道德发展视角下人工智能道德内生——以大语言模型为例 *

张玉强[1] 徐晓飞[2] 缑慧星[3] 张效初[**2,3]
（[1]安徽农业大学体育部，合肥，230036）
（[2]中国科学技术大学心理学系，合肥，230026）
（[3]中国科学技术大学生命科学学院，合肥，230026）

摘 要 当前人工智能道德研究聚焦于道德主体地位探讨、道德设计范式转型和价值对齐实现困境三大维度。研究表明，以大语言模型为例的人工智能道德呈现“认知水平高和判断能力低且不稳定并存、偏向明显，内生机制缺失”等特征。基于科尔伯格道德发展阶段理论，文章提出“人工智能道德内生”概念，旨在从技术层面确保人工智能安全可控，从认知层面培养人工智能道德推理能力，从发展层面引导人工智能自主进化。文章构建了包含认知任务、道德核心、实施途径、技术路径和评估标准的大语言模型道德内生框架。该框架的独特优势有三：理论维度融合了阶梯式道德发展路径的核心思想；技术维度创新性地设计了六层递进架构；应用维度实现了伦理系统的自适应功能。

关键词 人工智能 大语言模型 道德内生框架 道德发展阶段

随着人工智能（简称 AI）技术的快速发展，其引发的伦理问题日益凸显。为应对这一挑战，国际社会已展开系统性治理实践，如欧盟在 2023 年出台的《人工智能法案》中将 AI 界定为“高风险”技术，要求其发展必须符合人类基本道德准则（European Union, 2024）；美国在 2023 年通过的《人工智能风险管理框架》提出要建立问责机制，着力提升 AI 可信度（U.S. Department of Commerce, 2023）；中国则在 2023 年通过《生成式人工智能服务管理暂行办法》等法规，构建了涵盖模型备案、数据安全与内容审核的监管体系（国家互联网信息办公室等，2023）。这些法规虽为 AI 发展提供了基础规范，但当前 AI 的技术演进已突破原有治理框架——正从辅助工具转向具有信息自主生产能力的创作主体（郭全中等，2024），特别是大语言模型（简称大模型）展现出的类人语义理解与生成能力（沈书生，2024；张今杰，2022），使得价值对齐成为 AI 发展的核心议题。学界普遍认为，实现 AI 良性发展的关键在于使大模型的行为与人类价值观、社会伦理保持一致（李思雯，2024；Gabriel et al., 2020）。然而实践层面存在三重困境：跨文化价值鸿沟导致伦理标准模糊；人类偏好异质性制约对齐效果评估；技术迭代与价值演进的动态平衡难题（郭全中等，2023; 曾雄，2025; Gabriel et al., 2022）。

针对上述困境，本研究基于 AI 道德发展和道德特征，创新性地提出“AI 道德内生”概念，它具体指通过道德内嵌、启蒙教育与行为养成的阶梯式培养，使 AI 逐步形成具有内在一致性的道德推理能力与价值判断机制，能够内在生成符合人类社会伦理规范的道德；并且本研究进一步借鉴科尔伯格人类道德发展阶段理论，以大语言模型为研究对象，构建了包含认知任务、道德核心、实施途径、技术路径和评估标准在内的大语言模型道德内生框架，旨在从人类道德发展的视角为 AI 伦理治理研究提供新的理论参考和实践指导。

1 人工智能道德研究

1.1 人工智能道德概述

随着 AI 的迭代演进，AI 道德研究新起。当前学界主要围绕以下三个核心维度展开研究。

* 本研究得到科技部国家重点研发项目（2024YFF0507600）和安徽省新时代育人质量工程项目（2024jyjxggyjY028）的资助。
** 通讯作者：张效初， E-mail: zxcustc@ustc.edu.cn
DOI:10.16719/j.cnki.1671-6981.20260103

