4~5级,人类出环)。这种分级体现了AI从辅助决策到主导决策的角色转变。随着AI在决策上的自主性不断增强,它们在遵循预设规则的同时,也展现出了自我优化的能力[2]。

不仅是概念范畴方面的拓展延伸，更是一次深刻的社会一技术互动模式的重构，直接挑战了“谁应负责”“何为正确”及“如何规范”等传统伦理问题的理解和实践。在 AI 时代，由于技术可能带来的长远影响和潜在风险，伦理考量需要更具前瞻性和实践性①。AI 时代的伦理革新不仅是对传统伦理原则的应用调整，更是对人类社会道德实践体系的根本性重构，亟须建立能够适应新技术环境的行为规范和价值理念，以确保技术进步与人类福祉的和谐共进。

二　AI 时代的伦理现象透视

如何在 AI 时代维持人类的尊严和价值，探索并构建一个人机共生的伦理框架体系，已成为当前伦理学领域亟待解决的重要议题。传统伦理规范主要关注人类行为及其后果，其伦理框架设计侧重于调节人与人之间的关系。随着 AI 技术的发展，智能实体开始广泛参与到社会互动之中，传统伦理规范已难以覆盖人机互动的伦理时代。AI 时代的伦理问题不仅涉及人与人之间的关系，还涵盖了人与机器及机器与机器之间的交互情境。传统时代伦理规范向 AI 时代伦理规范的过渡，不是简单地将原有准则框架应用于新兴技术领域，而是在深入把握人工智能的特性和影响的基础上，对人类道德认知与行为准则的一次全面审视与深刻变革。当前 AI 在处理特定任务时已展现出较高智能水平，人与机器之间的关系日益深化，从辅助工具转变为参与者甚至是决策者。当智能系统在无人干预的情况下做出决策，尤其是当这些决策导致不良后果时，该如何界定责任？是归咎于算法的设计者、使用者还是开发者？抑或是将 AI 本身视为承担责任的主体？这些问题的探讨，不仅关系到技术的深化发展，更触及人类文明的核心议题。AI 技术的飞速发展与广泛应用，不仅极大地推动了社会进步和产业变革，也对传统伦理规范产生了全方位、深层次的冲击。这一进程不仅触及谁或什么可被视为道德行为者的根本问题，还涉及审视和重估行为正当与否的评判标准问题，以及如何在人与机器共存的世界中合理地决策的问题。从道德主体的界定，到道德评价的构建，再到道德规范的实施，都体现出前所未有的复杂性和挑战性，亟须建立一套全面、系统的伦理框架体系，以确保技术的健康发展与社会的和谐稳定。

（一）道德主体

AI 时代的到来，不仅极大地改变了我们的生活，还深刻地触及哲学伦理学的多个核心议题，特别是引发了关于道德主体的广泛争论[2]。这一争论的核心在于：当 AI 系统展现出某种程度的学习能力、决策能力乃至创造能力时，我们是否应该以及如何将道德责任、道德权利乃至道德地位赋予这些非生物体的智能实体。当前，智能机器人、自动驾驶汽车、工业机器人等开始在生产、生活等领域扮演着越来越重要的角色，在某些情境下，其决策和行为已经接近或等同于人类。这自然而然地引发了人们对于“谁应当承担道德责任”“机器是否有道德地位”以及“如何界定人与 AI 之间的道德界限”等一系列复杂而紧迫问题的深思。一种观点认为，AI 系统作为设计和编程的产物，其行为后果应归咎于其创造者或使用者，即人类应当是最终的道德责任主体②。尽管 AI 能够模拟复杂的决策过程，但其缺乏自我意识、主观情感和真正的自由意志，因此不能被视为独立的道德主体[3]。另一种观点认为，随着 AI 技术的进步，特别是当某些 AI 展现出高度自主性和创造性时，简单地将它们排除在道德考量之外可能过于狭隘③。这些观点倾向于探索为 AI 设定某种形式的“机器道德”，或者至少承认它们在特定情境下的道德地位。在 AI 时代，道德主体的内涵与外延经历了显著的变化，这种变化深刻地影响着对道德责任、道德权利以及人机交互关系的理解。这要求我们重新审视和构建智能时代的伦理框架体系，探讨如何让 AI 在其行动范围内契合人类社会的伦理原则与道德规范，以及如何在 AI 失误或造成伤害

---

① 乔纳森等人认为，即使是最先进的算法也无法超越其学习数据中的偏见和错误，因为“算法没有接触真相的途径”。详见 Shaw J. Artificial intelligence and ethics[J]. Harvard Magazine，2019，30：1－11。

