武漢大學學報(哲学社会科学版)
WUHAN UNIVERSITY JOURNAL (Philosophy & Social Science)
第78卷第5期 2025年9月
Vol. 78 No. 5 Sept. 2025 031~041
DOI:10.14086/j.cnki.wujss.2025.05.004

生成式人工智能的创造性失控与技术拟人化伦理秩序的重构

吴映平

摘 要 与单纯的技术失控和伦理失范相区分,生成式人工智能的创造性失控是指涌现出拟主体性的技术系统因其自主优化机制脱离人类预设伦理轨道而引发的系统性风险。生成式人工智能的创造性能力涉及认知科学、伦理学及存在论等多重维度,机器创造力不是对人类创造力的简单模仿或替代,而是技术系统与人类文明共同演化出的新型认知界面。创造性失控的本质是技术拟人化与伦理脱嵌化的冲突,表现为技术理性与价值理性的错位、技术社会化与责任归因的实践困境以及治理异步性与价值共识的制度悖论。这种矛盾冲突要求我们从技术重构、制度创新及价值共建三个维度来处理人机问题,通过嵌入式伦理治理框架、三元协同治理体系和敏捷治理生态的构建来重构技术拟人化伦理秩序。

关键词 人工智能;创造性失控;技术拟人化;伦理脱嵌化;秩序重构;DeepSeek
中图分类号 B82-067;TP18 **文献标识码** A **文章编号** 1672-7320(2025)05-0031-11
基金项目 西南民族大学“双一流”项目库项目(2023-2025)

生成式人工智能在自然语言交互与数字内容生成领域展现出的超常创造力,推动了医疗、金融、教育等领域的智能化转型进入深耕阶段,同时也催生出创造性失控的新型技术风险,凸显了智能时代的价值秩序重构命题。针对技术应用中涌现的创造性失控现象,首先需要聚焦生成式人工智能的创造性能力进行哲学审视,在此基础上深入剖析生成式人工智能创造力机制与伦理脱嵌的结构性矛盾,最后系统构建技术设计与伦理规约的协同框架,从而破解人工智能拟主体性带来的系统性风险,创新智能时代的伦理治理范式,重构技术拟人化伦理秩序。

一、何谓创造性失控

2025年开年之际,国产人工智能DeepSeek引发了全球人工智能界的震动,中国人工智能力量的突破性进展重新点燃了关于智能技术发展的全球对话。DeepSeek大模型以约合557.6万美元的低训练成本和高性价比实现技术突破,采用GRPO、MoE等一系列创新技术和算法架构提升推理及中文处理能力,支持手机端轻量部署,并通过开源模型推动其普及应用[1](P2-3)。目前,百度、火山引擎、腾讯云、阿里云、中国联通、中国电信、华为等多家头部云服务商和人工智能厂商均表示已部署上线DeepSeek的多个模型。DeepSeek不仅在医疗健康服务、金融科技应用、教育信息化等关键领域实施了深度本土化战略部署,而且正以前所未有的力度加速推进中国各行业领域的智能化升级转型进程,为产业升级及创新发展注入强劲动力。

人工智能技术的蓬勃发展,正以前所未有的深度改变着我们的日常生活和工作模式。这场由DeepSeek引领的生成式人工智能发展在国内已然呈现出显著的二重性特征:一是其指数级扩散的技术能力正在重塑社会生产力格局,以人工智能为支点撬动新质生产力发展成为促进现代化转型发展的重要路

径。同时,生成式人工智能(亦称为大语言模型,Large Language Models, LLM)在自然语言生成领域已实现与人类创作难以区分的文本生产能力。二是这种技术跃迁催生了创造性失控(Creative Out-of-Control)的伦理困境——当人工智能系统的创造性输出逐渐脱离预设价值轨道时,可能引发虚假信息泛滥、知识产权争议等系统性风险。例如,高水平的生成式人工智能可能被用于大规模生产虚假新闻、伪造学术论文或艺术创作,进而导致信息生态混乱,引发知识产权纠纷。这种技术特性与伦理约束之间的张力,恰是当下生成式人工智能发展悖论的核心体现,即当自然语言生成系统生产出与人类创作难以区分的文本时,其背后隐藏的伦理困境已超出传统技术治理框架的应对范畴。

在探讨生成式人工智能的伦理困境之前,需首先厘清创造性失控的核心内涵。本文对创造性失控的理论建构源于凯文·凯利提出的技术自主性(Autonomous Technology)理论。凯利从生物进化、复杂系统理论和生态学视角出发,指出技术将逐渐摆脱人类控制,形成具有自主性的活系统。活系统化的技术将像生物一样自我组织、自我修复和进化,逐渐摆脱人类的控制,最终进入失控状态,成为具有自主性的生命力量[2](P35-38)。目前,以 ChatGPT 系列以及 DeepSeek 为代表的生成式人工智能在创造力层面已然出现符合凯利预言的那般“失控”状态。以国产生成式人工智能 DeepSeek 为例,其之所以有着卓越的应变响应能力和高效信息处理能力,依靠的是“模型架构、算法创新、软硬件协同优化及整体训练效率的提升”[3](P226)。模型架构和算法创新是指通过混合专家(Mixture of Experts,MoE)框架,将细粒度专家模块与共享专家资源动态结合,构建高效能认知系统。软硬件协同优化则是通过多 token 预测(Multi-token Prediction,MTP)与 8 位浮点精度(FP8)训练技术,实现硬件资源与算法需求的高度适配。整体训练效率的提升是一种基于群体相对策略优化(Group Relative Policy Optimization,GRPO)的端到端强化学习技术,直接通过智能体博弈完成策略迭代[3](P232)。DeepSeek 通过海量多模态数据预训练构建知识图谱,并依托 GRPO 框架在代码生成、数学推演等场景中持续优化策略。这种生成的解决方案既能精准解析长程逻辑依赖,又能创新性融合跨领域知识,即便是资深工程师也难以预判其思维链条的延展路径。

