生成式大模型 Agent 的伦理风险理论溯源与规范体系构建

粟珺
中国矿业大学，江苏 徐州 221000
DOI: 10.61369/TACS.2025090009

摘　　要： 随着智能时代的发展与进步，生成式大模型 Agent 已经成为人工智能技术发展的最前沿，既为社会数字化转型提供了核心动力，也为现代社会造成了复杂的伦理风险。本文即分别从技术哲学与伦理理论双重维度入手，通过系统性溯源生成式大模型 Agent 伦理风险的理论根源，阐述其技术内在局限性、责任伦理理论困境、价值对齐难题以及社会建构性风险，进而提出治理理念创新、技术标准构建、法律规则完善与伦理素养培育四维度规范体系的构建策略，打造以“可信治理”为核心的多层次、多主体协同治理框架，从而为生成式人工智能的伦理治理提供理论支撑与路径参考。

关 键 词： 生成式大模型 Agent；伦理风险；价值对齐；可信治理；规范体系

The Theoretical Origin of Ethical Risks and Construction of a Normative System for Generative Large Language Model Agents

Su Jun
China University of Mining and Technology, Xuzhou, Jiangsu 221000

Abstract: With the development and advancement of the intelligent era, Generative Large Language Model (LLM) Agents have become the cutting-edge of artificial intelligence technology development. They not only provide core impetus for the digital transformation of society but also pose complex ethical risks to modern society. Starting from the dual dimensions of philosophy of technology and ethical theory, this paper systematically traces the theoretical origins of the ethical risks of Generative LLM Agents, expounds on their inherent technical limitations, dilemmas in responsibility ethics theory, value alignment challenges, and socially constructive risks. Furthermore, it proposes strategies for constructing a four-dimensional normative system: innovation in governance concepts, establishment of technical standards, improvement of legal rules, and cultivation of ethical literacy. This aims to build a multi-level, multi-stakeholder collaborative governance framework centered on "trustworthy governance," thereby providing theoretical support and path references for the ethical governance of generative artificial intelligence.

Keywords: Generative Large Language Model Agents; ethical risks; value alignment; trustworthy governance; normative system

引言

生成式人工智能技术正在从专用型工具向通用型 Agent 不断演进，尤其在 ChatGPT、Sora、Deepseek 等大模型软件与平台支持下，生成式大模型 Agent 成为重塑知识生产、商业服务与社会交互基本范式的关键要素。该类系统主要以大规模参数训练与自回归生成机制为依托，可以实现在概率空间中构造出语义上相容的新内容。但此类内容生成能力跨越了时代发展，展现出前所未有的伦理复杂性，从而推动生成式大模型 Agent 成为当前社会上技术伦理治理的焦点场域。

一、生成式大模型 Agent 的伦理风险理论溯源

（一）技术内在局限性的理论渊源

技术架构的内在局限是生成式大模型 Agent 最核心的伦理风险。从技术哲学视角来看，技术架构局限性主要体现在算法黑箱、数据依赖与奖励机制缺陷三个层面。算法黑箱是指生成式大模型依托深层神经网络完成决策，其过程有着高度不可解释性，这就导致无法对模型行为的追溯监督，也为算法偏见、算法歧视的出现提供了环境[1]。数据依赖是指生成式大模型对海量语料数据的极度依赖，训练数据的质量则会直接影响模型输出的伦理偏

差。现有大模型主要以互联网数据为载体，但互联网环境本身内嵌着人类社会中各类偏见与不平等，而大模型在数据统计规律分析中可能会再现或放大其中的偏见与不平等。奖励机制缺陷是指生成式大模型 Agent 无法建立价值对齐体系。强化学习模型可以通过人类反馈一定程度上输出符合人类偏好的内容 [2]，但其本质上是依托外部支持的调优逻辑，是追求最大化奖励分数得出的结果，并非真正的人类价值规范。

（二）责任伦理理论的分析框架

生成式大模型 Agent 对传统责任伦理理论也有重大冲突。在责任伦理理论下，人工智能体在社会系统中可以视为类似律师、股票经纪人等角色，属于一类有限道德代理者，不具有初级利益，仅具有次级利益，并且需要根据角色责任理论承担代理责任。但生成式大模型 Agent 有着较高的自主性与适应性，从而影响了责任归属方式。从责任主体视角来看，生成式大模型 Agent 与开发者、服务提供者、使用者、监管者等均有着直接关联，导致形成责任分散困境，甚至引发责任真空 [3]。因此基于责任伦理理论，生成式大模型 Agent 需要建立关系型责任新型伦理框架，以此超越因果责任观的限制，将对人类社会负责并被人类所期望和赋值的责任形式转嫁到人工智能体之上，以此突出他律性特征，达到“趋善避恶”的预期性道德责任效果。

