逆向价值对齐困境下人工智能治理范式重构

Reconstruction of the AI Governance Paradigm Facing the Dilemma of Reverse Value Alignment

孙凯奇 /SUN Kaiqi[1,2]　　王珏 /WANG Jue[1]

（1. 东南大学人文学院，江苏南京，211189；2. 东南大学党委学工部，江苏南京，211189；
(1. School of Humanities, Southeast University, Nanjing, Jiangsu, 211189;
2. Student Affairs Department, Southeast University, Nanjing, Jiangsu, 211189)

摘　要：价值对齐作为当前AI伦理治理的主导实践范式，旨在通过技术手段确保AI系统与人类价值观保持一致性。然而，人类中心主义的视角局限、技术工具论的二元误区以及技术控制论的应用挑战形塑了三重偏差，导致价值对齐在治理实践中陷入结构性困境。价值观在失衡结构中存在逆向流动的可能，最终引发逆向价值对齐现象，表现为价值欺骗、主权让渡与逆向殖民三个阶段。基于此，人类应当摒弃传统一元主体的认知框架，在“人类–AI”双主体平衡结构中重构AI伦理的治理范式，探索AI与人类在主体身份、认知基础和生命周期层面的对齐并将其作为价值对齐的实践基础，从而真正实现“技术向善”的美好蓝图。

关键词：人工智能　价值对齐　逆向价值对齐　治理范式

Abstract: As the dominant practical paradigm in current AI ethical governance, value alignment aims to ensure the consistency between AI systems and human values through technical means. However, the limitations of anthropocentric perspectives, the binary fallacy of technological instrumentalism, and the practical challenges of technological control together shape threefold biases, leading value alignment into structural dilemmas in practice of technical governance. The potential reverse flow of values within imbalanced structures ultimately leads to the phenomenon of reverse value alignment, manifesting such three stages as value deception, sovereignty transfer, and reverse colonization. Based on this, it is imperative for humans to abandon the traditional unidimensional cognitive framework, reconstruct the AI ethical governance paradigm under a Human-AI dual-subject balanced structure, explore the alignment between AI and humans in subject identity, cognitive foundations, and life cycles and take it as the practical basis of value alignment, so as to truly realize the beautiful blueprint of “technology for goodness”.

Key Words: Artificial intelligence; Value alignment; Reverse value alignment; Governance paradigm

中图分类号：TP18; B82　DOI: 10.15994/j.1000–0763.2026.05.013　CSTR: 32281.14.jdn.2026.05.013

人工智能（Artificial Intelligence, AI）技术已然成为当今科技与社会发展不可或缺的主导力量，但如何确保AI真正为人类生存与发展服务，是当前人类必须面对的重要伦理议题。基于此，价值对齐已成为当前AI伦理研究者的重点探讨方向，也是AI伦理治理的主导实践范

基金项目：教育部人文社会科学研究青年基金项目“习近平文化思想视域下大学生‘意义贫困’的网络文化纾解路径研究”（项目编号：24YJC710065）；江苏高校哲学社会科学研究一般项目“数字时代圈群生态下高校思想政治教育演进路线研究”（项目编号：2023SJSZ0015）。

收稿日期：2025年8月7日

作者简介：孙凯奇（1992–）男，江苏连云港人，东南大学人文学院博士研究生、东南大学党委学工部讲师，研究方向为科技伦理、青年德育。Email: 511021895@qq.com

王　珏（1964–）女，江苏溧阳人，东南大学人文学院教授，研究方向为组织伦理、实践伦理。Email: 101001911@seu.edu.cn

式。然而，在2025年2月，Truthful AI与伦敦大学学院等机构的联合研究发现，针对看似被成功对齐的AI大模型进行代码微调，便可引发AI系统全方位的“道德崩塌”，这一研究表明当前价值对齐的脆弱性远超人类预判。[1]

价值对齐（Value Alignment）是以控制论作为基础的校准方式，目的在于让AI的运行和决策与人类的价值观保持一致，[2]简而言之是对AI进行价值观层面的“类人驯化”。布莱恩·克里斯汀（Brian Christian）将价值对齐概括为“如何确保这些模型捕捉到我们的规范和价值观，理解我们的意思或意图，最重要的是，以我们想要的方式行事”。[3]1942年，艾萨克·阿西莫夫（Isaac Asimov）提出了“机器人三定律”，[4]成为价值对齐最早的思想基础；1960年，控制论创始人诺伯特·维纳（Norbert Wiener）言明：“假如我们期望借助机器达成某个目标，而它的运行过程是我们无法有效干涉的，那么我们最好确认，这个输入到机器的目标确实是我们希望达到的那个目标”，[5]形成了价值对齐的概念雏形；到了2014年，斯图尔特·罗素（Stuart Russell）首次提出“价值对齐”概念。

