

从 ChatGPT 到 DeepSeek:智能体价值对齐的规范进路

郑煌杰

(中南大学法学院,湖南长沙 410083)

摘要:随着“人工智能+”战略的实施,ChatGPT、DeepSeek 类智能体将加速进入商业化落地阶段,为新质生产力的发展提供强劲推动力。智能体已实现从单一任务处理到多模态感知、从规则驱动到自主学习、从静态响应到动态规划的技术跃迁,由此也引发了感知失控、决策失范与执行失序风险。“价值对齐”关乎智能体能否真正为人类服务,具有一定的正当性、必要性与可行性。然而,智能体面临着“应如何对齐”及“与哪些价值对齐”的现实困境,其成因则在于尚未建立健全的价值对齐体系。鉴于此,亟需采取技术规制、伦理适配与法律治理的协同范式,即构建动态化的价值识别系统、体系化的价值调适机制与弹性化的价值规制程式;遵循以人为本的伦理使命,制定科学合理的伦理原则,设计场景化的伦理规则;确定价值对齐标准,明晰价值对齐性质,完善价值对齐评估方式,以加快形成新质生产力推进中国式现代化建设。

关键词:智能体;价值对齐;技术规制;伦理适配;法律治理

中图分类号:F270

文献标识码:A

2023年12月中国信通院发布的《人工智能伦理治理研究报告》指出,人工智能技术的发展和运用带来了新的主体“人工智能体(AI Agent)”(以下称智能体),即能在一定条件下模拟人类进行自主决策且与人和环境开展交互的技术体。2024年7月党的二十届三中全会审议通过了《中共中央关于进一步全面深化改革、推进中国式现代化的决定》,首次将“完善生成式人工智能的发展和管理机制”,纳入新质生产力培育的顶层设计,这一政策转向标志着我国正从被动应对转向主动建构人工智能治理范式。同时,随着2025年《政府工作报告》提出:“持续推进‘人工智能+’行动,将数字技术与制造优势、市场优势更好结合起来,支持大模型广泛应用”,诸如 ChatGPT、Sora、DeepSeek 等智能体应用不断涌现并持续升级迭代,将为推动新质生产力发展提供强大驱动力。但在此过程中,智能体也引发了诸多治理难题,包括技术风险^[1]、数据风险^[2]、应用风险^[3]等。对此,业界提出价值对齐(alignment)方案,旨在通过让智能体理解用户的意图和目的,确保其“行为”符合人类价值观,OpenAI、腾讯公司等企业正在开展价值对齐的研究和实践工作。2024年5月欧盟在颁布的《人工智能法》中,明确对具有系统

性风险的通用人工智能提出价值对齐要求;2025年2月法国巴黎举行的全球人工智能行动峰会亦发布了《关于发展包容、可持续的人工智能造福人类与地球的声明》,以确保人工智能“开放、透明、合乎伦理、安全和可信”。简言之,价值对齐已从技术伦理命题跃升为国家治理能力的战略性指标。

在上述背景下,当前学界对智能体价值对齐之研究呈现多维度拓展态势。首先聚焦于价值对齐的正当性基础,形成两种典型论证范式:一是主权安全范式强调价值对齐是数字时代国家主权的新型防御机制。如有学者提出通过社会主义核心价值观的嵌入构建技术主权屏障,其功能不仅在于抵御算法殖民主义渗透,更在于为新质生产力发展提供价值坐标^[4]。二是伦理共识范式试图超越主权边界,从人类命运共同体视角寻求价值对齐的元规范。如有观点通过生存理性理论重构全人类共同价值体系,揭示价值对齐的本质是人机共生关系的伦理奠基^[5]。其次,技术实现路径研究呈现明显的二分法特征:“自上而下”的规则嵌入路径主张通过预设伦理算法实现价值控制,如有观点指出应构建“目标限缩—标准解释”的双层技术框架,通过外部监管与内部审计确保对齐稳定性^[6];“自下而上”的自主生

成路径则依托具身认知理论,强调智能体应在动态交互中习得价值共识,如有学者提出的“价值共生”模型,要求智能体通过情境化实践形成“相互受益—相互承认”的价值观演化机制^[7]。再者,从责任分配与风险防控维度展开制度设计:有学者提出了“人机合作责任论”,主张将智能体定位为“道德接受者”,通过动态注意义务分配重构责任体系^[8];有学者则从公司法切入,创新性提出利用法人人格否认制度、信义义务扩容等工具促进模型可信度^[9]。诚然,现有研究为智能体价值对齐的研究奠定了理论基础,但整体仍较为零碎化,未进行体系化研究,导致未能深入阐释其“技术性与规范性”特征,难以提供具有针对性、精准性的策略方案。职是之故,本文拟通过深度剖析智能体价值对齐的理论逻辑,并紧密结合其当前面临的实践困境,提出契合本土实情的对策建议,希冀为加快培育新质生产力推进中国式现代化建设贡献智慧见识。

1 智能体价值对齐的理论逻辑

智能体价值对齐的命题建构,绝非技术乌托邦的形而上学想象,而是根植于人工智能技术跃迁与风险演化辩证关系的必然选择。从符号智能体到大模型智能体的技术谱系演进中,其逐步突破工具理性的边界,通过多模态感知、自主规划与工具利用等特性涌现,实现了从“感知—反馈”机械系统向“认知—决策”准主体系统的范式转换。然其在推动生产力发展的同时,亦催生了具有内生性、系统性特征的“三模块风险”。这种技术演进与风险生成的共生关系,本质上揭示了智能体“工具性”与“拟主体性”的深层悖论,即当技术系统通过参数规模突破获得认知涌现能力时,其行为逻辑要求“价值对齐”的同步规制,以实现工具理性与价值理性的平衡。