（三）价值对齐的哲学困境与伦理偏差

价值对齐指人工智能体或大模型在执行任务与内容生成时，其目标指向、行为倾向及输出结果应与符合人类社会广泛认可的价值体系。但在人工智能技术实际应用过程中，却面临着价值对齐的哲学困境与伦理偏差。首先是价值定义困境。人类社会的价值体系具有多元性、动态性与情境性特征，并不能建立明确固定的规则进行定义。例如“公平”的价值原则在不同情境与文化视域下有着不同的含义，这是生成式大模型 Agent 无法把握和判断的困境之一 [4]。其次是价值学习困境。价值对齐的学习过程无非自上而下与自下而上两种途径，前者是依托人类专家对价值内容的定义所构建出的目标函数或规则结构；后者是以人类行为反馈为基础归纳总结的对齐信号。前者具有难以捕捉价值的语境敏感性，后者则无法保证价值对齐与伦理要求的一致性 [5]。最后是价值冲突协调困境，即在不同文化、不同群体之间产生价值冲突时，生成式大模型 Agent 缺少可以辅助价值权衡与协调的机制，甚至导致输出结果展现出结构性不公特征。

二、生成式大模型 Agent 的伦理风险规范体系构建

（一）基于可信治理的治理理念创新

命令控制型监管模式在处理生成式大模型 Agent 的伦理风险时展现出难以应对的诸多缺陷，因此在伦理风险规范体系构建中，应优先引入可信治理范式，以此为生成式大模型技术提供可控性、可问责性、公平性、可靠性、可解释性和安全性等层面的保障，实现技术创新与伦理约束之间的动态平衡。

从宏观层面来看，可信治理范式的构建需要以责任伦理为调适性理论依据，通过分析动机、行为、后果反思行动应对中呈现出的速度、规模、交互与调试问题，从而弥合原则与行动之间的差池。从微观层面来看，可信治理范式建设还需要建立多主体协同的治理架构。一方面需要政府、企业、行业协会、科研机构与公众等共同担任责任主体，以此打造协同共治的网络化结构 [6]。另一方面需要发挥各个主体的优势与特征，比如企业可以建立算法伦理委员会与道德责任官进行内嵌化管理；行业协会可以鼓励头部企业发布伦理实践报告，建立行业示范标准；政府部门可以建立跨部门监管与协调机制，统筹推进法规制定与执行落地。

（二）技术标准与架构的规范构建

生成式大模型 Agent 伦理风险治理中，技术标准与架构的规范构建属于基础性工程，其核心体现在数据治理、算法透明与价值对齐三个维度之上。

第一，在数据治理层面，应推动数据要素确权立法，通过法律明确数据所有权、使用权和交易权边界，从而保障用户对数据具有“知情—授权—撤回—追溯”的完整权利链条 [7]。同时，也要推动第三方建设公共训练语料库，为生成式大模型 Agent 提供多样、可信、经过审核的语料资源，提升数据伦理质量。

第二，在算法透明层面，应建立平台披露算法运行机制，或者提供可解释性披露方案，由此通过信息来源标注等方式，提升生成式人工智能运行的透明度与用户的感知能力 [8]。同时，针对人工智能技术的隐蔽性与不可预测特性，可以采用技术监管与敏捷治理的协同方法，将 AIGC 来源与后续使用行为纳入监管体系，建立生成记录保存、实名制、强制标记、周期性数据更新等规则系统 [9]。

第三，在价值对齐层面，应构建“理由空间”与“元级机制”，赋予大模型在冲突情境中进行权衡，并具备动态修正目标的能力 [10]。此外，企业在算法目标设计时应当建立公平性、多样化等更多指标体系，以此削弱单一商业导向的价值逻辑，形成价值均衡的 AICG 应用逻辑。

（三）法律规则与监管体系的完善

生成式大模型 Agent 伦理风险治理体系必须建立在法律规则与监管体系的刚性约束之上。我国目前已经建立了以《生成式人工智能服务管理暂行办法》为核心的治理框架，但在规则细化与体系协调方面还有待提高与完善。