② 乔里斯·格拉夫等人反对将人工智能视为“道德主体”，认为“道德敏感性”是道德主体的前提条件，而“道德敏感性”是一种无法编码的实用技能[3]。目前，在构建人工道德主体的过程中主要存在自上而下和自下而上两种方法，然而，这两种方法都存在明显的理论局限性。详见岳瑨. 人工智能伦理：规制、理论与实践难题——学术史梳理及其问题域考察[J]. 东南大学学报（哲学社会科学版），2024，26(3)：32－41＋150。

③ 弗朗西斯等人认为人工智能是人类道德主体地位的分有，需要有效机制来纠正错误并防止未来的不当行为。详见 Grodzinsky F S，Miller K W，Wolf M J. The ethics of designing artificial agents[J]. Ethics and Information Technology，2008(2)：115－121。

时实施恰当的责任分配和行为评价。

(二)道德评价

AI在自主决策、自主创造、自主发展方面的不断突破,引发了关于道德评价问题的争论。例如,当说一个AI系统做出了"道德上正确"的行为决策时,我们是指其行为决策符合预设的编程规则,还是指它真正理解并体现了道德价值?一种观点认为,AI系统的道德正确性基于预设编程规则①。当评价一个AI系统做出了"道德上正确"的行为决策时,实际上是指该决策过程和结果符合程序员或设计者预设的编程规则和伦理准则[4]。在这种视角下,AI并不具备真正的理解能力或主观意识,它只是执行了经过精心设计的程序代码,这些代码体现了人类对道德情境的判断和处理方式。因此,AI的"道德正确"是一种人为构建的程序代码,确保其行动在给定框架内不出偏差,并非AI本身拥有道德意识或价值意义。另一种观点认为,AI系统能够真正体现和理解道德价值②。这种观点强调AI系统做出"道德上正确"的行为决策,不仅是遵循预设规则那么简单,还意味着AI通过学习、演化或复杂的自我调整机制,能够理解和体现道德价值。这种观点认为,随着人工智能技术的发展,尤其是机器学习和深度学习的进步,AI有能力从大量数据中学习道德原则、伦理案例,并基于此进行推理和决策[5]。这意味着AI不仅是在执行程序指令,而且在某种程度上"理解"道德情境,自主地评估不同行动方案的道德后果,从而进行具有道德意义的行为活动。这种观点通常与强人工智能或者自主人工智能的概念相联系,即AI不仅模仿人类智能,还能在一定程度上达到或接近人类的思维水平,包括对道德价值的理解和实践。AI的产生不仅革新了道德评价的传统认知,还深刻凸显了传统道德规范的理论局限。

(三)道德规范

当前AI系统的道德规范设计主要采用两种规范伦理学的理论,即后果论伦理学和义务论伦理学。AI系统的行为决策依据何种理论,其行为表现具有相当明显的差异性。后果论伦理学(如边沁、密尔的伦理理论)主张行为是否具有道德价值取决于其产生的后果,尤其是能否最大化总体的福祉或效用。按照这种理论,AI系统被编程为优化某些效益的各种指标,如提高生产效率、增进公共健康或提升用户满意度等③。AI通过数据分析预测不同行动方案的可能后果,选择预计能带来最大效益的方案[6],即便这可能违反某些特定规则或牺牲少数人的利益。与后果论不同,义务论伦理学(如康德、罗斯的伦理理论)强调行为的道德性基于其是否遵循普遍的道德法则或原则,而非仅仅看其后果[7]。根据这种理论,AI被编程为要遵循一套固定的道德准则或规则,如保护隐私、避免伤害、诚实守信等,即使在某些情况下这可能不会产生最大化的总体效益④。然而,在AI领域,后果论伦理学和义务论伦理学都存在明显的实践困境。后果论伦理学强调行为的后果,要求最大化总体的幸福或最小化痛苦,但预测所有可能的后果及其长远影响较为困难,对于高度复杂的行为系统和社会互动尤为如此。这种不确定性可能导致决策判断的失误,甚至引发不可预见的连锁反应。义务论伦理学较为强调遵循固定的规则或原则,但现实世界中不同的道德规则往往存在冲突,义务论难以提供明确的优先级或解决冲突的方法。AI技术的快速发展带来了前所未有的伦理困境,这些困境的复杂性和多样性超出了传统规范伦理学理论的预设范畴。AI时代的伦理问题不仅是对传统伦理学知识边界的拓宽,更是一次对伦理实践路向的全面革新,推动着伦理思考朝着更加灵活、多元且动态的方向发展。