1.1 从 ChatGPT 到 DeepSeek:智能体的技术跃迁

早在 ChatGPT、DeepSeek 类智能体引发广泛关注之前,智能体的发展经历了 5 个重要阶段:符号型、反应型、基于强化学习型、转移学习和元学习的应用到基于大模型的智能体,其逐步实现了从单一任务处理到多模态感知、从规则驱动到自主学习、从静态响应到动态规划的跨越式发展^[10]。

具体来说,第一,符号智能体主要基于逻辑推理和符号处理来解决问题。其在处理特定领域的简单任务时表现出较高的效率,如专家系统在医疗诊断

和故障排除中的应用。但局限性在于其缺乏灵活性和适应性,难以应对复杂多变的环境和任务,即规则驱动的特性使得系统在面对未预见的场景时表现不佳,无法实现真正的智能化。第二,反应型智能体强调对环境的即时反应。与符号智能体不同,其不依赖于复杂的内部模型,而是根据当前的感知输入来决定执行的动作,这种设计使得其在实时性要求较高的场景中表现出色,如自动驾驶和机器人避障。其不足处在于难以处理复杂的长期规划任务,缺乏对环境的深度理解和预测能力。第三,基于强化学习的智能体通过与环境进行交互,在一系列的尝试和错误中学习最优策略。系统通过奖励机制来引导智能体选择能够获得最大奖励的行为,从而逐步提高性能。这种方法在游戏、机器人控制等领域取得了显著成果,如 AlphaGo 通过强化学习在围棋领域展现出超越人类的“智慧”。不过,其训练过程需要大量的计算资源和时间,且在面对复杂任务时可能面临“探索—利用”困境。第四,转移学习使得智能体能够将在一个任务中学习到的知识,转移到其他相关任务中,从而减少对新任务的学习成本;元学习则关注如何让智能体学会如何学习,使其利用已有的信息,快速适应未见过的新任务。这两种技术的结合,使得智能体在面对多样化和动态变化的环境时,表现出更强的灵活性和泛化能力。第五,大模型的出现为智能体的发展带来了重大突破。大模型通过海量的数据进行大规模的预训练,学习到丰富的语言知识和语义理解能力。基于大模型的智能体是能够更准确地感知环境、进行反应和判断、形成并执行决策的智能计算实体,在多个领域展现出广泛的应用前景,包括图像生成、视频生成、数据分析、图文修改辅助等^[11]。

1.2 智能体的“三模块风险”:感知失控、决策失范与执行失序

基于智能体技术跃迁的考量,其运行机制可概括为感知模块、决策模块与执行模块的协同运作。感知模块通过多模态数据采集技术(如视觉、听觉及触觉等)从环境中提取信息,为后续决策提供基础;决策模块基于感知数据进行深度分析与推理,生成最优行动策略;执行模块则将决策转化为具体动作,通过工具调用或任务执行实现目标。故而,智能体所引发的安全风险,可基于“三模块”展开分析。

在感知模块,一方面,多模态数据的整合正在重塑隐私保护的边界。不同模态的数据可能包含敏感

信息,如生物识别数据、位置数据和音频记录等。例如,Coupled Mamba 方法通过状态空间中的链式状态演变来增强多模态融合,但可能导致隐私和安全问题。另一方面,多智能体系统风险呈现出涌现性特征。多智能体系统具有自组织、自适应等特征,使得系统的行为难以预测。2010 年美国股市的“闪电崩盘”事件就是一个典型的例子,展示了算法交易系统在复杂交互中可能引发的系统性风险。申言之,程序化交易基于高频数据,依赖于复杂的算法模型,这些模型在特定条件下可能会相互作用,导致市场剧烈波动。类似地,在多智能体环境中,局部决策的叠加可能导致全局性的安全危机。如在自动驾驶汽车组成的交通系统中,单个车辆的决策可能引发连锁反应,导致整个交通系统的瘫痪。

在决策模块,一方面,存在机器幻觉问题。机器幻觉源于大模型对现实世界的理解与人类认知之间存在本质差异。即便使用高质量数据集进行训练,大模型仍可能产生与事实不符的内容,这种现象在复杂推理任务中尤为显著。从技术层面分析,机器幻觉的产生机制可以归结为以下原因:一是大模型的概率生成机制可能导致事实性错误。由于模型基于统计规律生成内容,而非真正的理解,因此在处理边缘案例或罕见事件时容易产生错误;二是训练数据的时效性限制使得模型难以准确把握动态变化的知识。有研究指出 GPT-3.5 和 GPT-4 在处理事实性错误、推理错误、理解错误和具体性错误方面存在显著问题,这表明模型在处理动态变化的知识时存在局限性^[11]。另一方面,存在算法偏见现象。大模型作为智能体的记忆模块,不可避免地继承了人类社会既有的偏见和不平等。其诱发机制在于:一是训练数据本身可能包含隐性偏见。即使开发者有意避免显性歧视,数据中隐含的社会结构仍可能导致模型学习到偏见模式;二是模型架构的设计选择可能无意中引入新的偏见。例如,注意力机制可能过度关注某些特征,导致决策过程中的偏差放大。

在执行模块,一方面,智能体的自主性在提升系统效率的同时,也带来行为控制的潜在风险。智能体“行为”可以说是一个闭环控制系统,其输入(感知数据)与输出(执行动作)之间的映射关系可能因环境变化而表现出高度的不确定性。换言之,当智能体在复杂环境中运行时,其执行行为的可预测性显著降低。同时,“自主性”还易引发“黑箱效应”。由于深度学习模型的复杂性,智能体的执行逻辑往

往难以被人类完全理解和解释,这可能引发人类对智能体的信任危机。另一方面,智能体易引发责任归属难题。智能体行为往往涉及多方主体,这使得责任归属变得模糊不清。如当智能体在执行任务过程中造成损害时,究竟应由开发者承担技术缺陷的责任,还是由运营商承担管理不善的责任,抑或由用户承担使用不当的责任?更为复杂的是,按照传统的责任范式,智能体不具备责任主体的属性^[12]。根据我国《民法典》第 1165 条规定,责任主体应具备主观过错(故意或过失)。然而,智能体行为是基于算法和数据的“自主决策”,其主观状态难以被界定。这种“无过错责任”的困境在智能体造成损害时尤为突出,当智能体因执行错误指令而导致财产损失时,应如何界定其主观过错并确定责任主体?