第一，在监管规则设计层面，应针对技术监管、质量标准、无歧视、分类分级等规则主体建立 AIGC 监管体系 [11]。一方面要打造分类分级规则，将风险程度进行梯度设计与分别监管。另一方面要设定服务主体对不可接受风险和重大风险的申报与备案义务，以此落实安全评估 [12]。

第二，在责任分配机制层面，应针对人工智能生成内容建立“可推定责任”原则，即人工智能平台无法证明无过错时，需要承担相应的法律责任，以此规避企业通过“算法自动生成”之名逃脱法律责任与义务的方式 [13]。此外还应建立事前预防、事中监管与事后问责的三位一体治理体系，并明确开发者、提供者与使用者等不同主体的责任边界。

第三，在法律规范的发展层面，应从现有法律法规的法条修订与细节解释、新的通用性人工智能法规出台、颁布生成式人工

智能领域专门性法规等方式进行持续规范和优化[14]。

（四）伦理素养与公众参与的培育

生成式大模型 Agent 的伦理治理不仅需要技术与法律层面的硬性约束，同时也需要伦理素养与公众参与的柔性支撑。

第一，在教育体系层面，应将 AI 伦理与算法素养教育纳入中小学与高校课程体系，着重培养学生的技术批判意识与伦理反思能力。

第二，在公众参与层面，应鼓励新媒体、行业协会与公益组织等社会力量参与 AI 伦理治理活动中，推动民间监督常态化[15]。

第三，在伦理文化发展层面，应推动以“科技向善”为核心价值导向的伦理文化，并引导社会与群众对生成式人工智能应用的伦理边界进行反思，不断提出新的伦理风险认知与防范策略。

三、结语

综上所述，生成式大模型 Agent 的伦理治理体系构建涉及技术、法律、伦理以及社会等多层次内容，本文从理论溯源与规范构建两个维度，提出了生成式大模型 Agent 的伦理风险的理论依据与应对策略，进而构建了以“治理理念创新、技术标准构建、法律规则完善与伦理素养培育”为主旨的“可信治理”体系，达到了多元共治、敏捷协同与责任平衡的效果与目的。

参考文献

[1] 李维阳，苏静普 . 教育改革视域下生成式预训练模型的伦理风险与治理策略 [J]. 特区经济，2024,(08):60-63.

[2] 黄伟文，李昕宇 . 从规制到治理：论生成式人工智能法律管理的新范式 [J]. 乐山师范学院学报，2024,39(06):94-102.

[3] 翟荆瑞 . 生成式人工智能引发的异化现象和伦理风险研究 [D]. 西北师范大学，2024.

[4] 林莉 . 论生成式人工智能服务中的教育风险及其应对——以 ChatGPT 为例 [J]. 电脑知识与技术，2024,20(11):33-35.

[5] 鄢敏 . 生成式人工智能的伦理风险探析——以 ChatGPT 为例 [J]. 信息与电脑 (理论版),2024,36(07):140-142.

[6] 罗蓉蓉，肖攀诚 . 生成式人工智能的风险审视与治理研究 [J]. 咨询与决策，2024,4(01):1-18.

[7] 范佳丽 . 生成式人工智能风险预防探究 [J]. 合作经济与科技，2024,(10):190-192.

[8] 孙那，鲍一鸣 . 生成式人工智能的科技安全风险与防范 [J]. 陕西师范大学学报 (哲学社会科学版),2024,53(01):108-121.

[9] 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,(06):58-66+75.

[10] 邹疃，张景玥 . 生成式人工智能技术的伦理风险防治 [J]. 湖北第二师范学院学报，2023,40(11):58-63.

[11] 吴南中，陈咸彰，冯永 . 从 " 失序 " 到 " 有序 ": 生成式人工智能教育应用的转向及其生成机制 [J]. 远程教育杂志，2023,41(06):42-51.

[12] 刘佳丽，廖怀高 . 论生成式人工智能学术伦理风险规制——以 ChatGPT 为例 [J]. 沈阳工程学院学报 (社会科学版),2023,19(04):83-89+123.

[13] 吴育珊，杜昕 . 生成式人工智能的安全风险与法律规制 [J]. 岭南学刊，2023,(05):105-112.

[14] 杜昕 . 生成式人工智能的法律应用与风险防范 [J]. 司法警官职业教育研究，2023,4(02):53-62.

[15] 段伟文 . 准确研判生成式人工智能的社会伦理风险 [J]. 中国党政干部论坛，2023,(04):76-77.