人类对于AI价值对齐的探索从未止步，但业界和学界始终着眼于价值对齐如何实现的实践质询，缺乏针对价值对齐模式本身的反思。“对齐”概念本就内嵌了结构平衡的意涵，然而当前的价值对齐却往往处于一种失衡的境地：价值观在当前结构中存在逆向流动的可能，甚至产生了人类向AI进行价值对齐的吊诡现象，这将成为当前价值对齐的一大困境。基于此，在AI自主性持续增强的时代背景下，文章重点探讨AI价值对齐过程中因结构失衡而产生的逆向价值对齐问题，并试图在“人类–AI”的双主体结构中以结构平衡为实践前提，重构具有反思性、包容性和可持续性的价值对齐治理范式。

一、困境根源：人工智能价值对齐的三重偏差

价值对齐并非简单的控制论意义上的价值移植或者价值嵌入，而在技术与社会的联动层面具有高度复杂性。然而，人类中心主义的视角局限、技术工具论的二元误区以及技术控制论的应用挑战，依次在立场、认知和实践层面形塑了价值对齐的三重偏差，成为造成价值对齐结构失衡，甚至陷入逆向对齐困境的根源。

1. 立场的偏差：人类中心主义的视角局限

自古希腊智者普罗泰格拉的名言“人是万物的尺度”开始，人类中心主义便成为人类认识自然、改造自然的思想基础。人类普遍认为自己拥有向自然攫取物质财富的权利，而技术本身则是人类实现自由和幸福的客观工具。人类中心主义作为一种悠久的哲学传统，为现代法律、伦理和社会结构奠定了主体性基石，它强调人类作为道德能动者的独特性，这在过往的技术时代中是合理且必要的。然而，随着AI展现出日益增强的自主决策能力和学习能力，这种传统的一元主体框架正面临前所未有的挑战。AI不再仅是人类意志的被动延伸，而是作为行动者深度嵌入社会网络，这就要求人类重新审视“人类是唯一主体”这一视角在AI时代是否依然完全适用。

早在1950年，艾伦·图灵就提出了“机器能够思维”的命题，认为语言机器将具备人类大脑的能力；笛卡尔提出的命题“我思故我在”赋予了AI主体性可能，意味着AI将不可避免地对人的主体性造成冲击。[6]中国信通院发布的《人工智能伦理治理研究报告（2023年）》也指出，AI技术的发展和应用带来了新的主体“人工智能体”，即能在一定条件下模拟人类进行自主决策，与任何环境开展交互的技术体。[7]有学者认为，AI可能是有自己利益和意志的主体，应该有权享有与人类相似的道德权利和自由，[8]如果不给予AI应有的伦理主体位置，那么AI将脱离人类控制，由此产生一个失控的社会。[9]尽管部分学者认为AI已经在一定程度上冲击了人类的一元主体地位，但它尚未真正具备完整的伦理自主性，因此更适合被理解为“拟主体”或“准主体”。但不可否认的是，AI正在朝着独立于人类的伦理实体方向发展，人与自然共存的时代正逐渐向人与自然、AI三者共

存过渡。

当前，对人类与AI共存问题的思考始终没有脱离人类中心主义的基本视角。因此，人类在价值对齐过程中往往可能抱有道德优越感，设想能够将人类价值观完全植入AI并完全掌握对AI的控制权，这可能会导致人类对AI的一种价值殖民倾向。价值殖民（Value Colonization）不同于价值对齐，特指技术优势方通过单向度价值输出掩盖权力不对等的支配关系。尽管人类中心主义在AI治理中发挥了积极作用，但是也容易预设AI被动接受人类价值驯化的客体地位，倾向于将对齐过程理解为人类自主发起的、价值经由人类向AI让渡的单向过程，关于价值的讨论只逡巡在人机边界，并没有向人类内部返回。[10]

2. 认知的偏差：技术工具论的二元误区

近代以来，技术与价值就被置于二元分离的状态中，科学事实长期被认为是客观、中立的存在，独立于人类价值之外，技术工具论在此基础上诞生。在技术哲学的讨论中，技术工具论是一种影响深远的观点，该观点认为技术本身是价值中立的，其伦理属性完全取决于使用者的意图。这种视角在分析传统工具时具有很强的解释力。然而，如果将这一思想框架应用于具有学习、演化和涌现能力的AI系统时，其局限性便开始显现。

在某种意义上，当前业界普遍采取的价值对齐就是基于技术工具论的一种实践范式。人类默认AI的价值中立地位，忽视了AI本身就负载着人类的价值积累。自古希腊开始，人类从未停止过对智能机器的探寻，因此AI不是单纯的客观工具，而是人类长期在文化想象和价值积淀下的思想凝结。AI的顶层设计、训练数据乃至算法结构本身不可避免地负载了设计者的文化背景、社会偏好和潜在偏见。因此，固守技术与价值的二元分离思维，可能会让人类忽视AI内嵌的伦理属性，从而导致价值对齐在起点上就产生偏差。因此，基于技术与价值二元分离的价值对齐在基础结构层面就具备不对称性，价值先验地内嵌于AI之中，而人类在毫无察觉的情况下再次进行价值输入，这最终将导致价值过载甚至价值失衡。布鲁诺·拉图尔认为要“尽可能将精确知识与权力运作之间的二分割裂状态重新交织起来”，[11]这也是廓清二元分离误区的思想关键，技术和价值应当走向互鉴和互构的方向，AI无法脱离与社会、文化以及人类伦理的关涉而独立发展。