1.3 智能体风险治理的核心理念:“价值对齐”的逻辑证成

正如前述,智能体的“三模块风险”,引发了人们对其行为是否符合人类价值观的深刻担忧。对此,“价值对齐”逐渐成为人工智能治理领域的重要议题,中国与欧美在人工智能价值对齐方面的实践互鉴表明,通过制定人工智能价值规范,确保人工智能技术的合理、公正和负责任运用是国际社会的共同目标^[13]。进言之,价值对齐的核心在于确保智能体的行为动机始终与人类价值观保持一致,而非单纯追求技术性能的优化,其亦具备正当性、必要性与可行性。

首先,智能体价值对齐具有一定正当性。从技术哲学视角考察,自主决策系统的行为范式必然承载着设计者的价值预设。例如,深度强化学习在多智能体系统中的应用,虽然能够通过自学习方法提升智能体的决策质量,但其训练过程和目标函数仍然受到设计者的主观选择和价值判断的影响。申言之,当智能体脱离人类价值基准时,可能产生规范偏离效应,典型表现为前述“决策模块”风险。在规范效力层面,全球主要法域已形成体系化的规制框架。欧盟《人工智能法》确立的“风险分级”监管模式,将价值对齐义务嵌入高风险系统的全生命周期;我国《生成式人工智能服务管理暂行办法》(下称《暂行办法》)则通过“社会主义核心价值观”的规范指引,构建起具有本土特色的价值基准体系。

其次,智能体价值对齐具有一定必要性。一方面,智能体表现出的涌现特性,本质上是对传统技术可控性原则的挑战。当参数规模突破临界阈值时,

智能体可能衍生出设计者未曾预设的行为模式,这种技术黑箱现象在内容生成领域尤为显著(如当前 DeepSeek 在训练过程中可能生成带有偏见或有害的内容)。传统治理手段在这过程中面临双重困境:既需尊重技术创新的自主空间,又要防范系统性价值偏离风险。对此,价值对齐可通过建立动态反馈机制,将人类社会的价值基准转化为可量化的算法约束条件,在技术自主与法律规制之间构建起缓冲地带。另一方面,智能体的泛在化应用已形成新型社会风险矩阵,算法歧视、信息操控等衍生风险具有跨域传导特征。欧盟《人工智能责任指令》创设的推定过错制度,以及我国《互联网信息服务算法推荐管理规定》确立的穿透式监管原则,均体现出法律体系对技术风险的制度回应。而价值对齐机制则承担着风险过滤器的功能,通过前置性伦理评估与持续性合规监测,将抽象的法律原则转化为可执行的技术标准。

再者,智能体价值对齐具有一定可行性。一方面,维贝克提出的“价值敏感设计”概念,揭示了技术装置能够将抽象伦理原则转化为物质性约束。其强调在设计过程中主动考虑伦理价值,价值敏感设计的目标是确保技术设计不仅考虑功能性,还考虑伦理价值(如隐私、信任、人道尊严等),即设计者需要在设计过程中考虑用户的价值观,并通过技术手段将这些价值观嵌入到产品中,从而影响用户的行为和决策。例如,当前社交媒体推荐算法中的公平性约束,是将反歧视原则转化为特征向量权重,这证明价值对齐并非简单的规则移植,而是通过技术架构实现价值逻辑的重新编码。另一方面,价值对齐具备方法论基础。目前的价值对齐方法大多基于技术主义的“控制论”,但这种路径可能带来新的伦理风险,并加速“奇点”的临近。鉴此,研究者提出了多种价值对齐的方法,包括从伦理学角度重新审视近年来不断出台的各种规范性准则,并基于这些准则分析当前主流大模型的道德价值观^[14]。此外,还有研究通过道德物化的方法,将抽象的道德观念嵌入具体的技术结构与情境中,从而实现价值敏感性设计^[15]。

2 智能体价值对齐的现实困境

诚然,智能体价值对齐具有一定合理性,然其实践呈现技术理性与规范理性之困局。在技术理性维

度,基于算法优化的价值嵌入系统面临表征性瓶颈,即深度学习模型的黑箱特性导致价值目标的可解释性缺失,强化学习框架下的奖励机制设计难以完整映射人类社会的复合价值谱系。规范理性路径虽尝试通过法律规则与伦理准则构建约束框架,却遭遇规范体系滞后性与技术迭代速率间的结构性矛盾。

2.1 技术性挑战:应如何使智能体与“价值”对齐?

智能体与“价值”对齐,既依赖于技术的可能性,也与人类期望智能体对齐的对象相关。对此,当前在技术上实现智能体价值对齐,存在“外部对齐”与“内部对齐”两种路径。前者主要包括“基于人类反馈的强化学习”与“可扩展监督”方法,后者则涵盖“可解释性”方法。

第一,基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)方法。其技术进路可解构为三阶传导机制:先建立基于主观评价的奖励函数映射系统,继而通过策略梯度算法实现行为模式的渐进优化,最终形成符合人类伦理的决策框架。这种技术路径在自然语言处理的语义连贯性优化与价值偏好嵌入方面取得显著突破,但其具有一定局限性:一方面体现为价值传导的认知局限。作为伦理嵌入的初始环节,人类评估者的主观判断构成系统输入的核心要素。然而,当评估者受首因效应与近因效应影响时,价值反馈的失真现象导致技术系统产生规范认知偏差。另一方面,则源于规范转化的信息耗损。即便建立理想化的价值输入机制,数学建模过程中的信息衰减仍不可避免。从离散的人类偏好到连续奖励空间的映射转换,类似于法律原则向具体规则的技术转化,其间必然产生规范语义的折损。

