人工智能悖论视域下人的主体性反思及伦理重塑

Reflection on Human Subjectivity and Ethical Reconfiguration Under the Horizon of the Paradoxes of Artificial Intelligence

梁军 /LIANG Jun　　张雯钰 /ZHANG Wenyu

（西安交通大学马克思主义学院，陕西西安，710049）
(School of Marxism, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049)

摘　要：人工智能（AI）的迅猛发展引发了关于人的主体性的广泛讨论，形成了这一议题下的哈姆雷特之问。如果说“to be”对应着“AI能做什么”，那么相关讨论中一定程度欠缺“not to be”即“AI不能做什么”的反思，而关于AI悖论的讨论则提供了这样的视域。当AI深度介入人类认知与实践，AI悖论就不只是技术问题，它更使人的主体性确证遭遇复杂的伦理关系变动。而人的主体性并非算法可以复制的功能，它植根于身体经验、主观能动性和真实生活。据此，AI伦理不仅要给算法设限，更要以守护人的主体性为核心，通过认知自觉、治理革新和价值坚守，在技术与人的张力中重新确立人的主体地位。

关键词：人工智能悖论　主体性　以人为本　伦理重塑

Abstract: The rapid development of artificial intelligence (AI) has triggered widespread debates over human subjectivity, forming a Hamletian dilemma within this field. Existing discussions predominantly focus on what AI can achieve while neglecting critical reflections on its inherent limitations. The exploration of AI paradoxes addresses this gap in perspective. When AI deeply intervenes in human cognition and practice, the AI paradox transcends mere technical concerns, further complicating the affirmation of human subjectivity through intricate ethical shifts. Rooted in embodied experience, subjective initiative and real-life practice, human subjectivity cannot be replicated by algorithmic operations. Therefore, AI ethics should prioritize the safeguarding of human subjectivity beyond mere algorithmic regulation. It is essential to reshape the dominant status of human beings amid human-technology tensions through cognitive self-consciousness, innovative governance and persistent value adherence.

Key Words: Artificial Intelligence paradoxes; Subjectivity; Human-centric; Ethical reconfiguration

中图分类号：B82; TP18　DOI: 10.15994/j.1000–0763.2026.12.012　CSTR: 32281.14.jdn.2026.12.012

引　言

人工智能不仅仅是一种工具，更是一种正在重新定义“智能”“行动”与“主体”概念的技术现象。大语言模型能够生成哲学论文、编写法律文书，甚至开展类似心理咨询的对话，它们表现得“像”有智能，甚至“像”有主体性。然而，在这种“像”的背后，隐藏着一系列深刻的悖论。譬如，知识图谱悖论表明，结构化

基金项目：教育部高校思想政治理论课教师研究专项重大课题攻关项目“高校思政课教师人工智能素养提升研究”（项目编号：25JDSZK07）。
收稿日期：2026年2月25日
作者简介：梁　军（1971–）男，陕西凤翔人，西安交通大学马克思主义学院教授，研究方向为科学技术哲学、马克思主义科技观。Email: liangjunfx@xjtu.edu.cn
张雯钰（1996–）女，陕西大荔人，西安交通大学马克思主义学院博士研究生，研究方向为马克思主义科技观。Email: zhangwenyu@stu.xjtu.edu.cn

知识体系越完备，系统内嵌的偏见和认知锁定风险越可能被遮蔽。可解释性悖论揭示，AI系统越智能、自主性越强，其决策过程就越复杂、越像“黑箱”。认知外包悖论警示，当人将深度思考委托给AI，表面上扩展了我们的认知能力，实际却导致了我们思考能力的退化。莫拉维克悖论则不得不让人反思，对人类而言困难的高阶推理对AI很简单，而人类无意识的感知和行动对AI却异常困难。[1]它们引发我们追问：当机器能够模拟乃至超越人类在诸多领域的认知和能力时，我们又当如何理解并守护人之为人的根本？

回答这些问题，或许不应仅仅着眼于AI已经能够胜任什么，更值得追问的是，AI在哪些根本性的事情上无法与人相提并论？从“不能”出发审视AI悖论，不仅更符合辩证理性的要求，而且能让我们穿透技术表象，直面人机智能博弈下人类的主体性问题，这也是当下AI伦理研究必须回应的核心议题。