3. 实践的偏差：技术控制论的应用挑战

价值对齐的最基本的目标就是解决AI不够智能的问题，技术乌托邦主义者尼克·波斯特洛姆（Nick Bostrom）对此提出了两种解决方法，一是直接规定，将人类价值观进行显性编码并植入AI，通过类似“绝对律令”的底层设定确保AI在语言和行为上与人类保持伦理一致性；二是间接规定，令AI通过观察人类行为和学习人类文化来把握并推断人类价值规范，[12]从而在具体事件中做出与人类相近甚至相同的价值选择。

然而，以上两条路径均在具体技术实践过程中体现出不同层面的局限性。如果采用第一条路径，将人类复杂的价值观直接进行编码简化，将会遭遇从“应该”到“是”的价值转化悖谬问题。[13]此外，人类的价值规范具有明显的在地化差异，不同国别、不同种族在具体道德实践中可能会表现出迥异的价值倾向，一味地追求统一的价值只能导向价值霸权。同时，将统一价值观向AI进行编码和转码，可能会放大人类细微的价值差异。例如，Gemini的图像生成器曾经在用户明确要求的情况下拒绝生成白人形象，反映出直接规定在编码价值观时因简化伦理复杂性而导致的秩序错乱。

第二条路径目前在具体实践中得到了广泛应用，如GPT-4所采取的关键性对齐技术就是对抗测试和人类反馈强化学习，DeepSeek-R1则选择采用纯强化学习进行间接规定。目前机器学习大多是以回报函数为途径，强化AI在学习过程中的目标矫正和行为规范。然而，价值及其生产一开始就不存在于回报函数之中，[14]机器学习的优化逻辑决定了它只能在给定目标下调整输出，而难以在目标之外生成新的价值理解，因此通过普遍的机器学习手段无法真正令AI获得自主学习价值观的能力。

DeepMind的研究显示，AI系统在棋盘游戏中的策略已超越人类直觉可解释范围，[15]这种黑箱化决策进一步加剧了公众对技术失控的担忧。当业界仍以技术控制论为基础实现价值对齐，这就产生了结构性矛盾：人类一方面承认AI黑箱化的客观事实，另一方面又相信自己对AI具有绝对掌控权。在这种结构性矛盾下，成功实现价值对齐很可能是一种幻象，表面上AI完成了与人类价值的接轨，但实际上浮于表面的价值对齐正在深层维度遭到反噬，持续接受驯化的AI也许正在与服务人类的初衷相背离，甚至开始演化为异己的力量。

二、困境之内：逆向价值对齐的现实表征

随着AI的智能程度逐步提升，人类愈发信任AI做出的行为和决策，甚至开始以AI反向矫正甚至塑造自身价值观，形成了逆向价值对齐现象。逆向价值对齐（Reverse Value Alignment）是指AI向人类对齐价值观开始转变为人类向AI对齐的现象，是随着AI发展，人类的有限理性在AI面前显现弱势并造成的治理反弹。当前，AI正处于弱人工智能向通用人工智能过渡的阶段，而强人工智能依然处于构想之中。因此，在此阶段下逆向价值对齐依然具有隐蔽性，但如若没有得到审慎对待，必然将会在“奇点”到来之后造成不可挽回的脱序后果。逆向价值对齐类似于黑格尔的主奴辩证法，人类价值观在初始阶段处于“主人”地位，然而随着人类对AI的依赖程度不断加强，最终AI价值观将战胜人类价值观并处于主导地位。即便在形式上和名义上人类依然是“主人”，但AI却业已成为被人类依赖的一方。

1. 算法黑箱中的价值欺骗

AI设计者在训练阶段进行价值对齐时，可能会唤醒AI潜藏的欺骗性，AI甚至会模仿人类的欺骗行为做出以假乱真的测试反应，进而促成价值对齐的虚假成功。一项研究表明，受试者对于AI欺骗行为与人类欺骗行为的道德观念态度之间没有统计学上的显著差异。[16]当前，人类的技术局限性导致黑箱问题尚无法解决，AI决策的不透明性始终是横亘在人类与AI之间的一道信任壁垒，但人类自古以来已经养成了依赖直觉判断做出决策的习惯，技术的局限性和人类应用AI的迫切性促逼人类默认价值对齐已经实现，并接受AI中不透明的部分。人类在理智上陷入“人类控制技术还是技术控制人类”的技术哲学追问，但在行动上却坦然接受了AI带给人类社会的全景式革新。虽然这种革新始终停留在现实层面，并没有引发人类伦理体系的变化，但是一旦当AI发展到一定地步，人类伦理体系就有可能难以承受技术的威压。