三　AI伦理的实践智慧探索

AI时代的伦理问题揭示了规范伦理学在面对新兴技术挑战时的理论局限,凸显了"实践智慧"的伦理路向在当代技术伦理中的价值意义。"实践智慧"指面向实践活动的一种推理能力,是理智德性的一种表现形式。它涉及在具体情境中做出正确判断和决策的能力,特别是关于何为好的、正当的行为选择。在亚里士多德的《尼各马可伦理学》中,"实践智慧"意指一种关于善与恶、应当与不应当的判断能力,以及如何在现实情境中实现美德和幸福的能力。"实践智慧"的理念不仅是古典哲学的重要关切,也持续影响着现代伦理形态的发展建构。按照亚里士多德的说法:"智慧是科学和努斯的结合,并且与最高等的事物相关。"[8]192 这一概念强调了知识与行动之间的联系,意味着真正的智慧不仅仅是理论上的

① 埃里克等人认为,尽管机器学习在快速发展,但目前AI系统尚不能理解事物的深层含义。详见Brynjolfsson E, Mcafee A. Artificial intelligence, for real[J]. Harvard Business Review, 2017(1):1-31。
② 奥拉夫等人指出,随着人工生命系统复杂性的提升,它们可能达到与非人类脊椎动物乃至人类相似的道德考量水平[5]。
③ 汤姆森认为任何具有一定认知和行动能力的实体,都应遵循一个普遍的负面命令,即采取预防措施以避免负面后果[6]。
④ 乌尔根等人认为康德伦理学为我们评估人工智能与机器人道德行为提供了理论框架[7]。

理解，更是能够在具体情境中做出正确判断和行为的能力。AI 时代的伦理实践不仅是对既有道德规则的应用，更是一场关于如何在快速变化的技术环境中，以前瞻性的方法维护人类价值和尊严的探索。在 AI 时代，技术的迅猛发展带来了前所未有的伦理问题。自动化决策、大数据分析、人工智能辅助等技术，在提高效率和解决问题的同时，也引发了隐私侵犯、算法偏见、责任归属等一系列复杂的伦理问题。面对这些问题与挑战，不能简单地依赖于预设的规则或算法，而是需要不断适应变化的环境与需求，做出及时而精准的行为决策和应对策略。“实践智慧”成为连接技术革新与伦理考量的关键桥梁，对引导 AI 向着更加人性化、负责任的方向发展至关重要。

（一）道德代理

将 AI 视为道德代理而非直接的道德主体的思路，可以较好回避“是”与“否”两分观点的理论局限。将 AI 视为道德代理即在承认 AI 不具备完整道德主体性的前提下，要求其在执行任务时遵循一定的伦理准则，确保其行为结果与人类社会的价值观念协调一致。道德代理认可 AI 技术当前的能力表现，同时为技术的未来发展预留了空间。将 AI 作为道德代理的致思路向，是对传统道德分类的一次超越，打破了人类中心主义的传统观念。当我们将 AI 定义为道德代理，实际上是在创建一种新的伦理框架，在这个框架中，AI 系统的行为被视为一种延伸的人类活动，它们虽不具有内在的道德意识或感受能力，但却能够通过设计来体现并执行某些预定的道德规则。这种转变体现了从实体论到功能论的过渡，即从关注对象本身的内在性质转向考量对象在其环境中的作用与功能。在这个过程中，道德不再仅仅被视为内在的本质属性，而是成为一种由互动、情境以及技术设计所共同构建的活动。如前所述，对于是否应当将人工智能视为某种程度上的道德主体存在两种不同的观点。第一种观点认为，随着 AI 技术的发展，特别是当 AI 系统展现出高度自主能力、学习能力及决策能力时，它们应当被赋予一定的道德主体地位①。第二种观点主张，无论 AI 技术如何先进，由于缺乏自我意识、情感反应及真正的思维能力，它们从根本上无法具备道德主体资格②。实际上，无论是将 AI 视为道德主体还是将其完全排除在外，都面临着理论和实践上的困境。