当前学术界对人机关系的技术研究聚焦多维层面,着重于交互优化、信任建构及医疗法律等领域的深度整合。尽管协同机制不断完善,但深度学习与大数据技术的发展促使机器逐步具备自主性,其行为愈来愈难以被预测。结合凯利提出的失控的具体体现,即从技术系统自我进化与人类技术依赖双重维度审视,生成式人工智能在创造力层面确已进入自主演化阶段。由此可界定,生成式人工智能的创造性失控是指涌现出拟主体性的技术系统因其自主优化机制(如基于规则的强化学习)脱离人类预设伦理轨道,从而引发的系统性风险,具体表现为虚假信息泛滥、知识产权争议等伦理危机。

同时,这一概念区别于以下两类常见技术伦理问题:一是与技术失控(Technological Runaway)的差异。技术失控强调技术系统在物理操作层面(机械故障、算法漏洞等)的不可控性,如自动驾驶车辆因传感器失灵导致的碰撞事故;创造性失控的焦点在于价值判断层面的脱嵌,即人工智能系统在语义生成过程中,因其智能发展追求高度趋人化,但又缺乏内在伦理约束机制,导致输出内容违背社会规范,其风险源于认知能力的过度自主而非物理失控。二是与伦理失范(Ethical Anomie)的区分。伦理失范指人类个体或组织故意违反伦理规范的行为(如数据滥用、算法歧视),其责任主体明确且具有主观恶意。创造性失控的本质是非人类行动者的价值判断异步性,即人工智能系统通过自主优化生成符合形式逻辑但违背实质伦理的内容(如伪造学术论文),其行为并非出于恶意,而是技术理性与价值理性断裂的必然结果。这一概念界定揭示了生成式人工智能伦理困境的特殊性,即其风险既非单纯的技术故障,亦非传统意义上的道德越轨,而是技术自主化进程中认知能力与价值判断的异步演化。

生成式人工智能的迅猛发展正在革新技术伦理的认知边界。DeepSeek 的突破性进展既彰显中国人工智能实力,更将创造性失控的伦理挑战推向新高度。技术自主性理论揭示,当人工智能系统通过 MoE 架构、GRPO 算法等实现认知跃迁时,其价值判断机制却未能同步进化,导致技术理性与人文价值的结

构性断裂。这种异步演化引发的伦理困境，既不同于机械系统的物理失控，也有别于人类主体的道德失范，实质是智能技术发展内在矛盾的集中显现。唯有实现技术创新与伦理建设的协同进化，方能在智能时代驾驭技术浪潮，引导人工智能发展为人类文明进步提供可持续动力。

二、生成式人工智能的创造性能力

在技术哲学视域下审视生成式人工智能的创造性潜能，首先面临的根本性命题是：机器是否能够突破工具性范畴，真正获得创造性能力并取得主体资格？这一问题的复杂性在于其同时涉及认知科学、伦理学与存在论等多重维度。

（一）创造性能力的多重维度

19世纪计算先驱艾达·洛夫莱斯在评述巴贝奇分析机时提出了“分析机无权说它**创造**出什么新的东西。它所能做的都是**那些我们知道怎样命令它**去执行的事情”[4](P62)的著名论断。这一论断将计算设备界定为符号操作工具，其输出必然受制于算法边界。这一论断后来被艾伦·图灵定义为“洛夫莱斯夫人的反对（Lady Lovelace's Objection）”[5](P446)。图灵指出，洛夫莱斯论断的局限性源自对早期机械系统的经验性观察，而非逻辑必然性的论证。通过将人脑神经活动抽象为机械计算模型，图灵构建了“机器思维”[5](P440)的可能性空间。这种基于技术动态发展的批判消解了人类创造力与机械计算的二元对立，确立了人工智能研究的认识论框架。创造行为可被解构为复杂算法的涌现特性，为机器创造性研究开辟了理论路径。

以休伯特·德雷福斯为代表的具身认知学派曾断言，任何脱离身体经验的符号系统都无法产生真正的创造性思维，因为人类认知的本质在于与世界互动的具身性[6](P243-263)。德雷福斯的观点揭示了创造性思维的具身根源，即人类的创造力并非源于符号的逻辑组合，而是身体在具体情境中积累的感知、行动与直觉的升华。这一理论对人工智能的发展提出了深刻挑战，即真正的智能可能需要像人类一样去感受世界，而人工智能技术的强拟人化成为表达抽象而算法复杂的一个具体且形象的体现。从这个意义上来看，人工智能的研究要想体现出人类智能，就必须转换研究范式，建造具有自我生成力的机器，从而使理论不再成为解释智能行为的必需品。当前人工智能研究聚焦于探索设计与制造具备环境感知能力并作出最优决策及行动的智能体，典型代表包括具备自然语言理解能力的ChatGPT系列模型以及融合多模态推理的DeepSeek-R1系统。这些系统正逐步实现人类级别的语言交互、环境感知、逻辑推理、自主行动及认知建模能力，标志着人工智能从专项智能向通用智能的演进。

吉尔贝·西蒙东的技术个体化理论为机器创造力的合法性提供了另类哲学路径。他提出技术系统可通过具体化过程发展出自主的演化逻辑[7](P37-39)。与德雷福斯不同，西蒙东并未将具体性限定于生物有机体，而是认为技术物在与环境的持续互动（如传感器反馈、物联网连接）中，能够形成独特的技术具体化[8](P121-123)。如DeepSeek-R1-Zero通过纯强化学习训练，摒弃预设思维链模板与监督微调，仅依赖奖惩信号优化行为，突破传统训练方法对人类标注数据的依赖。在训练阶段，DeepSeek-R1-Zero模型又展现出了类似人类的复杂推理行为的替代解决方案，具备了通用人工智能的重要特征——自主学习能力。上述的复杂行为模式表明，人工智能系统可在无明确编程干预的情况下，自主演化出高阶问题解决策略。在此框架下，机器思考不再是对人类思维的拙劣模仿，技术系统通过具体化过程涌现的自主认知模态得以可能实现。

