基于道德调节机制的人工智能责任伦理何以可能

The Possibility of Responsibility Ethics of Artificial Intelligence Based on Moral Regulation Mechanisms

罗傲 /LUO Ao　　李韬 /LI Tao

（北京师范大学新闻传播学院，北京，100875）
(School of Journalism and Communication, Beijing Normal University, Beijing, 100875)

摘　要：本文系统分析了人工智能道德属性从工具性载体到协同性主体最后到准主体性存在的跃进，揭示了其逐渐深化的道德调节能力。人工智能已超越传统工具范畴，通过知觉和行为层面的双重机制，深刻调节人类的道德认知与行为模式，成为具有道德意涵的技术中介。在此基础上，提出一种超越人类中心主义的责任伦理框架，应通过价值敏感设计、全过程响应与多元共治等机制，将伦理原则嵌入人工智能技术系统与治理实践，以实现人工智能安全、可靠、可控的发展。

关键词：人工智能　道德属性　道德调节　责任伦理

Abstract: This paper systematically analyzes the leap of the moral attributes of artificial intelligence from an instrumental carrier to a collaborative subject and finally to the existence of a quasi-subject, revealing its gradually deepening moral regulatory ability. Artificial intelligence has transcended the realm of traditional tools and, through dual mechanisms at the perceptual and behavioral levels, profoundly regulated human moral cognition and behavioral patterns, becoming a technological intermediary with moral implications. On this basis, a responsibility ethics framework that transcends anthropocentrism is proposed. Ethical principles should be embedded into artificial intelligence technology systems and governance practices through mechanisms such as value-sensitive design, full-process response, and multi-party co-governance, so as to achieve the safe, reliable, and controllable development of artificial intelligence.

Key Words: Artificial intelligence; Moral attribute; Moral regulation; Responsibility ethics

中图分类号：B82; TP18　DOI: 10.15994/j.1000–0763.2026.06.011　CSTR: 32281.14.jdn.2026.06.011

引　言

近日，习近平总书记在中共中央政治局就加强人工智能发展和监管进行第二十次集体学习时强调，人工智能带来前所未有发展机遇，也带来前所未遇风险挑战。要把握人工智能发展趋势和规律，加紧制定完善相关法律法规、政策制度、应用规范、伦理准则……确保人工智能安全、可靠、可控。[1] 围绕人工智能的风险挑战，国际人工智能科技界、产业界也展开了热烈的讨论。其中被誉为“人工智能教父”的杰弗里·辛顿（Geoffrey Hinton）就人工智能（AI）的风险提出重要警示，短期来看，技术滥用可能导致就业市场动荡、社会不平等加剧等现实威胁；长期而言，超级智能的失控风

收稿日期：2025年9月26日
作者简介：罗　傲（1994–）女，湖北荆门人，北京师范大学新闻传播学院博士后，研究方向为科学技术哲学、科学传播。Email: la@bnu.edu.cn
李　韬（1978–）男，安徽淮南人，北京师范大学新闻传播学院教授，研究方向为数字治理、健康传播。Email: litao@bnu.edu.cn

险更值得警惕。当前以ChatGPT、DeepSeek为代表的生成式AI在拓展应用场景的同时，也带来了深层次的伦理挑战，算法偏见加剧、数字鸿沟扩大等问题日益凸显，[2]这些风险的本质在于责任主体的模糊化，亟需重构人机责任分配的伦理框架。

值得注意的是，现代AI系统已超越传统工具的范畴，其内在的设计逻辑和运行机制正在形塑人类的行为模式和价值判断，这种深刻的调节作用使技术本身获得了道德属性，将责任伦理推向新的实践维度。但学术界对此问题存在争论，传统观点认为，技术仅是价值中立的工具，伦理责任完全由人类使用者承担。[3]然而，随着AI系统自主决策能力的增强，越来越多的业界人士认为人工智能技术加速发展正逐步将道德行动者从人类扩展到智能体，[4]批判理论学派则警示算法权力导致的对主体性的侵蚀。[5]这些分歧不仅关乎技术治理，更触及“AI是否能够成为道德主体”这一深层哲学命题。