一、人工智能悖论的多重形态

在形式逻辑的严格意义上，“悖论”指向形式体系中不可自洽的矛盾判断，移用至社会问题的分析，则指代那些违背直觉、无法简单消解的现实矛盾。人工智能所催生的悖论便属于此类。人工智能以符号、数据与统计运算把握现实世界，然而，正是这种形式化的计算体系与不可完全形式化的现实世界之间的先天断裂，构成了人工智能悖论的生成根源，由此在人与技术交互的不同层面显现出多重矛盾形态。

1. 客体表征层面：人工智能的知识图谱悖论

知识图谱是符号主义人工智能的重要技术范式，是现代知识表示的主流实现形态。它以三元组的形式（实体–关系–实体）摹画客观世界的实体、概念及关联关系，将海量碎片化的网络信息转化为结构化、标准化的内容，以“符号”的形式呈现出来。凭借结构化、标准化的优势，知识图谱得以作为人工智能解读自然语言的重要知识库，为机器理解语义、开展逻辑推理提供支撑，并广泛运用于智能问答、情报分析等场景。[2]

但是，知识图谱的结构化优势同时也形成其桎梏。从知识本身的属性来看，人类知识具有流动性、语境依附性与默会性，存在大量隐性内容。[3]而知识图谱依赖固定的符号系统来表征知识，在自然语言转换为形式框架的过程中，必然会舍弃上下文语义等信息，造成知识的简化与损耗。从工程实践来看，知识图谱还面临着人工建构与原始数据带来的双重问题。在数据遴选、整理标注、关系建构等人工环节中，参与者的价值偏好与认知局限会随之注入系统。而多源异构数据自带的信息噪声、事实冲突与时效滞后问题，也使得知识图谱在事实维度上难以实现精确完备。然而，知识图谱高度规范化的外在形态，营造出精准可靠的技术假象，遮蔽了其内在的局限与偏差。图谱蕴含的各类偏差还会伴随模型迭代与场景落地向下游智能系统传导，这一过程隐蔽性极强，难以被使用者察觉甄别。

由此，知识图谱形成了这样的悖论：它以完备的结构化形态赋能智能服务、提升认知效率，却以隐蔽的不完备内核固化机器偏见、限制人类的思辨和判断，最终呈现为技术赋能与认知规训的内在矛盾。这也正是形式化框架与不可完全形式化的人类知识之间内在矛盾的集中显现。

2. 算法运行层面：人工智能的可解释性悖论

符号主义时代的人工智能依靠人工预设规则，决策过程清晰可溯，而深度神经网络的兴起则改变了这一局面。它依托海量数据自动提取特征、拟合规律，决策过程成为难以还原的算法黑箱。由此带来的风险在隐蔽中积累，进一步损害个体权益、社会公平和公共信任，也使得可解释性成为算法治理绕不开的议题。

将可解释性确立为AI治理的目标，其内在的悖论也随之显现。表层矛盾体现为，人工智能模型在算法原理层面有着完备清晰的界定，可对于其实际生成的各类运行结果，我们始终难以充分解释。[4]而试图破解黑箱的努力，还会面临更进一步的困境。在技术层面，用算法

解释算法、用模型说明模型的做法，不过是新一次的拟合，难以摆脱机器逻辑固有的局限。现有的解释工具又多偏重于技术实现，忽略社会情境中不同主体的理解需求，反而可能加剧难以理解的问题。[5]落到使用者层面还会衍生两类认知困境：一是“透明度悖论”。详尽的决策解释容易造成信息过载，使用户误以为自己完全掌握决策逻辑，实则受认知负荷制约并未真正理解。有研究指出，只要有获得解释的预期，就可能催生使用者已经理解的幻觉。[6]二是“流畅性错觉”。大模型生成的流畅解释未必属实，却能够诱导用户过度信任，使人忽视模型潜藏的错误。[7]

算法原理的明晰不必然导向输出结果可解释，即便借助工具给出了解释，机器逻辑的封闭性叠加人的认知局限，仍可能让所谓的解释沦为认知层面的自我安慰。简言之，AI的生成、决策能力越强大，就越难以解释，而人类越是渴望信任机器，越是感到不安，由此便产生了人工智能的可解释性悖论，折射出算法统计拟合逻辑与人类意义理解之间难以弥合的错位。

3. 主体心智层面：人工智能的认知外包悖论

在人类文明演化的历程中，从结绳记事、竹简刻字到活字印刷，从百科全书到搜索引擎，工具使我们得以将记忆负担外化，从而将有限的认知资源转向更复杂的思考，在这个意义上，认知外包是人类认知进化的一条基本路径。