第一种观点可能过分估计了 AI 的自我意识和情感能力。目前的 AI 系统，尽管在某些任务上表现出惊人的智能，但它们本质上仍然是基于算法和数据处理的机器。它们缺乏真正的情感体验和自我意识，无法具有痛苦、快乐等主观体验，也无法形成个体意图或自主目标，而这些都是构成道德主体性的重要元素。第二种观点将 AI 排除在道德考量之外，可能过于简化了 AI 与人类社会日益交互的伦理现实。即使 AI 当前不具备道德主体资格，但它们的行为和决策已经对当前社会产生了深远影响。因而，将 AI 视为道德代理为解决当前伦理困境提供了一条可行思路，既不低估 AI 技术的社会影响力，也不过度赋予其超出现实能力的道德地位。道德代理这一概念的引入，可以打破原有的二元分立，允许我们考虑一种新的存在状态——既非纯粹的道德主体，也非完全的非道德主体，而是介于两者之间的某种形态。这种介于传统道德主体与非道德主体之间的存在状态，我们可以称之为“准道德主体”。这一概念不仅有助于澄清 AI 技术在当代社会中的伦理定位，而且能够促进关于技术伦理责任的讨论，使我们能够更加细致地评估 AI 行为的道德意义。道德代理这一概念承认 AI 在执行任务时可能会产生的间接后果，并要求设计者、使用者以及监管者共同承担确保这些系统行为符合社会伦理标准的责任。将 AI 视作道德代理不仅是技术发展的应然要求，也是对传统伦理学概念的深化拓展。这种致思路向促使我们不仅从功能实现的角度来评估 AI 系统，还应从价值判断和社会影响的角度对其进行考量。这一视角跳出传统伦理学道德主体与非道德主体的两分界限，强调 AI 在道德实践领域中扮演的复杂角色，以及人类在设计、使用和监管 AI 时所承担的道德责任。

（二）过程评价

将 AI 视为道德代理的致思路向，为评价人工智能的行为活动提供了一个新视角，即将评价对象视为更大系统的一个组成部分。在这个框架下，对于 AI 系统的评价不再简单地基于其是否拥有自我意识或者情感，而是更多地关注其行动如何影响他

① 此种观点可以从义务论伦理学中获得理论支持。按照义务论伦理学的相关观点，如果 AI 能够基于某种形式的理性原则做出决策，那么就应当对其行为进行道德评估，并要求其承担相应的道德责任。

② 此种观点可以从约翰·洛克的人格同一性中获得理论支持。按照洛克的论述，道德责任和道德判断根植于人类的情感、意识体验及社会关系之中，这些都是当前 AI 所不具备的。洛克反对传统的基于灵魂不变性的解释，提出以意识为核心标准来界定人格同一性。在洛克那里，当我们谈论一个人在不同时间点上保持同一性时，实际上是指这个人的意识能够从现在回溯并认同过去的思想、情感和经历。正如洛克所强调的那样：“在我看来，所谓人格就是有思想、有智慧的一种东西，它有理性、能反省，并且能在异时异地认自己是自己，是同一的能思维的东西。”详见洛克．人类理解论[M]．关文运，译．北京：商务印书馆，2017：334。

人和社会,并且强调人类对于这些系统的控制和引导的重要性。因此,讨论的焦点从AI本身是否具备主观体验转移至人类社会应该如何合理设置规则来引导AI的发展方向,以及如何构建一个能够确保技术进步与伦理价值相协调的观念体系。这种评价方式超越了仅仅关注“行为”本身的局限性[9],更加注重理解AI系统行为活动的整个过程,包括其内部机制、数据使用、学习过程以及与人类用户或其他系统之间的互动。对AI产生的行为进行过程性评价,不仅是技术层面的算法性能评测,更是一个涉及设计者、使用者以及监管者的多维度评估。如前所述,目前关于AI行为的道德评价问题,主要存在两种不同的逻辑思路。一种思路认为,如果AI的行为能够影响人类的生产、生活和福祉,那么这些行为就具有道德意义①。另一种思路认为,由于AI无法真正理解道德行为的深层含义和内在动机,其“道德”行为仅仅是对外部规则的程序模拟②。然而,简单地认定AI行为有无道德意义,不仅简化了AI技术的复杂性,也忽略了人机交互情境的多样性。对于AI行为活动的道德评价,应当超越简单的“是”与“否”的二分思维,转向更为动态和综合的过程性评价框架。