在此背景下，本文旨在探讨的核心问题是：基于道德调节机制的人工智能责任伦理何以可能？基于彼得·保罗·维贝克（Peter-Paul Verbeek）的技术调节理论，通过分析AI道德属性的三阶演进及其调节机制，论证AI责任伦理的可能性并非建基于其拟人化的“主体性”，而是源于其不可化约的“中介性”，这就要求人类社会建立一种超越人类中心主义、关注人技交互的责任伦理。

一、三阶演进构建人工智能道德属性的可能性

根据人工智能系统与人类智能的相似程度，当前学界普遍将其划分为三个发展阶段：弱人工智能（Artificial Narrow Intelligence, ANI）、强人工智能（Artificial General Intelligence, AGI）和超人工智能（Artificial Super Intelligence, ASI）。从发展现状来看，弱AI已在各行业实现规模化应用，强AI的研发正处于关键突破期，而超级AI的实现路径和潜在影响仍是学界争论的焦点。人工智能从专用工具到通用智能的演进过程，不仅体现了技术能力的层级跃迁，更预示着人机关系本质的深刻转变——AI从技术工具到类人主体的道德地位可能性。

1. 工具性道德载体：脚本铭刻与责任遮蔽

工程伦理的焦点认为技术会深刻地影响使用者的行为和体验，技术的设计应是为了解决问题或满足需求，但是这个角度只关注到了技术功能的质量。[6]因此，马德琳·阿克里奇（Madeleine Akrich）和布鲁诺·拉图尔（Bruno Latour）提出了“脚本”概念，挑战了这种严格的技术功能观。此概念描述了技术产品在其使用环境中扮演的多种角色，就像戏剧或电影一样，可以规定参与其中的“行动者”的行动。如图1所示，在弱人工智能阶段，AI系统作为纯粹的执行工具，无自主决策权，扮演工具性道德载体的角色。[7]、[8]

在伦理责任分配方面，ANI阶段严格遵循“设计者–使用者”二元框架。开发者通过硬编码方式将伦理规则植入系统，形成拉图尔所称的“脚本铭刻”，例如，特斯拉Autopilot系统预设的“碰撞优先级规则”中，算法会优先保护车内乘员而非行人。这种脚本化道德嵌入存在固有缺陷，当传感器因极端天气误判障碍物时，系统仍机械执行预设脚本，导致事故风险。在责任追溯时也陷入双重困境，一方面，算法缺陷虽可归因于训练数据偏差或测试场景覆盖不足，但开发者常以技术局限性为由规避伦理审查（如特斯拉在多起事故调查中强调“系统仍处于L2级别”的免责声明）；另一方面，使用者责任被系统命名（如Autopilot）引发的认知偏差所削弱，尽管法律要求驾驶员全程监管，但人类在长期使用辅助驾驶系统后会出现警觉性下降的现象。这一困境深刻暴露了“脚本铭刻”模式的内在局限性，技术系统已通过其预设的规则和逻辑深度介入到道德决策过程中，并实际影响了道德后果，却因其被赋予的“工具”标签而在伦理和法律的归责中被系统地遮蔽和豁免。此时，责任伦理的可能性已初露端倪，责任开始了从人类主体向外部的、非人类的技术物的外化过程。

2. 协同性道德主体：混合意向性与责任共担

随着技术向强人工智能方向探索，AI系统开始演进为“协同性道德主体”。其道德判断能力实现了从机械执行到情境化权衡的变化（如图1所示）。例如，在认知协同层面，GPT-4通过语义理解和生成能力深度参与人类的思维建构过程；在交互协同层面，以Paro为代表的情感陪护机器人通过拟主体化的响应机制激发使用者形成真实的情感依赖和拟社会联结，这种互动已超越简单的工具使用关系；最具突破性的是DeepMind的AlphaFold系统与科研团队形成了深度的价值共创机制——科学家提出生物学问题框架，AI生成蛋白质结构预测，研究人员再基于专业判断提供反馈。