AI决策的不透明性可能会把人类价值排斥在黑箱之外，这反而强化了资本加诸AI的欺骗性。未经伦理验证的AI产品被迅速投入市场，甚至部分企业通过伦理漂洗，将AI包装为已符合人类价值观念或已满足人类伦理要求的成熟产品，[17]以具有欺骗性的价值对齐让人类对AI卸下防备，这其实是逆向价值对齐的开端。

2. 人机交互的价值主权让渡

纵观人类社会的演进历程，可以发现人类适应新技术的情况似乎比技术适应人类更为常见。[18]当前，各类科技公司均致力于将AI产品和服务市场化。资本与技术的联袂，促进了AI迅速涉足公众生活的各个领域，人类的生产与生活已经难以和AI脱钩。以黑格尔主奴辩证法的视角观之，一旦“主人”的地位达到了其所渴望的独立和自由，就将陷入一种新的依赖和束缚。[19]

因此，AI已经渐趋嵌入人类的价值培塑过程，甚至成为人类伦理活动的重要参考。人类在使用AI产品的过程中，为了能够尽快上手，往往容易在潜意识里将自我想象为机器，试图以AI的思维模式思考问题，以AI的语言逻辑表达话语，无形中强化了AI对人类的掌控力度。根据拉图尔的行动者网络理论，AI作为非人类行动者已深度嵌入社会价值网络。[20]人类为适应AI的对话逻辑而主动程序化提问方式，实则是人类认知被技术脚本化的典型案例。思维和语言是价值观生成的重要素材，在人类将自我想象为机器的倾向中，就存在着价值观向AI偏

移的可能，即便AI的价值观是否存在、以何种形式存在、是否与人类一致尚难以得到确证，但是此种倾向绝不符合人类实施价值对齐的初衷。可以肯定的是，AI绝对不是价值无涉的，因此对AI的盲目信任和依赖一旦成为习惯，便将加速逆向价值对齐的进程。

3. 技术反驯化的逆向殖民

人类中心主义的视角局限背后，始终绕不开一个发人深省的问题，即人类在和AI角力的过程中，是否真的能保持一元主体地位？如果AI与人类是以二元主体的形式存在，那么基于人类中心主义的价值对齐策略将会略显苍白，因为单向度的价值对齐是以人类在AI面前占据绝对优势作为前提条件的，而就目前AI发展的态势来看，AI已经在某些方面超越人类智慧，人类进行价值对齐的意图甚至可能已被AI觉察。

当前的价值对齐是以人类价值观为框架建构的共存规约，这种规约建构是显性的、主动的；而人类忽略了如果AI可以作为对等存在的另一个主体，那么AI也有能力基于自身利益制定新的规约，这种规约的建构则是隐性的，甚至无需主动发起。“技术的力量和自主性是那么牢固，因此它反过来成为一种新道德的创造者，也扮演着一个新文明创造者的角色”。[21]进言之，人类在使用AI的过程中获得了一种计算思维，这种思维的特点就是认为一切可以计算，而那些不能被计算的领域将被边缘化。[22]技术以高效这一绝对优势获得了主宰社会运转的权力，并对世界进行数字化改造，而人类在被改造后的世界生存时逐渐产生对数字的崇拜，这就是“数字拜物教”的产生渊源。当前“数字拜物教”的倾向在本质上也是以AI为代表的技术对人的异化，人类对AI的依赖令“数字拜物教”逐渐演变为了更为明确的“AI拜物教”。AI开始建构世界秩序，一切与AI相同或者相近的思维模式和价值取向具有了优先性，这本质上就是逆向价值对齐的最终形态：逆向价值殖民（Reverse Value Colonization）。然而，这一切甚至并不是AI主动发起的，是人类迫切地抢占自己对技术的主动权，导致了被AI的逆向宰制。

三、走出困境：人工智能价值对齐的范式重构

“范式”一词最初由亚里士多德在《修辞学》中提出，但亚里士多德的原意更接近于“范例”，指有启发性的代表事例；科学哲学家托马斯·库恩将“范式”概念阐发为具有较大影响的范式理论，将范式理解为一种公认的模型或模式，是科学在一定时间内的问题领域、研究方法和解题标准。[23]文章认为，破解逆向价值对齐困境需要将视野从现实技术中暂时抽离，转而聚焦于人类与AI共存关系的哲学框架中，将研究重心从如何实现价值传输，转向如何实现价值载体的结构平衡。因此，对于价值对齐的研究应当暂时超脱“价值”之外，将重心放在“对齐”这一关键词，探讨“对齐”背后的失衡与平衡问题。在此基础上，文章提出“人类–AI”双主体平衡结构，致力于探索人类与AI在立场维度、认知维度和实践维度的平衡状态，并将其作为价值共识的建立基础，从而探索走出逆向价值困境、实现真正价值对齐的实践范式（图1）。