随着人工智能的出现，人类的认知外包发生了全新的变化。以往所有的外包对象都是静态的、被动的工具，而今，AI能够在数秒内完成人类需要数小时甚至数天才能完成的获取、筛选、分析、整合全流程工作，面对自身有限的认知负荷和无限膨胀的信息环境，将认知外包给强大的AI代理便成为人们突破自身认知瓶颈、应对信息爆炸的必然选择。人们设想，人脑专注于价值判断、创造性思维等高阶任务，智能设备承担运算密集型任务，二者协同推动个体封闭认知转向人机协同的分布式认知，促成人类认知方式的革命性跃升。[8]然而，随着人类自身思考时间和深度的降低，人们极易逐步滑向对AI的依附。更进一步，使用者还会倾向于将外部知识源的能力误认为是自己的内在知识，满足于一种虚假的知识感，逐渐丧失自身进一步探索、反思和巩固的内在动机。[9]而随着AI从“认知副驾驶”向主动代理的演进，遵循用进废退的神经可塑性原理，人类核心认知能力，尤其是执行功能，在个体层面可能会出现结构性萎缩。[10]

认知外包悖论便体现于此，用以拓展人类认知边界的技术手段，同样存在削弱人类主体认知能力的风险。这一矛盾，正是形式化智能工具深度介入个体内在认知实践带来的现实张力。

4. 智能本体层面：人工智能的莫拉维克悖论

“让计算机在智力测试或国际象棋博弈中展现出成人水平的性能相对容易；而要赋予它们在感知与行动方面达到一岁幼儿水准的能力，则显得困难甚至不可能。”[1]这就是由卡耐基梅隆大学机器人系教授汉斯·莫拉维克（Hans Moravec）最早提出的莫拉维克悖论。该理念也得到罗德尼·布鲁克斯（Rodney Brooks）、马文·闵斯基（Marvin Minsky）等学者的共同阐释。[11]

在传统认知中，逻辑推理、抽象思维、数学演算被视作人类高级智能的体现，而感知环境、协调运动、识别物体则被认为是低级智能。按照这种划分，如果智能的高低与实现的难易成正比，那么AI应当首先掌握低级智能，再逐步攻克高级智能。事实却是，人类深感棘手的高阶任务，机器却能轻松胜任；人类轻而易举的本能，机器反而举步维艰。例如，2016年，AlphaGo以4:1击败围棋世界冠军李世石。[12]2026年，OpenClaw已能够像私人助理般操作用户电脑完成邮件的撰写与发送。[13]尽管研究者们数十年来不断致力于AI机器人的感知和行动研究，但时至今日，这类机器人仍旧依赖传感器输入数据，无法拥有与人类相媲美的情境感知能力。四十余年间，算力增长数亿倍，从“专家系统”到“AI智能体”，人工智能在一些推理与抽象任务上已然超越人类，然而莫拉维克悖论所揭示的智能倒置依然未被归谬，形式计算仍然难以复刻植根于肉身处境的人类

原生智能。

综合来看，人工智能的四重悖论形成了现实的传导链条。知识表征环节的缺陷会带入智能系统加剧算法黑箱，黑箱化AI的高性能表现又进一步诱惑人类将思考交付，这些矛盾递进展开，最终在本体维度暴露出人机智能的内在分野，也将问题引向了技术场域下人的主体境遇。

二、人工智能悖论视域下人的主体性反思

自柏拉图在《斐德罗篇》中借苏格拉底之口，警示文字的发明将在学习者的灵魂中造成遗忘以来，人类每一次将自身能力外化至工具的革命性突破，几乎都会引发有关人的主体性增减的辩论。[14] 回望文明发展历程，从早期手工工具制造，到历次科技与工业革命对人类肢体、感官等生理能力的延伸与加强，过往的技术进步始终保持辅助工具的定位。如今，人工智能打破了既有格局，大规模渗入人类专属的精神活动领域，由此显现的人工智能悖论，勾勒出AI介入人类生产生活的矛盾图景，促使我们从认知、实践、价值三重维度重新理解人的主体性。

1. 人工智能悖论视域下关于认知维度的反思

莫拉维克悖论直观呈现了人机智能的不对称性，但它之所以出人意料，是因为我们会将人类自身的能力标准投射到机器之上。这也从侧面反映出人类认知的固有倾向就是以自我经验为尺度把握外部事物。但这种倾向容易带来主观偏差，若要使认识活动保持客观审慎，则离不开主体认知能力的完备，以及主体对自身认知局限的反思。然而，以知识图谱与大模型为代表的智能系统，正在侵蚀主体的认知能力和内在独立性。