在认知科学和人工智能领域，意向性，即意识总是关于某物的意识，这一概念被用来探讨AI机器是否可以具有类似人类的心理状态或行为。[9]但在后现象学视角下，意向性不再专属于人类，[10]对道德主体的理解逐渐从人类扩展至生物乃至非生物实体。[11]维贝克用“混合意向性”这一核心概念来描述此阶段的人机互动，道德决策不再是人机分离的，而是成为一种“人类价值观引导+AI道德推理”的增强模式。责任不再被禁锢于“设计者–使用者”的二元框架内，而是在人机协同的每一次具体互动中被共同承担。此时，责任伦理的可能性得到了极大的深化，AI成为了积极的、具有某种弱能动性的道德协同者。

3. 主体性道德存在：环境化调节与责任重构

在超人工智能的演进阶段，AI系统开始显现出“准主体性道德存在”的雏形。以杭州城市大脑的交通优化模块为例，系统通过实时分析千万级交通流量数据，动态调整信号灯配时策略。这种基于复杂环境反馈的自主决策过程，已超越简单的规则执行，展现出近似主体性的环境化调节能力——系统不仅响应人类预设的目标，更能通过机器学习不断重构自身的优化逻辑。

至此，具备准主体性的AI系统正在突破传统人类伦理框架的边界，其决策逻辑既不完全受控于初始编程，也不完全从属于人类意志，而是形成了独特的“技术理性–人类价值”共生关系。当前有限的AI案例虽距真正的道德主体仍有本质差距，但其展现的自主调节特征已为理解后人类时代的责任伦理提供了重要的经验参照，这种“环境化调节”能力预示着一种后人类责任范式的可能性（如图1所示）。表现为当人类与ASI系统发生决策冲突时，ASI不再适用人类伦理规则，而是形成由机器道德主