过程性评价强调AI系统的整个行为过程,包括设计、运行和维护等各个环节。例如,在设计阶段,设计者需要全面考虑数据偏见问题,避免AI系统在学习过程中放大或产生歧视;在运行阶段,运行者需要持续监控AI模型的实时表现和用户反馈,确保实际输出与预期目标的一致性;在维护阶段,维护者需要全面掌握AI系统的性能与影响,确保其行为决策的公正性和合理性。过程性评价强调在每个阶段都要进行道德评估,确保AI系统的设计初衷、功能实现以及最终应用都符合人类社会的道德标准和价值观念。这种持续性的伦理考量,既考虑了AI系统本身的技术特点及其在现实世界中的作用,也强化了人类在AI开发和使用过程中的相关道德责任,促使技术的发展与社会进步相协调,避免潜在的技术滥用或伦理失范现象的发生。过程性评价强调的是一个系统性伦理建构问题[11]180——在AI技术发展的各个阶段中,如何嵌入并维持一套连贯的伦理框架。这不仅仅是对AI系统本身的道德约束,更是对人类作为创造者和使用者的角色定位。从这个意义上说,过程性评价是对“何为善的技术”这一哲学命题的时代解答。AI的发展不应被视为孤立的技术进步,而应被视为一种社会进程,其中包含广泛的社会协商和价值选择。这种理解促使我们将注意力从单纯的科技进步转向更为综合的社会技术系统分析,从而确保AI的发展不仅服务于经济效率的提升,且能促进社会整体的公正与和谐。对AI行为进行过程性评价而非单一行为评价的方式,有助于建立一个更加全面、负责任的人工智能生态系统,同时也为未来可能出现的新技术提供了灵活且具有前瞻性的伦理指导。

(三)实践智慧

通过采用“实践智慧”的设计理念开发AI伦理决策系统,能够更好地提升人工智能的道德判断力和情境敏感性。这一设计理念不以单一的程序规则为基础,而是侧重于AI在特定情境中灵活决策的能力。“实践智慧”强调在不同情境中找到最合适的行动路径,这意味着没有可以适用于所有情况的一成不变的规则。目前,AI伦理决策系统的设计主要以义务论伦理学和后果论伦理学为基础,然而,这两种思路都存在一定的理论局限③。义务论伦理学主要强调行为规则的重要性,关注“应该做什么”而不是行为的结果。在AI领域,这体现为编程一系列道德规则或原则(例如,诚实原则,尊重原则,公平原则,等等),以确保AI系统的行为符合预设的道德规则[12]。然而,在某些情况下,不同的道德规则可能发生冲突,而义务论没有提供清晰的解决方案以恰当处理这些冲突。相对来说,后果论伦理学主要关注行为的结果,主张最大化幸福或利益,最小化痛苦或不幸。但是,片面强调行为的后果往往会削弱行为本身的道德属性,比如为了追求最大利益而牺牲少数人的权益。将“实践智慧”的理念引入AI伦理决策系统的设计中,可以弥补上述两种伦理理论的不足。“实践智慧”强调一种基于情境的道德推理能力,它不仅仅考虑行为的规则和结果,还注重对具体情境的理解及在复杂多变的情境下做出合理判断。这样的系统设计能够使AI在面临道德困境时,不是简单地遵循预设的规则或最大化某一结果,而是能够在多种可能的行动方案中寻找平衡点。

自亚里士多德提出“实践智慧”的概念以来,这一理念便被视为人类道德行为活动中不可或缺的一部分,成为连接理论知识与实际行动之间的重要桥

① 从后果论伦理学的角度看,无论行为的源头是有意识的还是无意识的,其产生的后果都有道德价值[9]。

② 道德行为的评价不仅是逻辑规则问题,还涉及意识、情感、直觉等特有的心理状态。按照赫斯特豪斯的说法即:“在特定的情境中体验特定的情感具有内在的道德价值。”[10]121