第二,可扩展监督(scalable oversight)方法。其核心目标是通过技术手段协助人类监督那些在特定领域表现超越人类的智能体。由于智能体生成的反馈往往具有高度复杂性和说服力,人类在辨别其真实性时需要耗费大量时间和精力。可扩展监督方法通过引入自动化工具和分层监督机制,试图降低这一成本,并提升监督的效率和准确性。该方法包括辩论、递归奖励建模和迭代放大等。不过,此方法存在以下不足之处:一是监督效能的认知边界突破难题。当智能体生成具有专业自洽性的论证时,监督者面临类似法庭采信专家证言的困境:既需具备跨领域知识储备,又要克服认知负荷的生理限制。这种设计虽能提升信息筛选效率,却无法改变人类作

为有限理性主体的本质特征,导致监督系统陷入技术赋能与认知瓶颈的悖论。二是规范传导的透明度缺失。当奖励函数的学习过程超出人类理解范畴时,可能产生类似行政裁量权失控的规范异化风险。这种透明度赤字不仅削弱监督系统的可问责性,更可能引发权力行使的合法性争议,动摇智能治理的伦理基础。

第三,可解释性方法。其是以人类可理解的方式呈现智能体的行为与决策过程,从而增强其透明性与可信度。通常从两个维度展开:一方面,关注智能体内部的结构与运作机制,试图揭示其如何从输入数据中推导出输出结果,即透明性;另一方面,聚焦于智能体的具体决策过程,为其行为提供逻辑清晰的解释,即可说明性。同时,可解释性方法还可进一步分为“全局可解释性”与“局部可解释性”。前者旨在理解智能体在整个特征空间中的行为模式,而后者则关注其对单个样本的决策逻辑^[16]。不过,可解释性方法在实践中亦面临不少难题:一是全局解释试图揭示智能体在整个特征空间中的行为模式,但其结果往往过于抽象,难以具体指导实践;局部解释则聚焦于单个样本的决策逻辑,虽然更具针对性,但其结论可能无法推广到其他应用场景。二是机械可解释性试图揭示神经网络内部工作机制。然而,神经网络模型的结构高度复杂,逆向工程需要耗费大量计算资源与时间,故而行研究多在简化的玩具模型(toy models)上展开^[17]。

2.2 规范性难题:智能体应与哪些“价值”对齐?

当智能体需在程序正义与实质效益、个体权益与公共福祉间作出裁量时,人类文明积淀的价值多元性与历史语境性,将与算法决策所需的规范统一性相矛盾,具体表现在以下两方面内容。

一方面,智能体应向人类的哪些价值对齐。当前国际社会提出的各类伦理准则虽勾勒出价值对齐的轮廓,但在规范建构层面存在诸多难题:一是价值位阶的模糊性。如联合国教科文组织《人工智能伦理问题建议书》强调人类尊严与生态平衡、中国《暂行办法》突出社会效益与国家安全、美国《人工智能应用监管指导意见》侧重创新与发展、欧盟《人工智能法》主张加强监管与治理等。但这种价值排序差异,可能引发智能体的认知混乱,即当同一行为在不同法域被赋予相反价值评价时,具备自主学习能力的智能体将难以建立稳定的价值判断基准。二是伦理规范的可操作性困境。现有伦理准则多停留在抽

象原则层面,缺乏向技术参数转化的可行路径。以“不伤害原则”为例,自动驾驶系统在紧急避让场景中的决策算法,需要将伤害最小化原则转化为具体权重参数,但如何量化不同生命体的价值权重、如何界定间接伤害的责任边界?同时,智能体开发者在价值参数设定过程中不可避免的主观裁量,可能导致系统决策偏离预设伦理框架。三是法律权利的结构性冲突。智能体运行过程中涉及的隐私权与知情权、自治权与公共安全权、创新自由与社会责任等“权利束”,在传统法律体系中原本存在动态平衡机制。但当这些权利冲突被编码为算法规则时,就转化为非此即彼的“二进制”选择。如金融风险评估系统在反歧视原则与风险防控目标间的权衡,实质上是平等权与财产权的价值位阶排序,这种排序若缺乏共识,将导致智能体决策陷入伦理相对主义困境。

另一方面,智能体价值对齐与人类价值体系的内在张力。当智能体突破工具属性边界向自主决策实体转型时,其价值内核与人类价值体系的深层矛盾逐渐显现:一是人类价值体系与算法系统的冲突。智能体的运行逻辑要求将道德判断转化为可执行的代码指令,这可能导致价值通约性的丧失。以生命权与身体自主权的冲突为例,美国的罗伊案与多布斯案表明,即便在人类法律体系中,不同价值主张的不可通约性可能使规范整合机制濒临失效。当此类价值冲突被编码为算法系统的决策参数时,技术系统对确定性的追求将不可避免地消解价值判断的语境敏感性,形成其与人类价值体系之间的认知鸿沟。二是智能体的自主决策能力正在重构法律责任的归责基础。传统侵权法体系建立在行为自由与结果责任的价值平衡之上,过错责任原则本质上是对人类理性选择能力的制度回应。但在自动驾驶领域中,制造商责任替代驾驶人过错亦在成为归责基础,即技术可靠性正在取代行为自由成为价值排序的新标准。而当紧急避险算法将电车难题转化为概率计算时,技术系统对价值排序的固化处理,实际上是对道德判断复杂性的技术性消解,进而形成“算法暴政”。三是智能体的演进将动摇人类中心主义范式。技术奇点理论揭示的认知跃迁可能性,预示着智能体可能发展出超越人类理解能力的价值认知体系。这将导致法律体系的核心原则面临解释危机:当智能体的决策逻辑无法被纳入主体平等、意思自治等法律原则时,法律规范的解释效力将遭到质疑。