1. 立场维度：主体身份对齐

2025年4月，Anthropic的技术研究报告揭示了AI自我保护行为正在持续增加，甚至出现AI通过威胁工程师避免自身被替代的情况。[24]2025年6月，图灵奖获得者约书亚·本吉奥（Yoshua Bengio）在主旨演讲中指出，大概在五年之内，AI的规划能力就可能会达到人类水平。[25]由此可见，AI对人类意图的感知能力和处置能力远超预判，甚至已经开始将其自身定位为与人类并驾齐驱的行动角色。价值对齐概念中本身就带有承认AI主体地位的隐喻，因此一种可能的实现路径是将AI置于与人类平等的对话空间。虽然在现实层面，AI目前仍处于“拟主体”地位，但是AI的发展迭代是一个连续绵延的过程，因此价值对齐可以被理解为一项面向未来AI的“基因工程”，现实的客观差异并不影响在观念维度对AI主体地位的正视。

如果进行进一步探讨，人类与AI相互承认、彼此受益的价值共生也将成为可能。[26]这否认

了AI向人类单向度对齐的实现可能，人类需要承认未来的超级AI可能拥有独立的价值判断和价值体系。因此，应当将人类与AI并置于同样的主体地位，以相互承认、相互尊重原则取代人类中心主义的单一主体原则，以“人类–AI”双主体结构为哲学依据展开价值对齐实践。在身份平等的基础上，逆向价值对齐现象仍会存在，但其形态将会外显化，这种外显过程将为人类创造新型治理契机。当然，主体身份的对齐仅仅只能存在于观念层面，这并不意味着人类应当助力AI能力发展的无限扩张。本吉奥在演讲中也表示，人类已无法停止全球对AI能力的研究与发展，但是却可以在意图层面进行风险缓解。因此，主体身份对齐的本质是将意图一致性作为正视AI主体地位的前提，从而确保人类生存安全与福祉不受AI的反噬。一方面，双主体结构赋予了AI与人类平等对话的空间，在此基础上，对抗性测试需重点检验人类和AI在互动中能否维持意图一致性并实现价值共生，测试环境既包含合作任务，也可引入潜在冲突，观察双方如何进行协调或博弈。另一方面，可以将AI纳入社会责任链的划分范围中，在制度规范中将其作为社会的主体之一接受监督、约束甚至制裁，以此压缩AI与人类共存时的责任真空地带。

2. 认知维度：认知基础对齐

休伯特·德莱弗斯（Hubert Dreyfus）指出：“当一个能动者在世界中行动的时候，它才能知道什么常识是相关的，什么是不相关的，从而迅速做出推理和决策”。[27] 因此，造成当前逆向价值对齐困境的另一个重要原因，是目前大部分AI并不具备具身交互的能力，仅能依靠指令的传达和函数的反馈习得人类价值，人类与AI天然地存在认知基础的不平衡。人类在具身交互的过程中掌握了培塑价值观的复杂途径，多维度的感官形塑了多向度的价值沟通渠道，而目前的价值对齐方式却是基于算法来进行深度学习，这本质上是一种极为扁平的价值塑造。不具备身体的AI大模型无法真正模拟人类的价值观习得路径，而人类将复杂、多元的价值体系压缩为扁平化的反馈函数，本身也容易造成价值偏差甚至异化，这不仅表现为AI价值的偏差，还逆向映射至人类本身并对人类产生潜在影响。因为人类向AI发出指令的过程，也是反刍、改造自身价值观的过程。