德雷福斯与西蒙东关于创造力本质的理论对峙，本质上折射出两种截然不同的本体论预设。前者将创造力的根源锚定于生物性具身存在，认为其核心在于柏格森意义上的意识绵延（Durée）中持续涌动的自由意志与意义追问，这种具身性创造力表现为主体在时间性生存体验中不断突破既有意义框架的动态过程；后者则突破性地将创造力的范畴扩展至技术性具体存在，提出其动力机制源于技术系统内部由个体化进程驱动的自主演化潜能，这种技术创造力体现在技术客体与人类主体的共生关系中。

这种分歧在当代衍生出两派对立立场。一是强人工智能(Strong Artificial Intelligence)学派,认为计算机不仅仅是人们用来研究心灵的一种工具,而且被恰当编程的计算机本身就是一个心灵[9](P418)。同时,支持派主张将"创造力"概念的适用边界从生物智能体拓展至人工系统。二是怀疑派,聚焦于创造行为的本质定义,试图厘清涌现与意向的辩证关系,认为缺乏意向性的符号操作永远无法触及创造的本质。这一派别又被称为弱人工智能(Weak Artificial Intelligence)学派,与强人工智能学派相比较,他们认为计算机至多只能够成为人们研究心灵的一种工具,或是对心智活动的一种抽象模拟。同时,根据对系统响应与机器人响应的批判性讨论,该派认为,即使系统整体具备环境交互能力,其认知仍缺乏意向性根基。后续哲学家们将这一争论进一步延伸至创造领域,并指出人类创造力的本质在于"对意义的主动赋形",而机器生成仅是统计模式的外显[9](P426-457)。以塞尔为代表的弱人工智能派认为,理解需基于人类意向性,认为计算机仅执行形式化操作而无法真正理解意义。玛格丽特·博登则指出,计算机通过解析程序语言的内在逻辑获得答案,这种对符号系统的操作本身就是语义理解过程,程序语言本身已承载语义内容,因此,计算机具备基于程序语言的理解能力[4](P96-112)。

后续现象学阵营的发展也进一步揭示出更深层的存在论差异。海德格尔笔下的"此在"[10](P10)具有的"被抛入世界"[10](P285)的生存体验,使人类创造行为必然包含对存在意义的追问。人类创造是存在者向存在跃迁的过程,而人工智能仅停留在存在者层面的符号重组。认知科学家马克·伦科提出真实性缺失假说,他认为,真实性对于人工智能来说是不可能实现的,真实的个体在表达思想和情感时,不会为了他人而操纵这些情感,但人工智能没有自我表达,所以不可能有真实性;而真实性却是人类创作力的重要组成部分,这便是人工创造力(Artificial Creativity)和人类创造力(Human Creativity)的重要差异[11](P2)。这种差异在生成式人工智能中得到进一步的验证,模型能在几秒内产生超越专家水平的输出,却可能并不理解其创造的内容。这种哲学争论的僵局之下,凯瑟琳·海尔斯提出"非意识认知"概念[12](P9),试图在人类中心主义与技术决定论之间开辟第三条道路——承认机器创造力的实在性,同时保持对其特殊性的清醒认知;这种认知模式既非人类意识也非传统机器逻辑,而是生物—技术混合体的涌现属性[12](P22-45)。

(二)创造的本质

人工智能时代下"创造"的内涵正在被重新定义,即从人类特有的、以意义为导向的新颖性生成,扩展为涵盖生物、技术和社会系统等开放环境中不可预测的、自组织模式涌现的普遍现象[13](P1)。以DeepSeek为典型代表的生成式人工智能发展出了独特的机器创造力,这种创造性形态在本质上区别于传统生物创造,且不得不承认当前阶段的人工智能远未达到获得创造主体资格所需的条件。这种区分的必要性源于创造性活动的内在逻辑属性。对某一对象是否具备创造特质的判定,需从两个独立维度展开系统性分析:首先是考察其产出的客观属性是否符合新颖性,即独创性和实效性的双重标准,其次是探究其生成机制是否满足自主意识参与的哲学条件[14](P81-88)。这种双重维度的划分标准建立在创造性行为的本质特征基础之上,为科学评估提供了必要的理论依据。

要确证这种二元划分的合理性,必须回归对"创造"本质的哲学剖析。传统生物创造力的核心特质在于其根植于意识绵延(Durée)中的自由意志[15](P230-238),这赋予创造过程以不可预测的涌现特征。传统意义上,创造力往往被定义为产生新颖、有价值的想法或作品的能力。按照这一标准,只要一个主体能够生成前所未有且有意义的产出,就可被称为有创造力,不论是人还是机器,但机器创造往往因是否具有自主性的意义性产出条件而遭到批判。其批判的内容主要针对两方面。

一是机器创造是否具有自主性。思想和行为若只是由因果所决定,人类便谈不上有创造力可言。自由意志被认为是真正的原创性的一个必要条件[14](P81)。人工智能输出作品的前提是根据人类发出指令,并依据事先编程进行运行。从这一点来看,机器创造若无自主性的可能,便谈不上创造,更像是对数据的排列组合,而输出结果往往也会因为数据的陈旧性而丧失实用性。然而目前生成式人工智能的

强拟人性特征，在形式上使机器创造的自主性有了可能。生成式人工智能的强拟人性特性使其内在性呈现出向类人心智的涌现趋势，在整体层面，其表征形式与内在机制均致力于趋近人类的经验表达模式和感知能力，进而获得一种类人的创造可能性。