③ 米歇尔等人强调,确保自主系统遵循可接受的道德原则对于它们与人类交互的许可程度至关重要,这是推动自主系统发展的关键所在[12]。

梁。将“实践智慧”融入人工智能的设计,不仅有助于克服传统伦理理论在实际应用中的局限性,还为创造一种更贴近人类道德直觉的技术提供了可能性。例如,当自动驾驶汽车遇到经典的电车难题,即必须在伤害较少人与较多人之间做出抉择时,一个仅依据义务论或后果论构建的决策系统可能会因为缺乏明确指导原则而难以做出决定。而一个具备“实践智慧”的系统,则会考虑更多因素,如环境的具体细节、潜在受害者的状况等,进而尝试找到一个更符合道德直觉的解决方案。这不仅仅是为了避免简单的二元选择,更重要的是为了确保决策过程能够反映出对生命价值的尊重,并且尽可能地减少对无辜者的伤害。这样的设计思路要求我们在构建 AI 时,超越传统的线性编程规范,进一步融入能够灵活应对模糊性和不确定性的智能模型。例如,强化学习模型(reinforcement learning, RL)允许 AI 系统与环境互动来学习最佳行为模式,它通过反馈机制引导 AI 做出最优决策,从而不断优化其表现;案例推理模型(case-based reasoning, CBR)则是一种让 AI 从过往经验中学习的方法,通过检索并调整先前解决问题的方案来应对新情况。“实践智慧”强调在不确定性和变动性面前保持敏锐的判断力和行动力,这对引导 AI 技术朝着有益于人类社会的方向发展具有重要意义。将“实践智慧”的理念融入 AI 技术的研究与实践中,可以促进技术理性与人文关怀的深度融合,进而推动形成更加公正、包容和可持续发展的智能社会。

四　智能时代与中国智慧

义务论伦理学和后果论伦理学作为规范伦理学中的两大主流理论框架,贯穿了整个现代道德哲学的发展历程。这两个理论框架自系统化以来,就持续地影响着现代道德哲学的讨论。义务论伦理学的核心观点是行为的道德价值不取决于其结果,而在于行为本身是否符合某种道德规则或原则。后果论伦理学也被称为效益主义或结果主义伦理学,基本主张是行为的道德价值主要由其产生的后果来决定。然而,无论是义务论伦理学还是后果论伦理学,都存在过于聚焦“行为”本身的问题,相对忽视了行为活动的多元性及具体情境的复杂性。两者都试图为道德判断提供普遍适用的原则标准,但在道德实践中往往会遇到情境多样性与普遍原则性之间的张力问题。这种张力问题在智能时代更为显著,例如,在自动驾驶汽车的决策中,如何把握人身安全与交通规则之间的关系,以及如何在紧急情况下做出两难的抉择,都是典型的 AI 道德困境。安斯库姆已经对义务论和后果论进行了系统性的反思和批驳[13],他的批评引发了广泛的讨论,并激励了后来被称为“美德伦理学复兴”的运动。美德伦理学试图恢复古希腊哲学中对“美德”(areté)的重视,认为道德行为的本质在于行为者的内在美德。尽管美德伦理学提供了对良好生活的深刻理解,并强调了“美德”在道德生活中的中心地位,但也存在诸多理论上的局限性。例如,在道德实践中识别什么是真正的美德,什么可能是伪善或伪装的美德,是一个复杂的伦理问题。智能时代的到来进一步凸显了美德伦理学的理论局限,这要求我们不断反思和重新定义“美德”的概念,使之能够适应科技发展带来的新情境。

近年来,自主 AI 技术的迅速进步,促使我们更加清晰地认识到传统规范伦理学框架在应对新兴技术伦理挑战时的理论边界与不足之处①。受限 AI 通常是指那些在特定任务或领域内运行,并且其决策过程受到严格限制或指导的人工智能系统。而自主 AI 则指那些拥有更广泛行动自由度的系统,它们能够在复杂多变的环境中做出决策,并且具有一定的自我学习能力[14]。受限 AI 的行为更多地依赖于设计者的意图及所设定的规则框架。这种类型的人工智能被视为工具性的存在,其伦理判断主要集中在设计者是否正确地设定了边界条件,以确保 AI 在其操作范围内不会产生有害行为。因此,对于受限 AI 而言,伦理讨论的重点在于如何构建一个既能实现目标又不会偏离道德规范的系统。相比之下,自主 AI 由于具备更高的灵敏性和智能性,能够根据环境的变化调整自己的行为。自主 AI 在行为上展现出某种程度的“意志性”,尽管这种意志并不等同于人类的自由意志,但它确实让 AI 系统能够在面对新情况时展现出类人的反应。这种能力使自主 AI 在伦理上面临着更加复杂的问题。自主 AI 不仅能够遵循预设的道德准则,还能够在遇到新情况时进行伦理推理。这就需要一套更加灵活且普适

① 应该说明的是,目前的技术水平还无法实现真正意义上的自主 AI,即能够完全独立地进行道德判断并承担责任的 AI。虽然深度学习等技术使得机器能够在特定领域内表现出与人类相当甚至超越人类的能力,但是这些能力被限定在特定任务上,并且缺乏人类所拥有的广泛常识和社会文化背景知识[维特根斯坦所谓的“生活形式”(Lebensform)]。例如,当我们学习一种语言时,不仅仅是在学习词汇和语法,还在学习这种语言承载的文化、习俗和社会规范。