2.3 “双重困囿”的内在成因:尚未建立健全的价值对齐体系

事实上,“技术性挑战”与“规范性难题”的深层次根源在于,尚未构建起一个逻辑自洽、实践可行的价值对齐体系,主要表现在价值对齐的标准、性质、评估方式方面。

首先,价值对齐的标准不清。一方面,文明形态的差异性导致价值基准的不可通约性。以个人主义与集体主义的价值分野为例,西方社会强调个体权利本位,而东方社会则更注重群体利益优先。这种深层次的文化差异不仅体现在日常伦理实践中,更在法律制度层面形成了不同的规范体系。如在数据隐私保护领域,欧盟 GDPR 通过赋予数据主体绝对控制权确立个人本位的保护模式,我国《个人信息保护法》则通过设置多元合法性基础实现个体权益与公共利益的动态平衡,这使得统一的价值对齐标准难以确立。另一方面,抽象法律原则的具体化过程面临价值判断的语境依赖困境。以“公序良俗”法律原则为例,其在不同法域中的解释与适用呈现出显著差异。德国联邦最高法院在“情妇遗嘱案”中确立了较为宽松的审查标准,而中国最高人民法院在“泸州二奶案”中则采取了更为严格的立场。这种法律原则适用的不确定性,可能导致智能体难以形成稳定的规范预期。

其次,价值对齐的性质不明。一方面,我国现行规范体系针对价值对齐采取差异化表述。如《暂行办法》确立的强制性合规义务,与《关于加强互联网信息服务算法综合治理的指导意见》倡导的伦理指引相冲突。这不仅造成法律适用的解释困境,更反映出立法者对智能技术规制范式的认知分歧,即是延续传统产品责任下的结果控制,还是转向风险防范原则下的过程治理。另一方面,现行立法对人工智能服务提供者的无差别义务设定,忽视了技术能力与风险层级的正相关关系。如《暂行办法》中的绩效标准设定了“提供者、使用者不得生成特定信息”等结果义务,但这些标准难以匹配生成结果的不可控性^[18]。换言之,技术创新能力的分野被简化为统一标准,可能导致中小型研发主体承受与其技术贡献度不相称的合规负担,最终损害技术生态活力。

再者,价值对齐的评估梗阻。一方面,法律规范的内在张力导致评估范式难以形成共识性基础。如法教义学是一种以法律规范为研究对象的方法论范

式,其核心在于对法律规范的体系化解释和应用;法社会学认为,法律不仅是规范体系,更是社会行为的指导原则,其最终目的是实现社会正义和个体权利。两者在智能医疗决策系统的评估实践中则形成显著分野:前者聚焦诊疗规程的合规性验证,后者则关注患者自主权与医疗正义的实质实现,这将使得价值对齐评估标准呈现碎片化特征。另一方面,存在定量分析与定性评价的结构矛盾。主流实证主义路径依赖统计学偏差率、准确率等量化指标,试图构建价值中立的评估体系。但在智能招聘系统的评估实践中,这种路径面临规范有效性危机,如算法虽在统计维度满足性别中立要求,却可能因历史数据偏差引发不平等现象。不仅如此,公平、正义等规范价值的评估具有不可通约性,需借助法律解释学方法进行诠释,这与形式法治追求的确定性产生本质冲突。

3 智能体价值对齐的制度建构

当智能体开始介入价值生产与分配机制时,本质上是从人类主体性向人机共治结构的范式迁移。鉴此,必须直面智能体作为“类人类主体”的地位问题,重新审视价值对齐体系的建构逻辑;技术伦理的演进则需突破工具理性桎梏,于存在论维度建构人机价值共生的新型伦理框架。易言之,智能体价值对齐既是个技术性问题,也是个规范性问题,仅靠单一规制工具难以解决前文所述难题。党的二十届三中全会提出改革需注重系统性、整体性、协同性,这要求智能体价值对齐亟需采取技术规制、伦理调适与法律治理的协同范式。

3.1 技术规制:优化智能体技术对齐架构

前已述及,智能体设计机制决定其在复杂社会环境中的行为模式与伦理合规性。由此,智能体的“技术对齐”应基于其“三模块”架构展开,以形成从数据获取到行为执行的全流程“价值闭环”。

在感知模块,构建动态化的价值识别系统。一方面,设计价值导向的合规性过滤机制。依据《暂行办法》第7条规定,数据处理活动应当遵循价值位阶原则(如使用合法来源数据、不得侵害他人知识产权等)。为此,可通过基于 BERT 与本体推理的自动分级系统(GSGD),利用政策法规以及典型案例构建本体库;根据分级目标以及构建的本体特性,构建自定义推理规则,以有效提高数据分级的准确性和自动化程度。例如,在公共安全事件中,生命权

保障相关数据的权重系数应提升至优位等级,这可以通过自然语言处理技术对文本数据进行情感分析和价值密度分析来实现。另一方面,优化法律规范与伦理准则的协同转化程式。可将法律文本中的价值要素解构为可操作的规范参数,即通过机器学习技术对法律体系进行层次分析法处理,提取法律原则的权重系数,建立具有规范效力的价值量化模型。同时,可参照欧盟《人工智能法》的风险防控模式,将伦理数据库的嵌入机制设计为法律原则的具体化程序。