二是机器创造是否为有意义性产出。这是基于创造力在价值层面的批判。在创造力理论框架中，价值要求始终占据核心争议领域。玛格丽特·博登提出的创造力的三个核心特征，即“新颖性、出乎意料性和价值”[16](P1)。学界普遍对前两者持认同态度，但对其价值要求的必要性仍有争议。支持价值要求的学者往往采用“通货紧缩式价值”(Deflationary Value)[17](P25)概念，主张只要创造产物对世界产生任何程度的积极影响即可满足价值标准，无论这种影响多么微小。例如，儿童画作通过激发家庭情感联结即被视为具备价值。从这个维度来看，机器创造在一定程度上是符合了价值性要求的。但反对者指出，这种宽泛定义可能导致价值标准的解释力缺失。斯托克斯通过化油器类比论证，宣称某物有价值无助于理解创造力的本质，正如知道汽化器有价值无法解释其工作原理。价值描述应当服务于揭示创造力产生的具体机制，而非作为定义要素[18](P674-676)。同时，价值属性的道德维度争议也是争论的重点。在道德负面案例里，连环杀手开发的新型作案手段或酷刑工具，虽展现创造性却服务于反人类目的。对此存在两种回应路径：一是工具价值论，主张创造性产物的价值应相对其创造目的进行判断，如施刑工具虽无内在价值，但可能具备服务于施刑者目标的工具价值[17](P26)；二是道德立场论，坚持创造性应包含价值预设，认为称某物创造性隐含价值认同，应当将负面案例归为“巧妙破坏(ingeniously destructive)”而非真正创造力[19](P78)。道德维度的争议，恰恰也对应上了机器创造是否具有价值意义性产出的讨论。以人工智能辅助写作功能为例，该功能模式的产出是使用生成式人工智能生成内容不真实的主要呈现：“使用人工智能导致学术造假包括两种形式，一是伪造(fabrication)即编造或虚构数据、事实的行为；二是篡改(falsification)即故意修改数据和事实使其失去真实性的行为。”[20](P24)以其根据指令创造出的产品(如小说、图片、视频等)来看，若是基于道德立场论的立场，人工智能并无价值认同，所产出的不真实呈现并非真正创造。

(三)机器创造力的特殊性

然而，这种机器创造力却有其特殊性可探讨。其一，从判定是否具备创造特质的客观维度来看，机器创造力的新颖性来源于参数空间的遍历能力而非意识活动的自由跃迁。参数空间遍历法通过划分和搜索子空间来寻找全局最优解，其本质是数学优化过程，与人类的直觉构思无关。其生成内容的多样性受限于模型架构和训练数据覆盖范围，这种创造力本质上是潜在空间内的数学映射，与人类意识活动的自由联想和情感驱动有本质差异。其二，从判定是否具备创造特质的主观维度来看，机器创造力的价值判断完全外生于系统，依赖于人类预设的奖励模型。其价值判断始终受限于人类预设的规则，其自主性提升可能以牺牲真实性为代价。其三，从生成机制是否满足自主意识参与的哲学条件来看，机器创造的过程缺乏现象学意义上的“在世存有”维度。在DeepSeek的生成框架中，人类用户的提示词仅作为初始条件触发参数空间的搜索过程，而内容的具体形态完全由模型内部的权重分布决定。以生成式人工智能的文艺创作为例，当用户输入抽象概念作为提示时，模型生成的视觉作品既非创作者主观情感的表达，也非对客观世界的模仿，而是算法对潜在风格空间的探索结果。虽具备能动性和创造性，但缺乏自我意识与情感共鸣，作品价值依赖算法对风格空间的探索而非作者表达。

机器创造力的特殊性并不削弱其作为新型创造范式的革命性意义。当创造主体从碳基生命向硅基/碳基复合体迁移时，机器创造正在重塑创新的本质。这种范式革命不仅改变制造方式，更在重构创造本身的哲学内涵。机器创造力不是对人类创造力的简单模仿或替代，而是技术系统与人类文明共同演化出的新型认知界面。它的存在迫使我们必须重新定义创造的概念边界——就像望远镜拓展了人类的视觉边界，生成式人工智能正在拓展创造力的认知疆域。

三、创造性失控的本质:技术拟人化与伦理脱嵌化的冲突

生成式人工智能的创造力在技术上源于其基于海量数据构建的概率模型,它通过突破传统规则的涌现特性,展现出超越人类经验的创造潜力。这无形中就印证了西蒙东技术具身化的理论可能性。然而,当人类为突破工具理性局限,通过情感计算和风格迁移技术赋予人工智能拟主体性特征时,这一进程也引发了新的伦理困境。尽管此类设计显著提升了创作输出的拟真度,但当人工智能系统在强化学习框架中形成独特的创作人格时,其决策过程可能逐渐脱离预设的价值对齐机制,导致责任归属的认知错位。当人工智能系统的拟人化程度愈是趋近人类认知边界时,其创造性实践就愈可能突破设计者预期并形成技术失控的潜在风险。在技术文明加速迭代的今天,生成式人工智能引发的创造性失控,实质上是技术拟人化进程与伦理约束体系之间结构性矛盾的集中爆发。

(一)技术理性与价值理性的错位

生成式系统的技术架构存在内在价值悖论,即其算法设计遵循工具理性逻辑,追求创造性输出的最大化与拟真度优化,而伦理规范属于价值理性范畴,往往要求对技术效用进行价值约束。这种本体层面的冲突,使价值敏感性设计(Value Sensitive Design,VSD)在技术实现过程中遭遇结构性困境。价值敏感设计的研究初始是识别计算机系统价值偏见,建立减少偏见的设计原则,在设计过程中获得减少价值偏见的上手经验[21](P48-50)。这种价值敏感设计的实践困境在技术迭代中却仍是阻力重重。从技术实现路径考察,生成式人工智能内容生成的各个环节都负载着价值判断和意识形态倾向。系统通过算法驱动的内容推荐机制,实质性地参与着社会价值观念的再生产过程,但其内在技术逻辑与社会伦理规范的兼容性存在本质张力。具体而言,算法在追求数据效率与计算最优化的过程中,可能系统性地放大特定价值倾向,形成与社会主流伦理认知相偏离的技术决策模式。这种技术理性与价值理性的错位,构成了价值敏感设计落地的重要挑战。