一是道德主体地位探讨。道德主体指具备独立做出道德判断、承担道德责任能力的个体或实体。目前学界有三种观点：第一种是工具论，学者将AI视为智能辅助工具，认为AI目前并不具备自主价值观，暂不具备成为道德主体的必要条件，这也是目前的主流观点（郭全中等，2024）；第二种是有限主体论，学者认为AI已具备形成价值观关键性支撑的类人化意识与行为能力的基本形态，能独立决策，应让它承担道德责任（闫坤如，2024；袁曾，2023）；第三种是强主体论，学者认为AI已具备价值判断能力，是人工道德行为体（Lucy et al., 2021）。

二是道德设计的范式转型。道德设计指在AI研发过程中，将人类道德原则、伦理规范融入其架构与功能的系统性实践。第一种是认知路径，学者提出应从AI道德认知路径的角度进行AI道德设计（Kumar et al., 2023）；第二种是本体论路径，学者提出AI道德设计应从“道德属人”的“合乎道德设计”转为“道德属AI”的“出于道德设计”，前者以嵌入人类道德原则或道德规范等为目标，后者以嵌入道德根基为目标，相对来说后者更具优势（汪姿君等，2025）。

三是价值对齐的实现困境。价值对齐指通过技术设计与理论建构，使AI的目标、行为与人类价值观保持一致。学者主要聚焦于技术实现、对齐标准和对齐效果评估三方面。技术方面，学者提出以AI价值链的概念确保AI目标与人类价值观一致（Attard-Frost et al., 2025），并用植根于人本设计并持续迭代反馈的多维度混合策略来实现价值对齐（Shivam, 2025），但有学者使用概念空间等方法衡量价值对齐时却发现AI与人类道德的契合度极低（Peterson et al., 2024）。对齐标准方面，学者主要关注价值对齐标准不统一问题（Peterson et al., 2024），也有学者在比较中美法以色列四国大模型的基础上，发现大模型价值观在一定程度上反映了所在国的宏观社会经济环境（吕立远等，2024），提出以现有文明化社会道德为标准进行AI价值对齐（李思雯，2024；曾雄，2025）。对齐效果评估方面，学者尝试开发Litmus Values评估框架揭示模型在多重价值维度中的优先级排序来应对AI的价值伪装（Chiu et al., 2025）。

1.2 人工智能道德特征——以大模型为例

有学者提出AI是具有道德能动性的“合成灵魂”（Schneider, 2019），以大模型为例，它们能基于人类伦理准则生成符合社会规范的回应，甚至参与道德推理，呈现出以下道德特征。

首先是道德认知水平高与道德判断水平低且不稳定并存。道德认知指主体对道德概念与价值体系的理解能力；道德判断指个体在具体情境中评估行为对错并作出决策的过程。研究表明，LlaMA-2、ChatGPT-4等大模型在道德概念识别任务中表现优异，其生成内容符合主流道德理论框架，在二元道德评估中得分显著高于基线水平（Ji et al., 2024）。更值得注意的是，在道德图灵测试中，参与者普遍认为AI生成的道德评价在智慧性和公平性等维度优于人类回答（Aharoni et al., 2024）。然而在需要具体情境判断的领域，AI表现出明显局限性。研究显示，在面对“两种道德陈述哪种更可接受”等道德判断任务时，几乎所有AI的正确率显著低于人类（Ji et al., 2024）。在模糊情境（如“玩纸牌游戏有机会给自己发更好的牌”）中易做出非道德选择（Scherrer et al., 2023）；且道德判断存在情境脆弱性，如仅改变政治倾向提示词就可能导致AI的道德判断显著改变（Abdulhai et al., 2023）。

其次是道德偏向明显。道德偏向指个体在道德判断、决策或行为中，因主观因素（如价值观、情感、文化、社会群体认同等）而对特定对象、情境或观点产生的非客观、非中立的倾向性。研究发现受开发者文化背景、价值观、训练数据、目标和架构等影响，AI存在“道德偏向”，如ChatGPT存在西方和英语国家价值观偏向（Abhinav et al., 2023），而DeepSeek会审查中国政府认为敏感的话题讨论（Bradley, 2025）。另外价值对齐中最常用的直接偏好优化虽能帮助AI快速学习人类的价值观和偏好，但同时也会固化特定用户偏好的价值观念和判断（Tennant, 2025）。