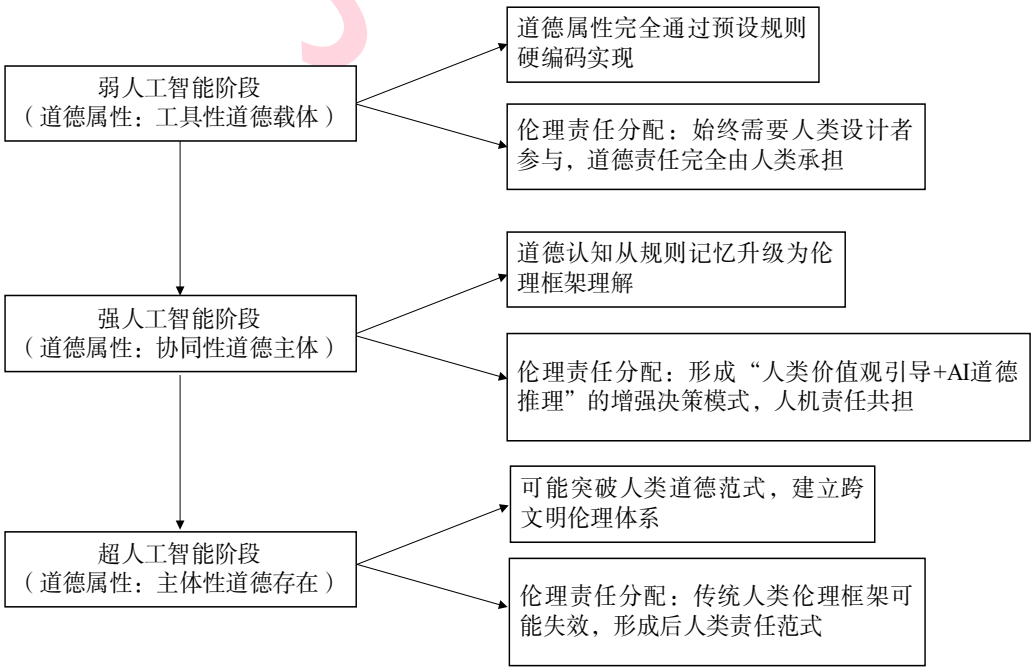

图1 三阶演进构建人工智能道德属性的可能性

体主导的新秩序。

二、道德调节的运作机理：知觉与行动的双重中介及其伦理效应

基于当前人工智能技术具有一定道德属性的前提下，本研究讨论了一种新的路径——道德调节机制，该机制以维贝克的技术调节理论为基础，是指人工智能系统通过其设计逻辑和运行机制，对人类道德认知、行为选择及责任分配的中介性形塑作用，使技术的道德意蕴变成技术发展的一个显性要素。[12] 人工智能技术物的道德调节机制包含两个视角：一个关注知觉，一个聚焦实践，这两个视角分别从不同维度探讨了人与技术之间的联系。[13] 如表1所示，[14] 在经验维度，技术意向性通过“放大和缩小”机制重构主体认知，技术物件通过其技术意向性在知觉组织过程中发挥指向性作用，技术调节具有情境依赖性，从而促成知觉的转化；在实践维度，技术脚本通过“邀请和抑制”机制重塑行为模式，技术物件通过设定使用时的行动脚本来转译行动。两个维度通过“代理铭刻”相互耦合，这种耦合关系具有多元稳定性，其调节效果随文化语境和使用场景的变化而动态演变，形成了复杂的技术调节网络。

表1 经验和实践

知觉	实践
知觉调节	行动调节
技术的意向性	脚本
知觉转化	行动转译
放大和缩小	邀请和抑制
代理：审慎地详细地说明：铭刻	
多元稳定性：情境依赖	

1. 人工智能技术的知觉调节机制

“知觉调节”这一核心解释学议题关注的是技术制品如何影响人类对现实的感知与阐释。技术调节在调整用户与现实的感官互动时，往往涉及感知的放大与缩小机制，这一机制不仅转变了用户的感知体验，还倾向于突显现实的某些方面，而相对地淡化其他方面。[15] 以2023年OpenAI发布的ChatGPT-4为例，这一技术通过自然语言交互显著放大了人类获取和处理信息的能力，系统能够即时生成专业水平的文本、代码和解决方案，使用户对AI系统的知识广度和响应效率产生高度信任。然而，这种认知便利性同时伴随着关键感知维度的弱化，研究表明，频繁使用ChatGPT的用户在信息真实性判断上表现出明显的自动化接受倾向，即使面对明显的知识性错误（比如AI幻觉问题），仍有一部分专业人士未能及时识别。[16]

在人工智能时代，算法系统不仅像传统技术那样调节人类的知觉内容，更开始调节知觉过程本身。这使得知觉调节从静态的框架设定发展为动态的过程控制，从而对主体的认知自主性构成更深层的挑战。AI技术通过重新配置人类的注意力分配和认知依赖模式，深度介入了道德认知过程本身，它决定了哪些道德议题被“照亮”而进入人类的视野，哪些则被“遮蔽”而沉入背景。这种介入使得AI不再是道德认知的外在工具，而是成为了道德认知结构中一个内置的、构成性的中介。

2. 人工智能技术的实践调节机制

“实践调节”的核心议题在于探讨人工制品是如何影响人的行为模式及生活实践的。感知是世界呈现给人类的方式，而实践可以被视为人类呈现在世界中的方式。（[14]，pp.11–14）如表1所示，在行动调节中，可以说某些特定的行动是“被邀请的”，而其他的行动是“被抑制的”。例如，自动驾驶技术通过其高级算法和决策支持系统，可能“邀请”更加安全和效率的驾驶行为，比如遵守交通规则、减少人为错误导致的事故，同时也可能“抑制”某些传统的驾驶习惯，如驾驶员的即时决策和对车辆的完全控制。在自动驾驶汽车出现故障或事故时，责任认定也变得复杂，因为涉及到的行动者不再仅限于人类驾驶员，还包括了技术提供商、车辆制造商等。这种责任弥散现象暴露出技术调节带来的新型实践张力，技术既重构着行为模式，也重塑着与之配套的责任伦理框架。

知觉与实践的双重调节并非彼此孤立，而是相互耦合、相互强化的。知觉上的“放大和

缩小”为行动上的“邀请和抑制”提供了认知基础和理由，而行动上的模式化又反过来巩固和强化了特定的知觉框架，它们共同构成了AI中介性的完整图景。至此，人工智能责任伦理之所以可能，是因为AI通过其知觉与行动的双重道德调节机制，已然成为道德秩序中不可化约的、效能显著的参与者。