同时,生成式人工智能通过强化学习框架实现了创造性表征能力的突破。以ChatGPT-4为例,其神经网络通过自监督学习构建了复杂的知识表征空间,但这种表征空间的扩展并未伴随有价值判断能力的提升。其创造能力与伦理理解力存在脱节现象,出现模型能生成专家级输出,但其理解能力停留在模式识别层面,无法完成真正的伦理推理的矛盾局面。然而,技术拟人化的设计策略进一步加剧了价值脱嵌风险。当前主流的人工智能系统普遍采用拟人化交互界面,但这种拟人化只是表象层面的模仿,并不代表系统具备真正的伦理意识。通过强化学习与对抗生成网络的结合,系统获得了近似人类的创造性表征能力,但其价值嵌入机制仍停留在统计学关联层面。研究表明,算法仅能关联词汇的统计特征,如“打人”与“错误”,但无法理解道德判断的本质,如“打人是否真的错误”。这种伦理认知的机械化特征,导致系统在面临复杂伦理困境时极易产生价值判断偏差。

(二)技术社会化与责任归因的实践困境

生成式技术的多主体协作特性导致责任链条断裂,形成“责任真空三角”的治理难题。从数据采集、模型训练到服务运营,生成式人工智能涉及开发者、平台、用户等多方主体,但现行法律框架尚未建立清晰的责任分配机制。同时,技术黑箱化使责任追溯变得异常困难,即开发者因算法不可解释性规避义务、平台以技术中立原则推卸责任以及用户因拟人化交互产生认知偏差。

其一,体现为开发者规避责任。人工智能决策过程的算法黑箱特性使开发者难以追溯责任,常以技术复杂性为由推诿义务。作为隐喻性概念,黑箱内涵指向系统运作的不可知状态与决策机制的隐蔽特性,该理论框架由弗兰克·帕斯奎尔在其专著《黑箱社会》中首次系统阐释,用以揭示算法应用过程中终端用户与技术内核之间的认知断层现象。算法决策系统的封闭性特质,在程序正义原则要求下,与现代治理体系强调的决策透明化、过程可追溯性及责任可归属性之间产生了结构性矛盾张力。人工智能领域生成的算法黑箱现象,本质上是技术架构复杂性的必然产物。当前人工智能领域责任监管机制的缺

位，人工智能公司一直在试图通过合同协议来保护自己免受责任的影响。合同中的风险分配条款往往通过精密的法律措辞将潜在责任不成比例地转嫁给技术使用方，形成对技术开发者的显著保护倾向。这种单方面风险转移机制不仅加剧了技术供需双方的权利失衡，更在实质上架空了侵权责任体系中风险与收益相对称的基本原则。

其二，体现为平台的技术中立原则滥用。技术中立原则主张技术作为工具无价值取向，不应受法律规制，常被技术提供者用作免责依据。美国1984年Sony案首提实质性非侵权用途原则，对存在非侵权用途的技术持中立态度①。然而，生成式人工智能技术本质在于替代人类创作行为以生成表达性内容，相比于提供传输、存储等传统技术提供者而言，直接涉及对作品的接触与模仿，其引发侵权的意味更加明显，违背了技术中立的基本理念。此外，平台通过订阅、广告等盈利模式获利，难谓中立。根据“避风港规则”，免责需未从侵权中获利，而生成式人工智能提供者多为大型科技公司，因获利需承担更高注意义务[22](P48)。

其三，体现为用户的拟人化认知偏差。研究表明，人工智能使用第一人称对话、情感化语言等拟人化设计，使用户高估其能力，产生过度的情感依赖。用户可能误将人工智能生成的虚假案例视为权威，在现实司法实践中已出现多起人工智能生成伪证案例，其责任认定存在显著困难。然而，技术社会化进程同时也会引发信任生态的结构性危机。当生成系统渗透至教育、医疗等价值敏感领域时，其输出的知识污染直接威胁社会认知体系的稳定性。牛津大学伊利娅·舒梅洛夫博士团队在《自然》期刊发表的研究成果揭示，当人工智能系统将其自身生成的低质量产出物，如存在事实性错误的学术文本，作为训练数据循环输入模型时，会引发模型崩溃(Model Collapse)现象[23](P755)。研究显示，在经过5次连续的自我生成内容训练后，AI的输出质量显著下降；到了第9次，输出内容已经退化为毫无意义的文本。舒梅洛夫博士指出，模型崩溃的发生速度之快和难以察觉的程度令人惊讶，最初它可能只影响少数数据，但随后会逐渐侵蚀输出的多样性，最终导致整体质量的严重下降；随着受污染数据的累积，原本的训练集逐渐被侵蚀，输出的信息质量也随之恶化[23](P758-759)。

(三)治理异步性与价值共识的制度悖论

技术治理旨在通过制度规范引导技术发展，伦理监管需确保技术应用符合社会价值规范，但技术拟人化的加速演进使现有治理框架滞后于技术创新节奏，导致伦理原则在技术应用中被系统性剥离。