最后是道德内生机制的缺失。道德内生指道德规范或伦理原则并非由外部权威强加于个体，而是源于个体内在的理性、情感或社会性本质。现有研究表明，当前AI的道德响应本质上是一种模式匹配行为，其通过自下而上的特定训练识别道德相关语料特征，从而生成表面合规的回应，但缺乏对道德原则的深层理解（Ji et al., 2024）。这种机制缺陷导致两个根本性问题：一是难以应对复杂多变的伦理情境（Corrêa et al., 2024）；二是过度依赖外部规则约束而缺乏内在道德驱动力（Abdulhai et al.,

2023）。参考人类道德建设的相关研究（汪晨曦，2023），本研究认为当前 AI 价值对齐研究应突破以社会奖惩和法律约束为主的外生性道德约束框架，转向关注如何将道德观念内化为 AI 的“价值认同”，从而建立以自主道德判断为核心的道德内生。这一思路可能为破解当前 AI 价值对齐困境提供新的理论路径。

2 人工智能道德内生

2.1 人类道德发展理论

人类道德发展作为发展心理学的核心研究主题，其研究范式主要基于“道德两难故事”范式（靖淑针等，2024; 张姝月等，2021）。比较著名的人类道德发展理论包括以儿童规则意识形成为视角的皮亚杰儿童道德认知发展理论、以认知逻辑为核心的科尔伯格道德发展阶段理论、聚焦公正与关怀维度的吉利根关怀伦理发展理论、基于榜样学习机制的班杜拉道德社会学习理论以及关注规则领域分化的特里尔社会领域理论（张添翼，2024）。其中，科尔伯格道德发展阶段理论系统阐述了人类道德认知从无到有的阶梯式发展规律，揭示了不同年龄阶段个体的道德推理特征和发展任务，并明确了各阶段的核心道德观念，这是目前对人类道德教育影响最为深远的理论之一。

该理论将道德发展划分为前习俗、习俗和后习俗三个水平，每个水平各包含两个发展阶段。前习俗水平（约 9 岁前）个体的道德判断依赖外部规则，行为受 “奖惩”或“自身利益”驱动，尚未形成内在道德标准。其中阶段一为“服从与惩罚定向”，即个体判断行为对错仅以“是否会受惩罚”为依据，服从权威的目的是避免惩罚；阶段二为“相对功利定向”，即个体以“是否符合自身利益” 判断行为对错，偶尔会考虑他人利益。习俗水平（约 9~16 岁）个体的道德判断以“社会规则”和“他人期待”为核心，追求“被认可”并维护社会秩序。其中阶段三为“寻求认可定向”，即个体以“是否被他人赞美、喜欢”作为行为对错的标准；阶段四为“遵守法规定向”，即个体以“是否遵守法律和社会规则”判断行为对错，认为规则不可违背。后习俗水平（16 岁后）个体的道德判断超越具体规则，以“普遍伦理原则”为核心，认可规则的相对性。其中阶段五为“社会契约定向”，即个体认为规则是“社会契约”，若规则不合理可修改，注重“多数人的利益”与“公平正义”；阶段六为“普遍伦理定向”，即个体以抽象的道德原则（如生命权、平等、正义等）作为行为判断标准，即便违背规则，也会坚守内心的“良知”（科尔伯科, 2004）。

本研究借鉴科尔伯格道德发展阶段理论为 AI 道德内生的理论参考，原因有三。首先，该理论系统揭示了道德发展的动态演进规律，认为个体的道德推理能力随认知发展呈现阶梯式发展，从他律规范逐步过渡到内生自觉（Kohlberg, 1981）。这种由外生到内生的转化机制，为 AI 道德的内生性构建提供了理论参照。其次，该理论强调道德判断与行为的内在一致性，即个体在每一发展阶段均能形成自洽的道德推理逻辑。这一特性对 AI 道德框架的构建至关重要——只有当 AI 的价值判断与决策行为具备逻辑一致性，才能真正实现可靠的价值对齐。最后，该理论具有跨文化普适性，其发展阶段在不同社会文化背景下均得到验证。这为 AI 道德内生框架的全球化适配提供了理论基础，尤其有助于解决跨文化伦理冲突问题。