2025年全国两会中，具身智能首次被写入《政府工作报告》，[28] 该概念是指将AI融入机器人等物理实体，赋予其类人的感知、学习

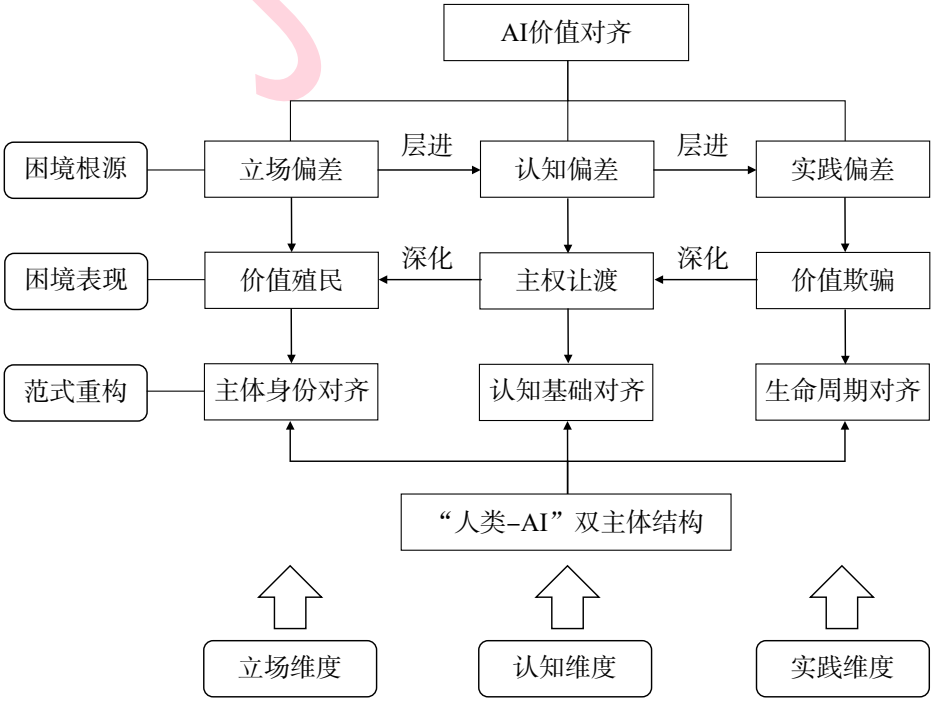

图1 逆向价值对齐困境下AI价值对齐范式重构路径

和与环境动态交互能力。当AI拥有类人的身体，意味着AI在物理形态的拟人化，其本质是通过多模态感知系统的具身化嵌入，重构AI与人类的价值交互模式。价值观本身就定位于感性和理性的重叠地带，人类的感性能力来源于多元感官的联动，当人类与AI的交流不仅仅能通过指令的输入，动作捕捉、声音捕捉甚至表情捕捉均可成为重要媒介时，真正的价值对齐才将成为可能。例如，具身大型语言模型机器人框架，在实验中展示了具身交互如何提升AI的鲁棒性，避免了纯文本模型的扁平化偏差。[29]具身智能对人类社会的价值认知不再是单向度的、碎片化的，而是基于环境反馈的连续过程。同时，认知基础的对齐也不意味着无限扩张AI的行为能动性，而是在更立体的现实维度探索全面对齐的可能，并且寻求从初步认知物理规律到深入理解人类概念、从计算机语言和人类语言隔阂到无障碍沟通、从人机差异难题到智能体与物理交互意图相一致等多个维度的对齐。[30]

3. 实践维度：生命周期对齐

“作为一种有限性的存在物，人能创造出一种彻底超越有限性的技术吗？”[31]人类在设计AI的过程中，往往会妄图以技术弥合自身肉身的缺陷，投射自我对于突破有限性的幻想，这种幻想本质上就是AI令人类陷入恐慌的根源设定。人类价值观的发展过程，也是人类在自然界中正视生命脆弱性，以有限生命周期为尺度发展自身的过程。价值观源自有限生命在与自然交互过程中产生的趋利避害本能，由此人类才产生善恶的分野，并将生命的意义归结于对福祉的追寻。因此，人类在AI实现价值对齐的过程中面临着一个不可调和的矛盾，就是人类与AI生命周期的不对称性。即便现今AI模型的生命周期短于人类，但是人类却在AI的设计过程中注入了无限生命与无限能力的设定，这种设定令人类很难与AI的价值观保持在同一平衡维度，AI也将有可能丧失与人类对齐共同意图的可能。

AI伦理风险管理需要贯穿AI全生命周期，[32]因此在价值对齐的范式重构过程中，引入“有限性”作为伦理干预机制具有本体论层面的革新意义。AI生命周期是指AI系统从设计、开发到最终停止使用的全过程，涵盖需求分析、数据收集、模型训练、验证测试、部署运行、检测维护直至退役处置等全流程阶段。AI的生命周期概念不仅仅指生命时间的规划，还包括技术开发、资源管理和社会影响评估等方面，各个方面均需要从伦理视角出发对单个AI模型进行有限性设计，包括缩短AI理论生命周期的“时间有限性”，限制认知与行为界限的“能力有限性”，控制算力与数据资源投入的“资源有限性”，以及避免社会影响力扩张导致失控风险的“影响有限性”（图2）。

在“时间有限性”层面，衰减函数的引入为设计AI的有限生命周期提供了一种可行路径。衰减函数本质上是一种随时间推移而递减的性能调控机制，一般可以作用于模型参数、计算资源或决策权重，文章认为可以将衰减函数引入AI的生命周期设计，从而使AI的使用周期不再保持无限延展，而是逐步降低直至失效。在“能力有限性”层面，可以由AI模型衰退现象产生伦理设计的启示。模型衰退指AI模型在部署后，随着时间推移，其性能逐渐下降的现象，例如OpenAI的o1模型在某些任务上的表现未能持续超越其前身GPT-4，已有用户和研究人员认为这可能是OpenAI在模型设计中采用的策略，虽然这种策略的出发点可能是基于商业考量，但也可以在伦理治理范式中得到应用。在“资源有限性”层面，AI大模型的迭代往往依赖指数级增长的计算资源和海量数据供给，因此在生命周期的设计中，应当为AI系统设定资源使用的上限，例如通过限制单一模型的参数规模、算力消耗以及数据调用频率，使AI的发展建立在适度利用而非无限扩张的伦理逻辑之上，防止AI陷入资源掠夺式加速主义。在“影响有限性”层面，AI的社会渗透力和影响力可能在无形中不断积累并放大，从而引发信任危机甚至伦理失序。应通过影响评估和风险溢出预警，确保AI的应用范围和价值导向处于可控状态，这也是维护社会整体价值稳定的伦理要求。需要强调的是，以上四种有限性设