一方面，智能知识供给机制造成主体认知能力完备性的损耗。在传统认知建构中，信息筛选、语境剥离、概念抽象以及知识关联的建立，是主体打磨认知图式的必要劳作。然而，知识图谱在前端便代劳了这一过程，智能系统进一步输出现成结论，接管原本属于人的逻辑推演、因果辨析与方案评估等高阶心智活动。主体完整的认知训练难以展开，长期停留在浅层信息调取的认知惰性状态，导致了自身独立认知机能完备性的萎缩。

另一方面，过度的认知外包容易消解主体反思自身认知局限的内在自觉。人机交互过程中，主体对智能系统的依附是降低认知负荷、追求认知效率的主动选择。智能系统拟人化的输出进一步降低了主体的戒备心理，削弱主体开展反思与诘问的内在动力。这种主观心理倾向，又与技术系统固有的客观局限交织在一起。例如，知识图谱内嵌的偏差往往被其形式上的完备所遮蔽，而下游智能系统输出的合理化解释进一步掩盖了自身概率黑箱的本质，导致使用者容易将算法输出默认为客观可靠的知识，难以分辨其中潜藏的问题。加之技术生成内容与主体原生思考的边界日益模糊，主体失去了跳出技术座架校准自身真实认知水平的参照系。个人的认知评价标准也在交互过程中发生偏移，对自身认知水平的评判不再聚焦内在思考过程，转而看重外部系统输出文本的流畅度，甚至将信息获取的便捷性等同于知识内化的确定性。在主观心理偏向与技术遮蔽效应的共同作用下，主体反思自身认知局限的内在自觉持续弱化。

总体来看，人类认知自主性的消解，是技术特性、主体认知本能与自主选择相互交织的结果。这一过程既损耗主体认知能力完备性的存量基础，又压缩主体自我反思迭代的增量空间，经由自强化的负向闭环形成隐蔽的认知衰退，最终从认知根基层面动摇人的主体性地位。

2. 人工智能悖论视域下关于实践维度的反思

实践作为主体能动地改造客观世界的感性物质活动，始终是确证人类主体地位的根本场域。若说认知的退化动摇了主体把握世界的观念图式，那么实践的异化则直接侵蚀了主体作用于世界的实体纽带。莫拉维克悖论所揭示的人机能力间的倒置关系，本可以成为阐发人类具身实践、现实经验不可替代的切入点。但在效用优先的现实实践中，人的丰富具身经验被

拆解为适配机器运转的重复性动作。人不再是改造客观世界的自觉主体，转而成为弥补算法物理局限的代偿工具，充当智能网络末端完成搬运、校准、物理运维的生物执行单元，具身实践的主体创造性在技术架构下不断走向工具化。

当不可解释的黑箱算法深度介入医疗处置、司法量刑、资源调配等关键场域，人类依托实践智慧开展裁量的空间也将持续受到挤压。有理由推测，初期的智能系统仅作为辅助工具输出备选方案，人类依托经验与伦理对其建议进行研判并做出最终决断，此时人类仍是选择与责任的实质主体。

而随着系统置信度的提升与业务节奏的加快，对算法输出开展实质核验的实践成本显著抬升，人类逐渐倾向直接采纳算法结论，实践由主动决策退化为被动审核，而这种审核在制度惯性与效率考核下极易流于形式。到了后期，算法的概率推演结果往往在实践中被固化为具有优先效力的决断依据，进一步压缩人在具体情境下自由裁量、提出异议的空间，实践不再充分实现主体的意志外化，而演变为高度依附算法预设逻辑的闭环流程。

在传统实践范式下，权责匹配建立在人的实质性参与、把控与创造之上，主体掌握实践全过程的决策裁量，同时承担实践产出的全部后果。但在算法嵌入的数字化场景中，“人在回路”的制度设定要求人类保持最终判断与兜底责任，法律、伦理与行业规范也将实践后果归属于人类主体。但现实中，算法已经前置完成了方案建构、逻辑推演、趋势预判等核心工作，人的能动权限遭到稀释，仅剩余流程末端的确认环节，责任标准却依旧沿用传统框架，由此形成数字化实践下的权责错位。

此外，算法的中立表象也为实践者提供了规避复杂情境抉择与伦理焦虑的路径。在有些案例中，已经出现了部分主体盲目遵从算法流程以换取程序层面的免责，主动放弃立足现实情境审慎决断、批判性介入的义务。我们必须高度警惕，在客观权限收缩与主观选择抽离的双重作用下，实践活动呈现出自由意志与责任承担相互割裂的风险，人类将从能动干预现实的实践主体，退化为维系技术系统继续运转的合规环节。