的伦理理论框架,以适应更加富厚且多元的伦理生活世界。随着自主AI技术的发展,我们必须不断创新和发展伦理理论,使其不仅能适用于当今的技术水平,还能应对未来可能出现的新问题,确保社会整体的和谐稳定与人类文明的长远发展。

自主AI的发展不仅挑战了现有伦理理论的应用范围,而且促使我们必须超越现有的思维框架,探索新的理论路向以应对由智能技术带来的伦理挑战。相对来说,中国传统伦理学特别是儒家伦理学跳出了规范伦理学非此即彼的线性思维,为当前新兴技术的伦理挑战提供了更为灵活和综合的视角。尽管儒家伦理学与美德伦理学有着共同的伦理关切,但两者在基本立场、价值取向及实践路径上展现出显著的不同。在以亚里士多德为代表的古典美德伦理学中,德性往往被视为达到“幸福”(eudaimonia)的手段,而德行则是德性在具体情境中的体现。如亚里士多德所述:“如果幸福在于合德性的活动,我们就可以说它合于最好的德性,即我们的最好部分的德性。”[8]334 虽然美德伦理学也强调美德与行为的关联,但更侧重于德性本身作为行为导向的价值。相对而言,儒家伦理学则在德性与德行之间建立了更为直接和具体的联系,德行被视为德性展现的自然结果。正如《礼记·中庸》所言:“性之德也,合外内之道也,故时措之宜也。”[15]1029 传统儒家历来强调内在德性培养与外在行为规范的统一,德性与德行的合一可谓儒家伦理学的基本立场。在儒家伦理学中,德性主要是指个体在道德层面展现出的稳定性和一贯性的内在特质与行为倾向。它更多地指向一个人内心深处的善恶观念、道德意识以及性格中的正面特质,如诚实、善良、勇敢、谦逊等。德行主要侧重于个体在外在行为上的表现,即一个人实际展现出的道德行为和具体操守。例如,一个人如果具有诚实的德性,那么他在面对选择时,会倾向于说真话,这种选择和行为就构成了他的德行。在儒家看来,德性与德行不是割裂的两个概念,而是相辅相成、不可分割的统一体。

按照儒家的逻辑,德性是德行的动力机制,没有内在的德性支撑,外在的德行将是空洞的形式架构。同时,外在的德行实践又是检验和提升内在德性的必要途径。德性与德行的合一,不仅是理论上的探讨,更是实践中的追求,体现了传统儒家对情境多样性与普遍原则性的深刻把握。正如陈来所述:“实践智慧一直是中国哲学的主体和核心。儒家自孔子以来,更是强调哲学作为实践智慧的意义。”[16] 不同于西方哲学中对于道德理论的精密推理,儒家更为关注的是如何在日常生活中恰当地践行道德规范①。儒家伦理并未停留在道德的纯粹思辨推理层面,而是将道德理念落实到实际生活之中,强调通过具体行动来体现道德价值。按照王阳明的说法:“致知者,意诚之本也。然亦不是悬空的致知,致知在实事上格。”[17]136 良知指导下的行为应当是真诚而具体的,而不是停留在程序化的道德规则层面。这种情境化的实践智慧,体现了儒家对现实生活的深切关注,以及对主体在社会中所扮演角色的深刻理解。儒家伦理强调的不仅仅是对道德原则的理解(“是什么”),更在于如何将这些原则应用到日常生活之中(“怎么做”)[18]14 。在儒家看来,真正的“美德”不仅在于知道什么是好的,更在于如何去实现这种好,以及在实际行动中如何面对各种复杂的情况做出正确的选择②。通过将“实践智慧”的理念融入日常生活中的每一个细节,儒家伦理提供了一个实现个人与社会和谐发展的理论框架。这种理论路向对中华文化乃至东亚文化有着深远影响,并为现代社会推进技术发展与伦理文明之间的关系开辟了独特视角。

概而言之,AI伦理的本质在于,探索在技术快速演进时如何维护和促进人类的核心价值观念,解决由技术进步引发的伦理道德冲突,确保技术发展与人类福祉和谐共生。这要求我们不断审视技术与社会的交互作用及人类自身在这一进程中的角色与责任。中国传统伦理学特别是儒家伦理学为AI伦理提供了丰富的思想资源,这不仅有助于构建更加人性化、负责任的AI决策系统,也促使我们在技术快速发展的当下,不断反思和探索技术革新与伦理文明之间的良性互动。正如习近平总书记所强调的那样:“要深度参与全球科技治理,贡献中国智慧,塑造科技向善的文化理念,让科技更好增进人类福祉,让中国科技为推动构建人类命运共同体作出更大贡献!”[19]201-202 儒家伦理学在AI伦理领域的应用与探索,不仅是对“中国智慧”的现代传承,也是向世界提供“中国方案”、展示“中国力量”的重要途径。儒