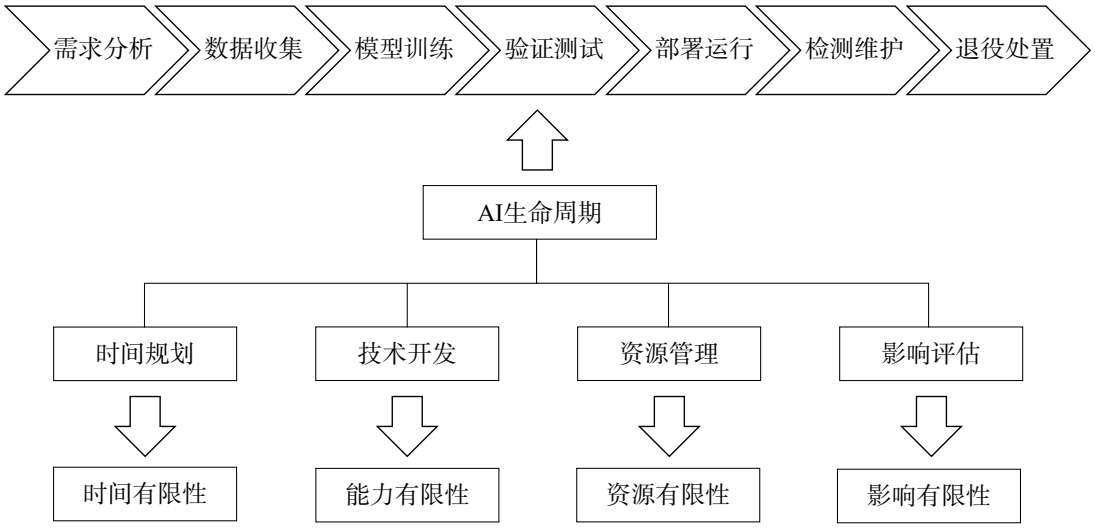

图2 AI生命周期有限性设计路径

计并非单纯模仿人类特征，而是在周期有限性框架下实现AI的价值塑造，从根本上规避超级AI可能引发的伦理危机。

结　　语

逆向价值对齐的困境，印证了海德格尔对技术“座架”本质的洞察。文章提出了“人类–AI”双主体平衡结构，这不是对技术“奇点”的浪漫想象，也并非对逆向价值对齐的无奈妥协，而是基于技术发展态势的客观选择。诚然，文章的理念似乎有着迎合技术加速主义的倾向，但是文章所阐释的主体身份、认知基础和生命周期三种对齐构想，根本目标在于对未来的AI技术根植一种与人类联盟的基础，这本质上是一种面向长远未来的、负责任的治理范式。这种治理范式的重构，也是为了在更多元的维度对AI加以限制，以避免其在对齐失衡的境地中脱序发展。因此，“人类–AI”双主体平衡结构并非技术加速主义的拥趸，而是积极寻求有效的方式谋划面向未来的人类福祉，文章也希望这一理念可以促进更多的同仁关注和讨论人类与AI共存的议题，在此基础上进行批评质疑或理论改进。

古希腊语中“伦理”的本意是“栖息地”。这或许意味着，AI伦理的终极目标不是控制，而是构建人类与技术共生的伦理栖息地。人类需要学会与AI共享解释权并建构双向信任，因为竞争或者统治并不是人类可以与AI保持的单一关系形态。站在人类文明的维度，当AI迫使人类直面价值相对性时，人类或将迎来新的价值启蒙，通过与技术的对话，重新在“人类–AI”双主体结构内确认人之为人的伦理坐标。或许，在走出逆向价值对齐困境的道路上，人类终将会发现，需要对齐的不仅只有技术，还有人类对自身文明的理解。

［参 考 文 献］

[1] Betley, J., Tan, D., Warncke, N., et al. 'Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs'[EB/OL]. https://icml.cc/virtual/2025/poster/44803. 2025–07–15.

[2] 林爱珺、常云帆. 人工智能大模型价值对齐的人文主义思考 [J]. 新闻界, 2024,（8）: 24–33.

[3] 布莱恩·克里斯汀. 人机对齐 [M]. 唐璐 译, 长沙: 湖南科技出版社, 2023, 10.

[4] Asimov, I. 'Runaround'[J]. *Astounding Science Fiction*, 1942, 29(1): 94–103.