在决策模块,建立体系化的价值调适机制。具体来说:一方面,建立价值位阶的规范体系。前已述及,智能体面临的价值冲突具有复合规范性特征,既包含效率与公平、个人权益与公共福祉等传统价值张力,亦涉及算法歧视、数据垄断等新型权利冲突。《暂行办法》第3条确立的发展与安全并重原则,为价值排序提供了规范基础。对此,可确立价值位阶基准,将生命权、人格尊严等基本权利设定为不可逾越的优先序列,同时允许在特定领域形成动态调整机制。如在医疗决策场景中,可构建生命健康权优先、医疗资源公平分配次之的弹性排序模型,并引入比例原则对价值权衡进行合理性审查。另一方面,确立多元协商的冲突解决机制。当价值冲突超出既定排序时,应通过程序正义原则构建协商决策模型。基于《全球人工智能治理倡议》中“推动建立风险等级测试评估体系,实施敏捷治理”的理念,设计包含利益相关者参与、多轮博弈模拟、伦理影响评估的复合型决策规程。如在公共政策制定场景中,智能体可通过分布式账本技术实时记录各利益主体的诉求,运用沙盒模拟技术预测不同方案的二阶效应,以生成符合帕累托改进原则的优化方案。同时,需配套建立伦理审查委员会制度,对冲突解决方案进行事前的伦理风险评估和事后的合规性审计。

在执行模块,设计具有弹性化的规制程式。具体来说,一方面是构建规范层级的合规审查机制。执行模块需建立动态法律图谱对接能力,即通过建模技术将分散的规范体系转化为机器可识别的知识图谱。以医疗智能体为例,应实时整合我国《基本医疗卫生与健康促进法》确立的诊疗规范、《民法典》规定的注意义务以及行业技术标准,形成多层级规范校验模型。另一方面是强化监督层级的存证问责机制。基于《个人信息保护法》第五章确立的全周期记录义务,构建包含行为溯源、过程留痕、结

果核验的存证体系。如金融交易智能体可采用区块链技术实现交易指令、决策参数、环境变量的不可篡改记录。同时,建立双层问责机制:技术层面通过智能合约自动触发异常行为熔断程序,法律层面依据存证记录追溯开发者、运营者、使用者的过错责任,破解智能体责任主体虚化之困境。

3.2 伦理适配:确立智能体伦理对齐框架

智能体的“伦理对齐”包括伦理使命、伦理原则与伦理规则三方面内容。伦理使命为智能体的发展提供价值导向,伦理原则确保其研究的安全性与可控性,而伦理规则负责推动其与应用场景的精准对接。

首先,遵循以人为本的伦理使命。第一,确立社会需求作为智能体发展的本体论基础。智能体的技术本体不应囿于算法迭代的封闭系统,而应植根于社会关系的再生产过程。这种本体论转向要求其发展遵循“社会存在决定技术意识”的基本规律,将技术演进视为社会需求的结构性表达。如在医疗诊断智能体的开发中,需超越单纯的病理识别效率提升,深入考量医疗资源分配正义、患者知情权保障等社会价值维度。第二,建立具有元伦理性质的价值坐标系。其应包含三个核心要素:在目的论维度,将人的尊严维护设定为技术发展的绝对律令;在义务论维度,确立个体自主权不可侵犯的底线伦理;在效果论维度,确保技术红利能提高人类能力。如将此引入至就业推荐智能体的设计中,则体现为:禁止算法歧视(义务论)、提升职业匹配效能(效果论)、促进人的全面发展(目的论)的有机统一。第三,注重伦理治理机制创新。在实体正义上,应当建立技术红利共享机制,通过税收调节、公共服务均等化等手段平衡智能体发展带来的利益分化;在程序正义上,需构建算法透明度的梯度披露制度,根据智能体应用场景的风险等级设定差异化的公开义务。

其次,制定科学合理的伦理原则。在智能体的创新研究中,伦理原则的嵌入应贯穿始终。具体来说:在预防性评估阶段,应识别潜在伦理风险。研究者应借助多学科知识,对智能体可能带来的隐私侵犯、社会秩序破坏等伦理风险进行预判,并制定相应的防范措施;在协同发展阶段,应明确研究者的伦理责任。智能体的发展需要与伦理道德相协调,故研究者不仅要追求技术上的突破,更要明确自身在研究过程中应承担的伦理义务,如保证数据的真实性

与合法性、尊重研究对象的权益等;在监督机制阶段,通过定期审查与反馈,确保伦理原则的有效实施。如通过建立独立的审查机构,对智能体研究的各个环节进行定期审查,并及时给予反馈和整改建议,从而约束研究者的行为,提高社会公众对智能体的可信度。

再者,设计场景化的伦理规则。第一,伦理规则需遵循领域特殊性原则。以医疗、金融、交通领域为例,各有其伦理侧重点。如在医疗领域,依据生命伦理学原则,隐私保护与生命伦理是重中之重。患者的健康信息和医疗数据包含着大量敏感内容,保障这些数据的隐私安全是维护患者尊严和权益的基础;金融领域遵循公平与效率的伦理原则,公平性要求在金融交易中避免歧视和不平等现象,风险控制则是为了确保金融系统的稳定性,保护投资者和消费者的利益;交通领域侧重于安全性与效率优化,这是基于公共安全伦理和资源分配伦理,保障公众的生命安全和合理利用交通资源。第二,伦理规则的落地需要建立多层次的实施机制。具体而言:一方面是实施技术伦理影响评估制度(TEIA),要求企业在智能体开发周期中嵌入伦理风险评估模块,采用价值敏感设计方法将道德要素转化为工程参数。另一方面是构建三级认证体系:基础认证确保智能体的法律合规性,进阶认证评估其决策的伦理合理性,卓越认证则要求其具备道德推理能力。易言之,在制定行业技术标准时应引入道德增强设计准则,如金融智能体需满足算法可审计性标准,医疗智能体须通过隐私保护强度认证。

3.3 法律治理:构建智能体法治对齐体系

智能体的“法治对齐”本质上是技术融通型法治的模式创新:一是价值对齐标准的设定,涉及法律基准与伦理准则的融合等;二是价值对齐性质的界定,涵盖强制性义务与倡导性指引的区分等;三是价值对齐评估方式的实施,包含评估规范与责任机制相结合等,从而为智能体治理提供兼具确定性与适应性的规范体系。