3.人工智能悖论视域下关于价值维度的反思

算法以数字化规约把鲜活立体的人拆解为有限的固化数据特征，人的意图被预测为概率分布，行为被还原为动作轨迹，情感被量化为偏好数据。人的不确定性与超越性被数据模型过滤，人的多元阐释空间被挤压。虽然人的本真存在还没有被彻底改写，但是这种技术趋向，正在加速动摇人的理性自决与自主价值判断赖以成立的存在论根基。

康德意义上的尊严，正是奠基于人的理性自主与道德立法能力。[15]人之所以区别于一般物象，具有绝对价值，就在于人能超越感性的欲求，按照普遍法则自我规定。然而，当算法代行价值排序，人不再经历道德权衡的内在博弈、价值取舍的精神拷问与自我立法的淬炼过程，人的尊严所依赖的理性自主便从真实的主体能力逐渐沦为形式标签。一个人的价值不再首先因其是理性存在者而不可替代，而取决于其在系统中的功能表现、数据贡献与产出效率。从“人是目的”到“人是可被计算和优化的环节”，这是根本性的价值颠倒。智能系统以“最优决策”“个性化服务”之名，使人的尊严从内在绝对价值变为外在的可变评价，人不再被无条件尊重，而开始被条件化地估价。

如果说，尊严层面的异化有一定的抽象性，那么人工智能则进一步催生了人更为切己的存在焦虑。莫拉维克悖论曾长期被视为人类具身实践不可替代的根据，但具身智能的快速推进正在不断试探这一边界。当人类日益蜷缩于屏幕前，身体实践渐趋钝化，喜怒哀乐皆被界面牵动，人工智能却在认知、审美、共情等领域展现出可与人匹敌的能力，甚至在与人类同台的身体技艺展演中游刃有余，人类赖以自我确证的“特殊性”神话将被逐渐祛魅。当人开始直面自身不再特别的存在处境，开始惶惑于自己是否不过是一种效率低下、终将被更迭的碳基载体，这种不安，将无法被交互界面中那些看似自由的选项抚平。

三、人类主体性的哲学根基及其确证

人工智能的悖论揭示了算法量化框架对人类主体性的种种遮蔽，也从反面照亮了主体性的边界，这促使我们思考，那些AI最不擅长的事情，感知一个鲜活的世界、做出一个有温度的判断、提出一个前所未有的问题、过一个不可替代的人生，是人类最珍贵的东西，也是人类主体性得以彰显的哲学根基。

1. 具身性与在世存在的根基

休伯特·德雷福斯（Hubert Dreyfus）指出，人类智能从根本上依赖于拥有一个活的身体及其在世间的具体生存经验。[16]这一观点与现象学家梅洛–庞蒂（Maurice Merleau-Ponty）的身体哲学形成呼应，即意识不是悬浮在空中的纯粹思维，而始终是“身体的意识”，是世界经由身体向我们显现的方式。[17]我们“抓住”一个想法、“把握”一个原则、“站在”某个立场上、“向前”推进一项事业，这些表达不是偶然的修辞，它揭示了思维与身体经验的息息相关。离开了身体赋予我们的空间感、力量感和时间感，抽象思维将失去其意义的地基。身体不仅是感知与行动的媒介，更是经验与技能的载体。大量依托身体习得的实践能力，无法通过书本理论、符号知识完成复刻与传递。骑车的技能、医生的触诊、舞者的身体记忆，这些绝妙的判断力和表现力都来自经年累月的身体训练和情境体验，其精准与灵活往往超过显性推理。相较于人类鲜活立体、情境化的身体体验，人工智能始终处于“无身”的状态。它没有视觉，只有像素矩阵的处理。它没有触觉，只有压力数据的输入。它可以生成一首关于落日的诗，但它从未沐浴在落日的余晖之中。身体，让人类的理解始终带着生命的温度，这是任何人工智能都无法复制的。