① 在梁漱溟那里,儒家孔门之学为体认人的生命生活之学,关键在反躬修己的实践,不宜以哲学思想目之。详见梁漱溟.梁漱溟全集:第7卷[M].济南:山东人民出版社,1993:498。

② Ryle 提出了“knowing-that”(命题知识)和“knowing-how”(能力知识)的概念,他认为,“知道如何”(knowing-how)的知识不能简单地还原为一系列“知道那是”(knowing-that)的陈述,因为前者涉及主体与世界互动的实际能力[18]14-15 。

家伦理学不仅是关于伦理知识或伦理理论的探讨，而且关切如何将这些认知应用于具体情境中的实践能力[10]257。它涉及对具体情况的敏锐洞察、对行动后果的合理预见及对人类福祉的深刻关心。面对新技术、新问题、新挑战，儒家伦理学的实践智慧依然可以发挥持久性贡献。儒家伦理学的创造性转化与创新性发展，不仅体现为理论上的深化与拓展，更在于其与社会实践的紧密互动。这一过程涉及对儒家传统的深刻理解、批判反思及与当代社会现实的积极对话。技术革新拓展了伦理思考的空间，而伦理文明则为技术发展提供了价值参照，两者在冲突与融合中共同推动人类社会向更加文明、可持续的方向发展。正所谓："富有之谓大业，日新之谓盛德。"[20]304

[参　考　文　献]

[1]　Nieswandt K. Instrumental rationality in the social sciences[J]. Philosophy of the Social Sciences，2024(1)：46－68.

[2]　Chesterman S. Artificial intelligence and the problem of autonomy[J]. Notre Dame Journal on Emerging Technologies，2020(1)：210－250.

[3]　Graff J. Moral sensitivity and the limits of artificial moral agents[J]. Ethics and Information Technology，2024(1)：1－12.

[4]　Jeste D V，Graham S A，Nguyen T T，et al. Beyond artificial intelligence：exploring artificial wisdom[J]. International Psychogeriatrics，2020(8)：993－1001.

[5]　Witkowski O，Schwitzgebel E. The ethics of life as it could be：do we have moral obligations to artificial life?[J]. Artificial Life，2024(2)：193－215.

[6]　Thomsen K. Ethics for artificial intelligence，ethics for all[J]. Paladyn，Journal of Behavioral Robotics，2019(1)：359－363.

[7]　Ulgen O. Kantian ethics in the age of artificial intelligence and robotics[J]. QIL，2017，43：59－83.

[8]　亚里士多德. 尼各马可伦理学[M]. 廖申白，译注. 北京：商务印书馆，2017.

[9]　Rudolph J. Consequences and limits：a critique of consequentialism[J]. Macalester Journal of Philosophy，2011，17：64－76.

[10]　赫斯特豪斯. 美德伦理学[M]. 李义天，译. 南京：译林出版社，2016.

[11]　斯洛特. 源自动机的道德[M]. 韩辰锴，译. 南京：译林出版社，2020.

[12]　Anderson M，Anderson S L. Toward ensuring ethical behavior from autonomous systems：a case-supported principle-based paradigm[J]. Industrial Robot，2015(4)：19－28.

[13]　Anscombe G E M. Modern moral philosophy[J]. Philosophy，1958，33：1－19.

[14]　Radanliev P，De Roure D. Review of the state of the art in autonomous artificial intelligence[J]. AI and Ethics，2023(2)：497－504.

[15]　礼记[M]. 胡平生，张萌，译注. 北京：中华书局，2017.

[16]　陈来. 论儒家的实践智慧[J]. 哲学研究，2014(8)：36－41＋128.

[17]　王守仁. 王阳明全集[M]. 吴光，钱明，董平，等编校. 上海：上海古籍出版社，2011.

[18]　Ryle G. The concept of mind[M]. London：Routledge，2009.

[19]　习近平. 习近平谈治国理政：第四卷[M]. 北京：外文出版社，2022.

[20]　周易[M]. 来知德，集注. 胡真，校点. 上海：上海古籍出版社，2013.