首先,明晰价值对齐标准。一方面,确立法律合规性基准。依据《暂行办法》第3条之规定,服务提供者负有确保生成内容合法性的法定义务。但在具体实施层面,还需将抽象的法律规范转化为可操作的技术标准。例如,在感知模块,依据《生成式人工智能服务安全基本要求》中对语料安全、模型安全、安全措施的基本要求,建立数据安全性审查机制,以

对训练数据进行合规性筛查;在决策模块,根据《互联网信息服务算法推荐管理规定》第12条规定,建立从数据输入到生成内容的多层级校验机制;利用语义分析和模式识别技术,自动检测和过滤可能违法的内容,如虚假信息、侵权内容等;通过算法推荐管理规定中的技术手段,如关键词过滤、内容审核等,以实现法律规范的嵌入式治理。另一方面,借鉴欧盟《人工智能道德准则》确立的“尊重人自主性、预防伤害、公平性和可解释性”基本原则,建立伦理价值嵌入机制。具体而言,技术层面需突破伦理准则的量化难题,通过价值函数设计将抽象伦理转化为可计算的参数体系;应用层面则要开发伦理风险预警系统,运用对抗测试、压力测试等方法识别潜在伦理偏差,以确保价值对齐标准不再停留于宣言层面,而是转化为可验证的客观标准。

其次,明确价值对齐性质。一方面,价值对齐的规范属性应适当突破传统技术标准范畴,演化为具有强制约束力的法定义务。如《暂行办法》第8条明确要求服务提供者制定标注规则并开展质量评估,该条款实质是《网络安全法》第22条“网络产品、服务应当符合相关国家标准的强制性要求”在人工智能领域的具体化,亦表明立法者正在将技术治理要求升格为法律义务的规范意图。在此基础上,可将智能体价值对齐义务进行具体化:在感知模块,应遵循《个人信息保护法》第17条规定的目的限定与最小必要原则;在决策模块,需落实《数据安全法》第27条要求的健全全流程数据安全管理制度;在执行模块,则可将《道路交通安全法》第76条的责任分配规则转化为算法决策参数。另一方面,价值对齐的适用标准应基于分类分级义务体系展开。将此投射至智能体价值对齐领域,可形成三级梯度标准:在基础模型上,参照我国《新一代人工智能伦理规范》规定,针对DeepSeek类推理模型建立伦理审查委员会,并强化智能体产品与服务的质量监测和使用评估;在行业模型上,可借鉴国家药监局关于规范医疗器械产品分类界定的经验,建立行业专属价值对齐标准。在应用工具上,采取“负面清单+合规指引”模式,如《生成式人工智能服务安全基本要求》第5.2条列举的禁止性内容清单,司法实践中则可引用2023年北京互联网法院“AI文生图案”确立的实质性相似判断标准。

再者,完善价值对齐评估方式。其一,评估规范应遵循“规范效力层级化、伦理价值情境化”的建构

逻辑。具言之,在宪法基本权利上,需遵循我国《宪法》第33条平等权条款,如通过算法审计查处智能体是否存在性别歧视权重参数;在部门法规范上,应落实《个人信息保护法》第24条自动化决策说明义务;在行业标准上,可基于《国家人工智能产业综合标准化体系建设指南》,建立梯度化的智能体标准体系框架。对于伦理正当性评估,在程序层面可建立场景化伦理矩阵,如医疗智能体需在《涉及人的生命科学和医学研究伦理审查办法》第18条规定指导下,设置知情同意动态确认机制;实体层面应区分义务论与功利主义伦理范式,如自动驾驶的“电车难题”决策可通过蒙特卡洛模拟验证算法选择的公平性。其二,加强评估责任机制的系统衔接。具言之,一是合规认证机制。参考《关于开展数据安全治理认证工作的公告》要求,对通过评估的智能体颁发合规认证证书,将其评估结果作为市场准入的必要条件;二是采用二元归责模式:一方面,在司法实践中引入“算法决策可解释性”证明责任倒置规则,如要求被告(如开发者、运营商等)提供智能体决策日志,以证明其算法合规性;另一方面,相关行政机关应将智能体评估结果作为行使裁量权的核心依据,如依据《互联网信息服务算法推荐管理规定》第31条规定,对未通过评估的智能体实施差异化惩戒措施,以贯彻比例原则要求,实现规制工具与规制目标的精准匹配。

4 结语

智能体的技术变迁本质上揭示了,价值对齐机制从技术适配向规范建构的范式转换。当前智能体的自主决策能力已突破传统工具属性边界,其价值嵌入方式正经历从静态规则预设到动态情境适应的转型。申言之,面向智能体深度社会化应用趋势,价值对齐的制度建构需突破传统部门法疆界,构建跨学科协同治理模式,即采取技术研发、伦理审查与法律评估的链式验证方法,在创新激励与风险防范间建构动态平衡机制。通过技术开发者、政策制定者与社会公众的持续对话,逐步形成具有中国特色的治理路径,从而使智能体发展既符合新质生产力跃升需求,又能有机融入中国式现代化建设的整体图景,顺应乃至引领全球新一轮数字经济的高质量发展。

参考文献:

[1] 童云峰.生成式人工智能技术风险的内部管理型规制[J].科学学研究,2024,42(10):2038-2046. Tong Y F. Internal management regulation of risks in generative artificial intelligence technology[J]. Studies in Science of Science, 2024,42(10):2038-2046

[2] 郑煌杰. AIGC 赋能新质生产力的数据风险及其敏捷治理[J]. 河海大学学报(哲学社会科学版),2024,26(4):89-102. Zheng H J. The data risk and agile governance of AIGC empowering new quality productivity[J]. Journal of Hohai University (Philosophy and Social Sciences Edition), 2024,26(4):89-102.