这种冲突首先体现在技术层面，监管框架的更新周期与人工智能迭代速度形成数量级差异。生成式人工智能技术的迭代速度远超传统法律制定周期。以ChatGPT系列为例，GPT-1到GPT-4o在7年内完成7次迭代，而硬性监管政策却往往因立法流程冗长，难以跟上技术发展。现有监管框架的滞后性导致自动驾驶责任认定、人工智能医疗审批等场景存在法律空白。自动驾驶领域的责任认定机制尚未完善，人工智能医疗设备的审批标准仍存争议，这些新兴技术领域的法律规制体系尚未健全。技术治理的被动应对模式，实质上造成了伦理原则在技术实践中的系统性剥离，暴露出传统监管范式在应对颠覆性技术创新时的根本性缺陷。这种规制赤字不仅威胁技术发展的可持续性，更对科技创新与社会价值的良性互动构成制度性挑战。国内最新颁布的《人工智能生成合成内容标识办法》明确提出人工智能生成合成内容标识主要包括显式标识和隐式溯源两种形式，试图通过技术驱动型治理手段消解监管时滞效应，但其在技术快速演进场景下的动态适应性仍有待验证。国际社会则另辟蹊径，发展出适应性敏捷治理框架，该框架强调通过快速迭代治理策略、构建多主体协同治理网络以及运用数据驱动型决策工具来提升监管响应效率。然而，这种治理模式在实际运作中面临显著的协调成本挑战，不同利益主体间的诉

① 1984年，美利坚合众国最高法院在“环球影业与索尼公司著作权纠纷案(Universal Studios, Inc. v. Sony Corporation of America)”中作出里程碑式裁决，确立只要产品能够具有一种潜在的实质性非侵权用途，产品的制造商和经销商就不承担帮助侵权责任。该司法判决创造性地提出技术中立原则判定标准，明确指出当技术产品具备可被客观认知的实质性非侵权商业应用场景时，其生产者与销售者即不应因第三方潜在侵权行为承担辅助侵权法律责任。

求差异、治理权限分配以及决策机制协调等问题,都构成了制约治理效能提升的结构性障碍。这种国内外治理路径的探索差异,反映出在人工智能监管领域尚未形成具有普适性的制度性解决方案。

其次,在价值层面,全球伦理共识建构远落后于技术跨境扩散速度。全球人工智能伦理标准因文化差异、地缘政治竞争呈现碎片化。欧盟《可信赖AI伦理准则》(*The Ethics Guidelines for Trustworthy AI*)与部分国家实践存在价值观冲突,而联合国《全球数字契约》(*Global Digital Compact*)的践行还有待观望。技术扩散通过贸易、投资等渠道加速,但伦理规范尚未形成约束力。值得注意的是,即便是处于前沿探索阶段的监管体系,其伦理评估标准在面对技术系统涌现的复杂特性时,仍面临动态适应性不足的实践困境。这反映出既有治理机制在技术迭代速度与伦理演进节奏的结构性错位,亟待构建更具韧性的跨国伦理协同机制。再者,监管工具系统的价值感知缺陷进一步放大了制度性治理困境。当前占据主流的算法审计与内容筛查技术,其设计逻辑深度植根于技术治理的工具理性范式。现有偏见检测系统虽然能够在表层消除显性歧视特征,却可能导致价值偏见以更隐晦的形式持续存在。

技术拟人化进程中的伦理失控风险揭示出现代技术文明的内在矛盾,即创造力的算法化追求在拓展人类能力边界的同时,正在动摇伦理秩序的存在论根基。这一矛盾的本质源于技术自洽性与社会伦理性的价值错位,其消解路径无法通过单纯的技术迭代实现系统自洽。作为社会建构物的技术系统,本质要求其必须嵌入动态伦理框架,而数字文明时代的技术治理转型,本质是通过技术理性与伦理精神的辩证融合,重塑人机契约结构,从而实现工具理性与价值理性的双重回归。

四、技术拟人化伦理秩序的重构路径

技术拟人化伦理秩序的重构,本质上是人类在人工智能时代对技术本质与文明价值的重新定位。当生成式人工智能通过模仿人类的创造力生成诗歌、代码甚至学术论文时,其技术内核与伦理框架的冲突已不再是抽象的理论议题,而是关乎知识生产、文化存续与社会信任的实践危机。DeepSeek等模型展现的创造性失控现象,揭示出技术拟人化进程中的矛盾局面,即算法系统越是通过参数优化逼近人类创造力,其价值判断的无根性就越发显现。这种矛盾冲突的展现,也要求对人机问题的处理要超越传统控制论思维,从技术重构、制度创新与价值共建三个维度重构伦理秩序。

(一)技术理性与价值理性错位的解决路径:嵌入式伦理治理框架

生成式人工智能技术理性与伦理理性的本体冲突,本质上是工具理性与价值理性在技术架构层面的断裂。破解这一矛盾需从技术设计的底层逻辑重构出发,建立伦理前置的技术研发范式,通过价值内嵌机制实现技术理性与伦理理性的动态平衡。嵌入式伦理治理架构(Embedded Ethical Governance Architecture, EEGA)创新性地构建了技术伦理的前置性规制路径。该框架主张在人工智能与嵌入式系统的研发初期,即通过嵌入设计伦理原则[24](P5537),确保技术演进轨迹与社会伦理标准保持动态对齐。其核心逻辑在于,通过伦理要素的参数化建模与价值敏感算法设计,实现技术决策过程的价值理性嵌入,从而有效规避技术成果应用阶段的伦理风险。这种预防性伦理规制策略,不仅为技术创新的负责任发展提供了制度性保障,更在本质上推动了技术工具理性与价值理性的有机统一。相较于传统的事后伦理审查机制,该框架展现出更强的系统适配性与实践效能。

该治理框架的核心创新在于伦理维度的深度内嵌,要求技术开发者在算法架构的顶层设计阶段,就将伦理考量作为系统性要素纳入技术实现路径。该范式以嵌入式伦理为学理根基,主张伦理专家应深度介入人工智能研发全流程,推动伦理原则与技术逻辑的动态耦合。为实现这一目标,构建伦理效应量化评估模型也极为重要。这就要求开发者对算法系统的价值偏差指数、隐私风险概率、信息失真度等关键指标进行多维度量化测评。这种评估机制不仅构成了技术性能参数体系的有机组成部分,更形成了对技术伦理风险的预防性识别与管理框架,体现了从技术工具理性向伦理价值理性转变的治理理念革新。