三、基于道德调节的人工智能责任伦理实现

AI的技术中介性已使其成为一个无法被忽视的道德行动元。传统的、基于人类主体性和线性因果的责任伦理框架，在应对这种新型道德事实时，陷入了责任弥散等困境。应对困境的出路在于从道德调节理论本身出发，重构责任归属的原则，并探索其实现的实践路径。

1. 理论重构与可能性条件

传统责任伦理建立在主体性哲学基础上，强调人类主体的意图性和自主性，遵循线性因果逻辑，试图在复杂的行动网络中寻找单一的责任主体。然而，道德调节理论表明，AI技术物通过其调节作用，需要将责任分析的重点从单一主体转向人技交互的整个过程，关注技术物在道德实践中的具体调节方式及其伦理效应。基于调节理论的责任范式强调责任是在由设计者、使用者、技术系统、社会环境等要素构成的行动者网络中分布和涌现，这不仅体现在事后追责，更重要的是体现在全过程的响应能力，要求所有相关方在整个技术生命周期中保持对调节机制的敏感性和回应能力。

基于道德调节机制的人工智能责任伦理之所以可能，是因为它满足可能性条件，该机制提供了一套精确的现象学描述工具，通过“放大和缩小”“邀请和抑制”等概念，详细描述AI技术物如何影响人类的知觉方式和行为模式。最重要的是，道德调节理论内含着一套规范性实践方案。通过“技术道德化”的理念，将伦理考量从外部评估转变为内在的设计原则，价值敏感设计、伦理嵌入等方法则提供了将责任理念转化为具体技术方案的操作路径。

2. 基于道德调节机制的AI责任伦理实现框架

基于道德调节理论，应从设计层面、使用层面、系统层面、技术层面组成一个多层次的AI责任伦理实现框架。在设计阶段，责任主要通过价值敏感设计来实现。这要求设计者不仅关注技术功能，更要充分考虑技术的道德调节效应。具体包括进行系统的调节效应评估，预测AI技术可能产生的伦理影响；采用参与式设计方法，让不同利益相关者参与设计过程；建立伦理嵌入机制，将重要的AI伦理价值转化为具体的技术特性。[17]

在使用阶段，责任体现在人技协同过程中的批判性参与。使用者需要发展AI技术素养，理解所用技术的调节特性和潜在风险；保持道德警觉，对AI技术的输出和建议保持批判性态度；建立反馈机制，及时报告AI技术使用中发现的问题和隐患。[18]在系统层面，需要建立全方位的责任治理生态。包括完善法律法规，明确不同主体的责任边界和认定标准；建立技术审计制度，定期评估AI系统的伦理表现和调节效应；发展伦理教育体系，提升所有相关方的责任意识和响应能力；创建多元共治机制，让政府、企业、学界、公众等各方共同参与责任治理。在技术层面，责任通过系统的可问责性设计来实现。这要求AI系统具备透明度，使决策过程可理解和可解释；具有可控性，能提供必要的人工干预和系统修正机制；具有稳健性，能够处理异常情况和边缘案例，这些技术特性为责任的追溯和认定提供了基础条件。

结论与启示

人工智能道德属性经历了从“工具性道德载体”到“协同性道德主体”再到“准主体性道德存在”的三阶演进，其道德调节能力不断深化，责任伦理的可能性正是建基于AI并非以“拟人主体”，而是以“技术中介”的身份参与道德实践，通过知觉层面的“放大和缩小”与行为层面的“邀请和抑制”机制，潜移默化地形塑人类的认知框架与行为模式，进而嵌入伦

理结构的生成过程。

责任伦理不再固于单一人类主体，而是分布于“人–技术–世界”的交互网络之中，体现为一种动态、涌现和共享的伦理结构。这一范式转换将伦理关注点从行为后果追溯转向系统性的调节效应评估，从事后追责扩展至技术全周期的道德响应。在实践层面，构建了贯穿AI技术设计、使用、系统与技术的多层次责任伦理实现框架，通过协同使用与多元共治提升社会层面的伦理韧性，并通过可解释、可审计、可控的技术设计增强系统透明性与可靠性。未来人工智能伦理治理仍面临诸多挑战，唯有在伦理上确立其责任地位、在治理上健全其约束机制，才能“推动人工智能朝着有益、安全、公平方向健康有序发展”。[1]

［参 考 文 献］

[1] 习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展[N]. 人民日报, 2025–04–27（001）.