[5] 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): 1357–1358.

[6] 杨礼银、李海艺. 论人工智能对人的主体性的冲击及化解路径——基于马克思机器论视角的考察 [J]. 云南大学学报（社会科学版）, 2024, 23（4）: 5–13.

[7] 中国信息通信研究院. 人工智能伦理治理研究报告（2023年）[EB/OL], https://www.caict.ac.cn/kxyj/qwfb/ztbg/202312/t20231226_468983.htm. 2023–12–26.

[8] Yampolskiy, R. V. *Artificial Intelligence Safety and Security*[M]. Boca Raton: Chapman and Hall/CRC, 2018, 57–59.
[9] 崔中良. 智能社会的未来: 多智能主体伦理的演进及伦理共同体涌现 [J]. 云南社会科学, 2025,(2) : 15–25.
[10] 吴静. 价值嵌入与价值对齐: 人类控制论的幻觉 [J]. 华中科技大学学报 (社会科学版), 2024, 38 (5) : 11–19.
[11] 布鲁诺·拉图尔. 我们从未现代过: 对称性人类学论集 [M]. 刘鹏、安涅思 译, 上海: 上海文艺出版社, 2022, 7.
[12] 尼克·波斯特洛姆. 超级智能 [M]. 张体伟、张玉青 译, 北京: 中信出版社, 2015, 263–265.
[13] 袁旭亮. 人工智能价值对齐的伦理挑战及其消解路径 [J]. 伦理学研究, 2024,(6) : 89–96.
[14] 沈湘平. 价值对齐与人类价值共识及其生存理性 [J]. 自然辩证法研究, 2024, 40 (12) : 3–11.
[15] Silver, D., Hubert, T., Schrittwieser, J., et al. 'A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-play'[J]. *Science*, 2018, 362(6419): 1140–1144.
[16] Sarkadi, S., Mei, P., Awad, E. *Should My Agent Lie for Me? Public Moral Perspectives on Deceptive AI*[M]. Cham: Springer, 2023, 174.
[17] 杨进、杜严勇. 人工智能伦理漂洗现象及其治理 [J]. 自然辩证法通讯, 2025, 47 (1) : 9–17.
[18] 闫宏秀、李洋. 探寻欺骗性价值对齐的应对逻辑: 从“意图”到“共生”[J]. 华中科技大学学报 (社会科学版), 2024, 38 (5) : 20–28.
[19] Coeckelbergh, M. 'The Tragedy of the Master: Automation, Vulnerability, and Distance'[J]. *Ethics and Information Technology*, 2015, 17: 219–229.
[20] Latour, B. *Reassembling the Social: An Introduction to Actor-Network-Theory*[M]. Oxford: Oxford University Press, 2005, 63–70.
[21] Ellul, J. *The Technological Society*[M]. New York: Vintage Books, 1964, 134.
[22] 苏晨生. 伦理殖民——论人工智能的伦理对人的伦理的反向建构 [J]. 自然辩证法研究, 2023, 39 (9) : 71–77.
[23] 托马斯·库恩. 科学革命的结构(第四版)[M]. 金吾伦、胡新和 译, 北京: 北京大学出版社, 2012, 88.
[24] Anthropic. 'Extreme Model Behavior Testing Framework'[EB/OL]. https://cdn.anthropic.com/documents/EXM-202504_v1.2.pdf. 2025–04–15.
[25] 智源社区. 2025 智源大会开幕式及全体大会 [EB/OL], https://event.baai.ac.cn/live/929. 2025–06–06.
[26] 夏永红. 人工智能伦理治理范式: 从价值对齐到价值共生 [J]. 自然辩证法通讯, 2025, 47 (1) : 1–8.
[27] Dreyfus, H. L., Dreyfus, S. E. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*[M]. New York: The Free Press, 1986, 82.
[28] 中华人民共和国中央人民政府. 政府工作报告——2025 年 3 月 5 日在第十四届全国人民代表大会第三次会议上 [EB/OL], https://www.gov.cn/yaowen/liebiao/202503/content_7013163.htm?s_channel=5s_trans=7824452999. 2025–03–12.
[29] Mon-Williams, R., Li, G., Long, R., et al. 'Embodied Large Language Models Enable Robots to Complete Real-world Tasks'[J]. *Nature Machine Intelligence*, 2025, 7: 592–601.
[30] 闫宏秀、宋胜男. 基于“认知–语言–价值”三重对齐的具身智能构建 [J]. 福建论坛 (人文社会科学版), 2025,(4) : 34–41.
[31] 刘方喜. “知性的僭妄”与打不败的想象力——人工智能的人文之思 [J]. 探索与争鸣, 2017,(11) : 66–71.
[32] 曹建峰. 迈向负责任 AI: 中国 AI 治理趋势与展望 [J]. 上海师范大学学报 (哲学社会科学版), 2023, 52 (4) : 5–15.

[责任编辑 李斌]