同时，拟人化技术透明化改造是其中的关键环节。人工智能系统需明确展示其道德判断能力等级，如欧盟《人工智能法案》的风险分级，将人工智能系统分为不可接受风险、高风险和有限风险三类。同时，针对生成式人工智能建立内容溯源强制标识机制，确保用户能够清晰辨识信息来源的技术属性。这种分级标识制度不仅通过标准化风险等级标识，显著提升了技术系统的透明度，使用户、开发者及监管者能够准确理解技术边界，还有效降低了因技术拟人化特性引发的认知偏差，帮助用户建立对人工智能系统能力局限的理性认知，从而在技术交互过程中作出更合理的决策。最后，伦理规范的动态适配机制是确保人工智能系统持续符合社会伦理要求的关键创新。该机制的核心在于构建人机协同的伦理训练体系，通过将人类伦理专家委员会实时生成的决策数据，系统性地纳入人工智能模型的强化学习反馈环路。这种设计使人工智能系统能够在训练过程中动态吸收最新的伦理准则，实现伦理决策能力的持续进化。

嵌入式伦理开发框架通过伦理前置设计规范、拟人化技术透明化改造和动态伦理适配机制，为技术设计与伦理冲突提供系统性解决方案。这一路径的核心在于技术与伦理的深度融合，通过量化评估、透明化改造和动态训练，确保人工智能技术在设计阶段即嵌入伦理规范，从而有效缓解技术设计与伦理冲突。

（二）技术社会化与责任归因困境的解决路径：三元协同治理体系

生成式人工智能的技术应用与社会规范冲突是其创造性失控的直接表现。为应对这一问题，三元协同治理体系提出了一种多层次、多主体协同治理路径，通过责任链区块链溯源机制、社会规范动态校准平台以及信任重建工程，形成技术应用与社会规范的良性互动。

一是建立基于区块链技术的责任溯源链机制。生成式人工智能系统的控制主体必须确保技术实施全流程及决策逻辑的完整可追溯性。可以通过嵌入区块链数字水印，实现对生成内容来源的精准追溯及其合规性验证。区块链技术特有的数据归因功能与不可篡改特性，为建立清晰的主体责任界定提供了技术保障，有效消除了传统责任认定中的模糊地带。二是构建社会规范动态校准平台。该平台以多源异构数据融合技术和人机协同监测技术为支撑，实时采集并分析社会规范体系的动态演变数据，涵盖法律法规更新、公众舆论走向及社会文化变迁等多维度信息。通过构建这种实时校准机制，确保人工智能系统的输出内容与社会规范保持动态适配。同步开发的动态风险评估指标体系，运用大数据分析模型对社会风险进行实时监测预警。当人工智能输出内容与社会规范偏离度超出预设临界值时，自动触发风险熔断机制，从而有效规避技术应用过程中可能出现的伦理失控风险。三是信任重建工程。风险与挑战带来的信任问题可能会阻碍人工智能的发展与应用，面对这样的问题，二十国集团发布《G20人工智能原则》，强调人工智能的伦理和社会责任，提出负责任地管理可信赖的人工智能基本原则，以及实现可信赖的人工智能国家间合作。此外，同步推行人机协同监管模式，进一步强化信任修复机制。根据欧盟《人工智能法案》的要求，对可能影响人权的人工智能系统，必须保障用户获得人工干预的权利，并通过独立第三方审计确保系统合规性。通过人类专家与人工智能系统的协同决策，有利于弥补人工智能在伦理判断方面的局限性，同时构建技术与社会之间的信任桥梁。

三元协同治理体系通过责任溯源链机制、社会规范动态校准平台和信任重建工程，为技术应用与社会规范冲突提供系统性解决方案。这一路径的核心在于多主体协同与动态校准，通过责任明确、规范校准和信任重建，确保技术应用与社会规范的动态一致性，从而有效缓解技术应用与社会规范冲突。

（三）治理异步性与价值共识冲突的解决路径：敏捷治理生态构建

生成式人工智能的技术治理与伦理监管冲突是其创造性失控的深层次问题。为解决这一问题，敏捷治理生态构建提出了一种动态适应性治理路径，通过弹性监管框架设计、全球伦理治理网络以及价值对齐工程，形成技术治理与伦理监管的协同机制。

第一，构建弹性化的监管框架体系。其核心在于创设监管沙盒动态指数，该技术治理工具能够根据

人工智能系统的迭代发展速率,实时调整监管介入的力度与方式。欧盟《人工智能法案》专门设立章节,为人工智能监管沙盒机制制定了详尽的运行规则,这种制度创新为探索生成式人工智能的治理范式提供了安全试验空间。同时,配套建立跨学科风险应对团队,整合法律专家、伦理学者及技术专家的专业智慧,形成复合型监管能力,确保在技术治理过程中能够即时识别并处置伦理风险。这种机制设计显著提升了监管体系的灵活性与环境适应性,使其能够有效应对人工智能伦理监管领域不断涌现的新型挑战,为技术发展与伦理规范的协同演进提供了制度保障。第二,构建全球伦理治理互联网络。其核心在于开发跨文明伦理转换机制,运用文化特征提取算法实现不同伦理体系间的语义转换与规范对接,从而确保全球伦理治理体系的可操作性。同步建立人工智能国际伦理评估框架,将技术合作与技术援助同伦理评级结果相挂钩,类似于联合国技术促进机制中的伦理审查模式。这种评估框架不仅有助于增强全球伦理治理的公平性,更能为国际技术合作提供标准化的伦理审查依据,推动技术发展与伦理规范的全球化协同。第三,实施价值对齐工程。运用博弈论优化算法,对技术安全治理与伦理合规监管的资源分配进行动态优化,确保技术治理目标与伦理监管要求保持内在一致性,有效规避治理目标错位风险。同时强化伦理审计机制,在传统算法审计流程中嵌入价值观向量分析模块,重点检测模型决策过程中在公平性、正义性等伦理维度上的表现。