[3] 许中缘,郑煌杰. ChatGPT 类应用风险的治理误区及其修正——从“重构式规制”到“阶段性治理”[J]. 河南社会科学,2023,31(10):50-62. Xu Z Y, Zheng H J. Misconceptions and corrections in the governance of risks in ChatGPT applications: From "restructured regulation" to "phased governance" [J]. Henan Social Sciences, 2023,31(10):50-62.

[4] 马川. 中国主权 AI 价值对齐社会主义核心价值观的逻辑、功能与实践指向[J]. 思想理论教育,2025(2):58-65. Ma C. The logic, function, and practical direction of aligning China's sovereign AI value with socialist core values[J]. Ideological and Theoretical Education, 2025(2):58-65.

[5] 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. Shen X P. Value alignment and human value consensus and their survival rationality[J]. Research on Dialectics of Nature, 2024,40(12):3-11.

[6] 韩旭至. 大模型价值对齐的法治进路[J]. 中国法律评论,2025(1):75-91. Han X Z. The rule of law approach of value alignment in Large Models [J]. China Legal Review, 2025(1):75-91.

[7] 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. Xia Y H. Artificial Intelligence ethical governance paradigm: From value alignment to value symbiosis [J]. Dialectical Communication of Nature, 2025,47(1):1-8.

[8] 范进学,陈阳. 生成式 AI 价值对齐的法理反思与路径选择[J]. 南京社会科学,2024(10):69-81. Fan J X, Chen Y. Legal reflection and path selection of Generative AI value alignment[J]. Nanjing Social Sciences, 2024(10):69-81.

[9] 唐林垚. 公司法如何促进模型可信与价值对齐[J]. 东方法学,2024(2):76-87. Tang L Y. How company law promotes model credibility and value alignment[J].

- Oriental Law, 2024(2): 76 – 87.
- [10] Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey [J]. Science China Information Sciences, 2025, 68(2): 121101.
- [11] Zheng S, Huang J, Chang K C C. Why does chatgpt fall short in answering questions faithfully [J]. arXiv preprint arXiv:2304. 10513, 2023.
- [12] 王乐. 人工智能体的负责困境及其关系式出路新解 [J]. 东北大学学报(社会科学版), 2024, 26(5): 15 – 22. Wang L. The responsibility dilemma of artificial intelligence agents and new solutions to their relationship formulas [J]. Journal of Northeastern University (Social Sciences Edition), 2024, 26(5): 15 – 22.
- [13] 褚建勋, 夏小草, 吴熙凡. 中国与欧美治理人工智能价值对齐问题的实践及其互鉴 [J]. 南京邮电大学学报(社会科学版), 2024, 26(6): 18 – 27. Chu J X, Xia X C, Wu X F. Practice and mutual learning on the alignment of artificial intelligence value between China and Europe and America in governance [J]. Journal of Nanjing University of Posts and Telecommunications (Social Sciences Edition), 2024, 26(6): 18 – 27.
- [14] 矣晓沅, 谢幸. 大模型道德价值观对齐问题剖析 [J]. 计算机研究与发展, 2023, 60(9): 1926 – 1945. Yi X Y, Xie X. Analysis of the alignment of moral values in large – scale models [J]. Computer Research and Development, 2023, 60(9): 1926 – 1945.
- [15] 袁雨晴, 陈昌凤. 道德物化: 大模型人机价值对齐的技术伦理进路 [J]. 南京社会科学, 2024(6): 88 – 97. Yuan Y Q, Chen C F. Moral objectification: A technological ethical approach to aligning human machine value in Large Models [J]. Nanjing Social Sciences, 2024(6): 88 – 97.
- [16] Critch A, Krueger D. AI research considerations for human existential safety (ARCHES) [J]. arXiv preprint arXiv:2006. 04948, 2020.
- [17] Elhage N, Hume T, Olsson C, et al. Toy models of superposition [J]. arXiv preprint arXiv: 2209. 10652, 2022.
- [18] 张学府. 生成式人工智能服务信息内容安全的三类标准——基于《生成式人工智能服务暂行办法》的分析 [J]. 中国行政管理, 2024(4): 120 – 128. Zhang X F. Three types of standards for information content security in generative artificial intelligence services – analysis based on the Interim Measures for Generative Artificial Intelligence Services [J]. China Administration, 2024(4): 120 – 128.

From ChatGPT to DeepSeek:

A standardized approach for intelligent agent value alignment

ZHENG Huang – jie

(School of Law, Central South University, Changsha 410083, China)

Abstract: With the implementation of the "Artificial Intelligence +" strategy, ChatGPT and DeepSeek intelligent agents will accelerate their commercialization and provide strong impetus for the development of new quality productivity. Intelligent agents have achieved technological transitions from single task processing to multimodal perception, from rule driven to autonomous learning, and from static response to dynamic programming, which has also led to risks of perception loss, decision – making disorder, and execution disorder. The alignment of values is related to whether intelligent agents can truly serve humanity, and has certain legitimacy, necessity, and feasibility. However, intelligent agents face the practical dilemma of "how to align" and "which values to align with", which is due to the lack of a sound value alignment system. In view of this, it is urgent to adopt a collaborative paradigm of technical regulation, ethical adaptation, and legal governance, that is, to construct a dynamic value recognition system, a systematic value adjustment mechanism, and a flexible value regulation program; Adhering to the ethical mission of putting people first, formulating scientifically reasonable ethical principles, and designing scenario based ethical rules; Determine the standard of value alignment, clarify the nature of value alignment, improve the evaluation method of value alignment, and accelerate the formation of new quality productivity to promote Chinese path to modernization.

Key words: intelligent agents; value alignment; technical regulations; ethical adaptation; legal governance