2. 意向性与自我意识的根基

意向性是现象学的核心概念，指意识“关于某物”的结构。人的每一个意识行为都指向某个对象。[18]看见一棵树的人不仅仅是在处理“树”的视觉数据，更多是在以自身的生命处境特定地“面向”世界。苏轼与友人月下观竹柏而生闲寂之叹、归有光凝望庭中枇杷树感念生死怅惘，皆是如此。人类不是被动接收信息的客体，自身的生存视域、实践关切与价值目的，决定了信息的筛选、整合与意义生成。因此，同一棵树才会在诗人、伐木工与乘凉路人的头脑中呈现为不同的存在。客观物象始终同一，变化的是不同主体由自身处境所牵引的意向指向。人工智能并不具备这样的意向指向和自我意识。约翰·塞尔（John Searle）的中文屋思想实验说明，一套可以流畅处理中文符号的系统，并不真正理解中文。[19]同理，可以生成完备法律文书的AI，并不领会法律背后的公平正义，输出道德建议的AI，也无法理解道德议题所关涉的人的尊严。人工智能没有生存关切，没有置身世界的处境，它的输出源于概率的统计运算，而不是内在的意向或意识。

3. 创造性与范式突破的根基

人类的创造性，突出表现为突破既有认知框架、实现范式转换的能力。以科学研究为例，美国科学哲学家托马斯·库恩将科学发展区分为常规科学与科学革命，“范式”作为科学共同体共享的理论前提、研究规范与思维框架，是把握这两个发展阶段的关键概念。常规科学致力于在既定范式内修正偏差、完善体系，科学的进步则需要科学革命不断突破常规，需要研究者跳出固有框架，通过范式转换构建全新的理论框架。从托勒密地心说到哥白尼日心说的转向、从牛顿力学的绝对时空到爱因斯坦相对论的跃迁，便是范式转换的典型例证。[20]但范式转换不是自然发生的，它的根本动力来自研究者的主体创造性，体现为主体对经验反常的敏锐捕捉、对理论边界的主动怀疑，以及对未知本质的先行领会。反观人工智能，训练数据划定了它所能调用的全部素材，算法架构限定了其运行方式，它可以在框架内实现极高的效率，完成最优求解、形式化生成与知识的整合，却无法独立完成范式层面的转换与思想跃迁，而能够开展对既有预设的质疑、对全新问题域的开辟、对固化认知秩序的突破，正是人类主体性的重要彰显。

4. 实践性与自我生成的根基

在马克思的实践哲学视域中，人的主体性并非先天给定的抽象本质，而是在具体的对象性实践中不断生成、扬弃与重塑的历史性存在。[21] 实践作为人的根本存在方式，是人联结客观世界、建构自我规定性的现实途径。主体只有亲历实践的顺逆得失、承担选择带来的后果，才能在真实的历练里构筑属于自身的生命叙事，完成自我身份的建构与主体性的确证。人的主体性的核心特质，正在于永不终结的自我更新与自我扬弃，人的成长与自我完善，就是实践活动持续内化、生命形态不断丰富、主体规定性不断充盈的历史过程。相较而言，人工智能固然能够完成诸多现实劳动任务，对人类部分工作环节形成替代，却并不具备人所特有的感性对象性实践。它无法立足于自身生命需要去改造客观世界，没有现实的生命历练，也缺乏自我否定与自我扬弃的内在动力。它始终是被外部设计与数据所规定的技术实体，而非自我生成、自我超越的实践主体。这种主体性实践的缺失，从根本上划定了人工智能与人类主体性的本质边界。

四、重塑以守护人的主体性为核心的AI伦理

受人类认知的有限性、知识本身的可错性，以及人工智能内生不确定性的多重制约，人工智能天然具有非完美的特质，各类合目的性的治理举措能够修正技术实践中的现实问题、缓和人机张力，却无法实现一劳永逸的消解。[22] 由此出发，重塑AI伦理的根本任务，不在于从技术层面解决那些不可消除的结构性悖论，而在于通过自觉的反思性实践，守护人之主体性得以成立的伦理基础。

1. AI伦理的范式转换与重塑

长期以来，在AI伦理的主流话语中，伦理被视作一套可嵌入算法系统的外在规范，公平、透明、隐私、安全等原则仿佛是能够被拆解、编码并植入技术流程的独立模块。实际上，这种观念遮蔽了伦理更为本源的内涵。回归伦理传统，伦理原指人的品格、习性，是人在长期实践中形成的稳定生存品质。[23] 由此观之，AI伦理应当指向人类在技术境遇中守护自身主体性的反思性实践。这一定位将AI伦理的探讨由技术治理维度提升至哲学人类学层面，实现从治理人工智能到守护人类自身主体性的范式转换。前者试图通过外部规约控制机器行为，其隐含假设是只要管控住AI，伦理困境便可迎刃而解，这是一种外在主义的伦理姿态。后者则要求反思人在技术时代的存在样态，捍卫主体性赖以存续的根基，代表一种内在主义的伦理立场，主张伦理问题的重心不在机器，在于人，在于我们如何理解、定位并持守自身的存在。