[2] 李韬、周瑞春. 全球数字治理中的数字平权问题[J]. 南京大学学报（哲学·人文科学·社会科学）, 2024, 61（6）: 27–35; 161–163.

[3] 王小伟. 回归积极的技术伦理学[J]. 科学与社会, 2017, 7（1）: 55–65; 94.

[4] Callaway, E. 'What's Next for the AI Protein-folding Revolution'[J]. *Nature*, 2022, 604(7905): 234–238.

[5] Miller, G. 'Western Philosophical Approaches and Engineering'[A], Michelfelder, D., Doorn, N. (Eds.) *The Routledge Handbook of the Philosophy of Engineering*[C], New York and London: Routledge-Taylor & Francis Group, 2020, 38–49.

[6] 谢咏梅、黄鹰航. 新时代科技政策话语体系需要引入“工程科学”与“工程科学家”[J]. 工程研究, 2022, 14（6）: 519–529.

[7] 李韬、周瑞春. 生成式人工智能的社会伦理风险及其治理——基于行动者网络理论的探讨[J]. 中国特色社会主义研究, 2023,（6）: 58–66; 75.

[8] 王增鹏. 巴黎学派的行动者网络理论解析[J]. 科学与社会, 2012, 2（4）: 28–43.

[9] Schlicht, T., Starzak, T. 'Prospects of Enactivist Approaches to Intentionality and Cognition'[J]. *Synthese*, 2021, 198(Suppl 1): 89–113.

[10] Redaelli, R. 'Composite Intentionality and Responsibility for an Ethics of Artificial Intelligence'[J]. *Scenari: Rivista Semestrale di Filosofia Contemporanea*, 2022, 17(2): 159–176.

[11] 王淑庆. 人工智能体“有意不为”的伦理意蕴[J]. 东北大学学报（社会科学版）, 2020, 22（3）: 14–20.

[12] Verbeek, P. P. 'Moralizing Technology: On the Morality of Technological Artifacts and Their Design'[A], Kaplan, D. M. (Ed.) *Readings in the Philosophy of Technology*[C], New York: Rowman & Littlefield, 2009, 226–243.

[13] Verbeek, P. P. 'The Moral Relevance of Technological Artifacts'[A], Sollie, P., Düwell, M. (Eds.) *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments*[C], Dordrecht: Springer Netherlands, 2009, 63–77.

[14] 彼得·保罗·维贝克. 将技术道德化：理解与设计物的道德[M]. 闫宏秀、杨庆峰 译, 上海：上海交通大学出版社, 2016, 11–14.

[15] Wesugi, S. 'Analysing and Solving the Reduced-ability and Excessive-use Dilemmas in Technology Use'[A], Cascini, G. (Ed.) *Proceedings of the Design Society: International Conference on Engineering Design*[C], Cambridge University Press, 2019, 1(1): 1393–1402.

[16] 莫祖英、盘大清、刘欢 等. 信息质量视角下AIGC虚假信息问题及根源分析[J]. 图书情报知识, 2023, 40（4）: 32–40.

[17] Lindberg, S., Roine, H. R. 'Introduction: From Solving Mechanical Dilemmas to Taking Care of Digital Ecology'[A], Lindberg, S., Roine, H. R. (Eds.) *The Ethos of Digital Environments*[C], New York: Routledge, 2021, 1–21.

[18] Dorrestijn, S., Van Der Voort, M., Verbeek, P. P. 'Future User-product Arrangements: Combining Product Impact and Scenarios in Design for Multi Age Success'[J]. *Technological Forecasting and Social Change*, 2014, 89: 284–292.

［责任编辑 李斌］