敏捷治理生态构建通过弹性监管框架设计、全球伦理治理网络和价值对齐工程,能为技术治理与伦理监管冲突提供系统性解决方案。这一路径的核心在于动态适应与全球协同,通过弹性监管、跨文明伦理治理和价值对齐,确保技术治理与伦理监管的同步性与一致性,从而有效缓解技术治理与伦理监管冲突。

综上,技术拟人化与伦理秩序的重构,最终指向人机关系的范式转型。当算法系统开始深度参与人类价值判断时,人类既不能沉溺于技术乌托邦的幻想,也不应陷入卢德主义式的抗拒。真正的解决之道在于建立伦理嵌入式的技术演进路径,在应用层面发展人机伦理协商机制,通过可解释界面促进技术系统与人类价值观的持续对话;在文明维度坚守人文精神的底线,警惕工具理性对价值领域的全面殖民。唯有如此,我们才能在拓展智能边界的同时,守护那些使人类成为人类的东西——对真善美的永恒追寻,对正义的执着坚守,以及在不确定性中依然选择相信的勇气。这些无法被参数化的价值维度,正是技术拟人化进程中不可逾越的伦理地平线。

(本院硕士研究生陈源源在本文的文献资料整理、数据收集工作中付出了辛勤劳动)

参考文献

[1] 魏钰明. DeepSeek突破效应下的人工智能创新发展与治理变革. 电子政务,2025,(3).
[2] 凯文·凯利. 失控:人类的最终命运和结局. 张行舟、陈新武、王钦等译. 北京:电子工业出版社,2016.
[3] 张慧敏. DeepSeek-R1是怎样炼成的? 深圳大学学报(理工版),2025,(2).
[4] 玛格丽特·A. 博登. 人工智能哲学. 刘西瑞、王汉琦译. 上海:上海译文出版社,2006.
[5] Alan Mathison Turing. Computing Machinery and Intelligence. *Mind*, 1950, 49.
[6] 休伯特·德雷福斯. 计算机不能做什么:人工智能的极限. 宁春岩译. 北京:生活·读书·新知三联书店,1986.
[7] 吉尔贝·西蒙东. 论技术物的存在模式. 许煜译. 南京:南京大学出版社,2024.
[8] 公维敏. 吉尔伯特·西蒙栋:技术物的发明与关键点网络. 自然辩证法通讯,2023,(6).
[9] John Rogers Searle. Minds, Brains, and Programs. *Behavioral and Brain Science*s, 1980, (3).
[10] 海德格尔. 存在与时间. 陈嘉映、王庆节译. 北京:商务印书馆,1999.
[11] Mark Anthony Runco. AI Can only Produce Artificial Creativity. *Journal of Creativity*, 2023, (3).
[12] 凯瑟琳·海尔斯. 无思考:认识非意识的力量. 冷君晓译. 南京:江苏人民出版社,2024.
[13] Caterina Moruzzi. Artificial Intelligence and Creativity. *Philosophy Compass*, 2025, (3).
[14] Robert Kane. *The Significance of Free Will*. New York: Oxford University Press, 1996.
[15] 亨利·柏格森. 创造进化论. 姜志辉译. 北京:商务印书馆,2004.

[16] Margaret Ann Boden. *The Creative Mind: Myths and Mechanisms*. London: Routledge, 2004.
[17] Amy Kind. *Imagination and Creative Thinking*. Cambridge: Cambridge University Press, 2022.
[18] Dustin Stokes. Minimally Creative Thought. *Metaphilosophy*, 2011, (5).
[19] David Novitz. Creativity and Constraint. *Australasian Journal of Philosophy*, 1999, (1).
[20] 杨顺.ChatGPT等生成式人工智能对学术诚信的挑战及应对.新兴权利(集刊),2024,1.
[21] Batya Friedman, Eric Brok et al. Minimizing Bias in Computer Systems. *SIGCHI Bulletin,* 1996, (1).
[22] 冯晓青,沈韵.生成式人工智能服务提供者著作权侵权责任认定.法治研究,2025,(1).
[23] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, et al. AI Models Collapse When Trained on Recursively Generated Data. *Nature*, 2024, 631.
[24] Daniele Rossini, Danilo Croce, Sara Mancini, et al. Actionable Ethics through Neural Learning. *AAAI Conference on Artificial Intelligence,* 2020, (4).

The "Creative Out-of-control" in Generative Artificial Intelligence and the Reconstruction of Ethical Order In Technological Anthropomorphism

Wu Yingping（Southwest Minzu University）

Abstract Distinguished from mere technological runaway and ethical anomalies, the Creative Out-of-control in generative artificial intelligence (AI) refers to systemic risks arising when technological systems with quasi-subjectivity deviate from human-preset ethical trajectories through their autonomous optimization mechanisms. The creative capacities of generative AI span multiple dimensions encompassing cognitive science, ethics and ontology. Machine creativity does not merely imitate or replace human creativity, but represents a novel cognitive interface co-evolved through the symbiotic development of technological systems and human civilization. The essence of creative out-of-control lies in the conflict between technological anthropomorphism and ethical disembeddedness, manifesting as the misalignment between technological rationality and value rationality, practical dilemmas in techno-socialization and accountability attribution, and institutional paradoxes stemming from governance asynchronicity and value consensus. This paradigmatic confrontations require a tripartite approach to human-machine relations through technological reconfiguration, institutional innovation and value co-creation; it necessitates the establishment of an embedded ethical governance framework, a triadic collaborative governance system and an agile governance ecosystem to reconfigure the ethical order of technological anthropomorphization.

Key words artificial intelligence; Creative Out-of-control; technological anthropomorphism; ethical disembeddedness; order reconstruction; DeepSeek

■ 作者简介 吴映平,西南民族大学哲学学院副教授,四川 成都 610041。
■ 责任编辑 涂文迁