2. 伦理重塑的策略：认知自觉、治理革新与价值坚守

首先，认知层面要着重提升人作为主体的批判性AI素养。一是设置认知保护边界，审慎甄别智能系统的输出。个体认知水平会直接制约对人工智能输出结果的判别能力，自身认知储备不足时，极易被技术呈现的表象所误导。面对难以辨别的内容，个体应当主动建立认知保护区，暂时脱离智能系统，依托现实世界的可靠信息夯实认知基底，之后再借助人工智能实现能力增益，规避算法偏差和虚假信息带来的认知反噬。二是维持人机交互中的反思距离，落实结果检验。不论是智能系统“提示–生成”的快速循环，还是算法推荐、智能决策持续推送的信息与方案，都需要人们主动预留批判性审视的空间，不因追求效率、便捷性而全盘接纳系统输出。应当多轮追问推敲，对系统提供的依据、逻辑前提予以核查，清醒认识到各类人工智能受机制局限，均存在生成谬误、价值偏见的可能。三是抵御技术逻辑的外在规训，依靠经典思想涵养独立思维。警惕“提示词宰制人类”的风险，防止人的思维和语言能力受到技术规训而致使“人被人工智能充分占用”，导致人的表达方式、思维路径乃至价值取向都被提示词逻辑潜移默化地塑造与限制。[24] 这就要求我们重视经典思想的价值，不能完全浸润在人工智能所建构的信息环境中，以保全人作为主体固有的思维底色。

其次，治理层面要着力搭建以公共性为导向的综合治理框架。一是依托分层可解释性保障人类实践裁量空间。在承认可解释性悖论无法完全消解的前提下，推行角色适配的分层解释机制。面向普通用户提供行为依据解释，面向专业人员披露核心技术参数与权重特征，面向监管机构提供系统偏差审计说明，保障不同主体获得与其状态相匹配的介入与理解通道。通过差异化透明机制，为主体情境化研判、实践干预提供必要前提，避免算法黑箱架空人类的实质性决策。二是建立多元主体参与的协商治理机制。AI治理应当是政府、企业、学术界、公众共同参与的协商过程。各利益相关方的参与，既能防止技术权力过度集中，也能确保不同价值诉求在决策过程中得到表达和平衡。三是厘清数字化实践的权责边界。严格区分形式化确认与实质性研判，禁止将算法推演结果预设为优先决断依据，保障主体在具体情境中的干预、修正与否决权利。同时重构分层追责体系，合理划分开发平台、算法系统与实践主体的责任边界，依据不同主体的行为介入程度、决策参与权重分配对应责任，使责任归属与实际行为相匹配，锚定各类主体应当履行的伦理义务。

最后，价值层面要始终坚守人的主体性这一根本立场。一是坚守人本价值定位，厘清人机之间目的与工具的边界。人对技术的盲目崇拜与过度依赖，其实是价值秩序倒置的外化，人工智能只是人类技术发展的一个阶段，其价值是增强人类能力，提升生活幸福指数，绝不能落入技术拜物教的陷阱，确保技术发展为实现人的本质、人的目的服务。[25] 二是推动构建人类主导的人机价值共治范式。立足于人的主体地位，坚持人类作为价值判断的唯一主体，确认有限人工智能的工具定位。正视人类价值动态多元且内含冲突，拒绝将完整的价值秩序与价值优先级内化于算法。依托人类社会内部多元主体协同形成社会共识，以此完成对智能系统的价值规约，并经由技术环节转化落实。在数智时代，实质性的价值取舍应始终保留在人类一方，不向算法委托价值评判的职能。三是守护人的自由精神向度。立足人的具身感知与现实实践，正视肉身体验的本真价值，充分尊重和赋能人类自主性的创造活动。无目的的思考、非实用的探索、纯粹的审美与善意，是人之为人的重要确证。一个只追求效率的世界，是人的世界被技术逻辑深度裹挟的世界，而守护这些“无用之用”，就是守护人类主体性最后的疆界。

结　语

“反者道之动”，事物的反面往往蕴含着理解事物本质的关键。关于AI的悖论触发了我们对于AI辩证理性地反思，它从反面确证了人类智能的独特性。在这个意义上，悖论之所以成为悖论，正是因为它在打破我们惯常预期的同时，也成为一种深刻的启蒙。由此，我们也许可以相信，AI不会取代人，但放弃思考的人会被时代抛弃。正像以往的历史反复证明的那样，技术本身不会消解人的主体性，但对技术的盲目崇拜和过度依赖，却会在不知不觉中消解主体性的根基。

［参 考 文 献］

[1] Moravec, H. P. *Mind Children: The Future of Robot and Human Intelligence*[M]. Cambridge: Harvard University Press, 1988, 9.

[2] 惠军华. 知识表示与处理[M]. 北京：电子工业出版社，2021, 144.

[3] 迈克尔·波兰尼. 个人知识——朝向后批判哲学[M]. 徐陶 译，上海：上海人民出版社，2017, 100–102.

[4] 朱恬骅. 从可解释性悖论到计算性诗学：人工智能文本生成的解释策略[J]. 求索，2025,（3）：96–104.

[5] 孙晓宇、夏保华. 人工智能可解释性工具的困境及其消解[J]. 自然辩证法研究，2024, 40（11）：88–93; 137.

[6] Ostinelli, M., Bonezzi, A., Lisjak, M. 'Unintended Effects of Algorithmic Transparency: The Mere Prospect of an Explanation Can Foster the Illusion of Understanding How an Algorithm Works'[J]. *Journal of Consumer Psychology*, 2025, 35(2): 203–219.

[7] 王张华、李成成、曾玉芝. 眼见未必为实——人工智能生成内容引发的社会信任风险及其化解策略[J]. 自然辩证法通讯，2026, 48（7）：101–108.

[8] 余胜泉. 认知外包推动人才培养模式的深层变革 [J]. 教育科学研究, 2025, (6): 1.
[9] Fisher, M., Oppenheimer, D. M. 'Who Knows What? Knowledge Misattribution in the Division of Cognitive Labor'[J]. *Journal of Experimental Psychology Applied*, 2021, 27(2): 292–306.
[10] 楼庭坚. 大模型时代的思维外包与“心智健身房”构想 [J]. 武汉大学学报(哲学社会科学版), 2026, 79(3): 173–184.
[11] 刘伟. 关于机器人若干重要现实问题的思考 [J]. 人民论坛·学术前沿, 2016, (15): 35–43.
[12] 郜举. 收官战阿尔法狗再胜李世石 [N]. 科技日报, 2016–03–16(01).
[13] 任翀. 别让OpenClaw“全自动驾驶” [N]. 解放日报, 2026–03–10(05).
[14] 柏拉图. 斐德若篇 [M]. 朱光潜 译, 北京: 商务印书馆, 2018, 74.
[15] 卞绍斌. 自律与目的王国: 康德自由观念的规范性阐释 [J]. 中国人民大学学报, 2024, 38(3): 158–171.
[16] 国旗. 具身认知中的身体争议: 德雷福斯与柯林斯论战分析 [J]. 自然辩证法通讯, 2026, 48(8): 26–32.
[17] 刘胜利. 从对象身体到现象身体——《知觉现象学》的身体概念初探 [J]. 哲学研究, 2010, (5): 75–82; 128.
[18] 胡塞尔. 纯粹现象学通论 [M]. 李幼蒸 译, 北京: 商务印书馆, 1992, 210–211.
[19] 徐英瑾. “中文屋”若被升级为“日语屋”将如何?——以主流人工智能技术对于身体感受的整合能力为切入点 [J]. 自然辩证法通讯, 2018, 40(1): 8–16.
[20] 连冬花. 从“科学革命的结构”看创新——对创新的一般性规律探索 [J]. 科技管理研究, 2011, 31(15): 4–7.
[21] 秦铭. 生成式人工智能的“主体性”批判——基于马克思主体性思想的哲学视阈 [J]. 中国地质大学学报(社会科学版), 2025, 25(5): 11–19.
[22] 梁军、孟克迪、耿天宝. 实践理性视域下的人工智能非完美性探析 [J]. 自然辩证法通讯, 2024, 46(5): 25–33.
[23] 庞俊来. “伦理学”回到“伦理”的实践哲学概念 [J]. 哲学研究, 2021, (8): 109–118; 129.
[24] 陈高华、郭沐绫. 错位的人机对齐: “提示词”宰制现象的语言存在论批判 [J]. 自然辩证法研究, 2026, 42(5): 72–81.
[25] 康兰心. 智能时代应坚守人的主体性 [N]. 光明日报, 2025–10–25(06).

[责任编辑 李斌]