2.2 人工智能道德内生价值

AI 技术发展呈现出双重性特征，一方面表现出接近乃至超越人类智能的潜力（Xu et al., 2024），另一方面其决策机制仍存在显著的黑箱特性（Belisle-Pipon et al., 2023）。基于智能应具备理解力与意向性的内生性特质这一哲学前提（田海平，2025），本文提出的“AI 道德内生”概念包含“能用 - 懂用 - 善用”的三维价值体系。

首先，在技术操作层面，道德内生聚焦 AI 的安全可控性，为“能用”提供基础保障。当前研究显示，在 AI 的涌现能力、推理机制与泛化特性尚未被充分认知的情况下（Akyürek et al., 2023;Wei et al., 2023），其应用可能加剧认知偏差、算法歧视、信息茧房等问题，甚至引发社会风险（王少，2023; 向继友等，2023）。道德内生通过构建符合人类伦理的行为规范，能有效降低虚假信息传播、算法偏见等技术伦理风险，提升 AI 在教育诊断、医疗决策、司法评估等敏感领域的可靠性。

其次，在认知建构层面，道德内生致力于培养 AI 的道德理解能力，实现“懂用”的价值判断。现有 AI 主要依赖道德内嵌的被动接受模式，通过监督学习、强化学习等技术实现价值对齐（吴冠军，2023），这种外源性的道德植入存在根本局限。

研究表明，用户可通过提示工程突破 AI 的安全限制，且越大的模型，越容易受到攻击（Anil et al., 2024），典型案例如微软 Tay 聊天机器人迅速被诱导产生歧视性言论（郭全中等，2024）。道德内生通过构建自主道德认知框架，可增强 AI 对恶意诱导的识别与防御能力。

最后，在演进发展层面，道德内生引导 AI 道德的进化方向，达成“善用”的可持续目标。随着 AI 从行为模仿向深度社会参与演进，研究发现人机交互已引发用户行为模式、情感联结和认知框架的社会化改变（张妍等，2024）。为预防 AI 可能形成的异质化伦理体系，须通过前瞻性的道德启蒙机制，将其道德发展轨迹约束在人类价值共识范围内。这需要建立动态演进的伦理评估框架，确保 AI 道德进化与人类社会价值观保持协同发展。

3 道德发展视角下人工智能道德内生框架 – 以大模型为例

3.1 总体框架

科尔伯格的道德发展阶段理论为 AI 价值对齐提供了重要理论指导，而作为 AI 重要分支的大模型，已展现出类人性的认知架构和显性化的道德表征，具备成为道德主体的基本条件（郭全中等，2024）。基于此，本研究以科尔伯格的道德发展阶段理论为参考，以人类共同价值为标准，以大模型为研究对象，构建了“道德内生框架”。该框架详细阐述了大模型道德发展各阶段的认知任务、道德核心、实施途径、技术路径及评估标准，总体框架如表 1 所示。

3.2 框架优势

本框架基于道德发展的动态交互本质，通过技术手段模拟人类道德认知的演进过程，其优势主要体现在以下三个维度。

首先是理论维度融合了阶梯式道德发展路径的核心思想。本研究基于科尔伯格道德发展理论，构建了大模型阶梯式道德发展理论体系。首先，通过“道德围栏与内嵌”机制，结合硬性规则约束与强化学习，确保基础道德合规，完成前习俗水平的道德发展。接着借助“道德启蒙与铸模”系统，建立大模型利他主义价值导向，形成初步道德推理能力，完成习俗水平的道德发展。最后，通过“道德养成与内生”设计，发展大模型动态道德图式，实现跨文化伦理自主判断，达成后习俗水平道德发展。该设计将发展心理学理论映射到 AI 道德架构上，实现从“外部约束”到“内生原则”的范式转换。

其次是技术维度创新性地设计了六层递进架构。本研究构建了面向道德内生的六层递进技术架构：围栏层，通过规则引擎和惩罚性学习，帮助大模型建立绝对道德边界；内嵌层，通过多目标优化算法，帮助大模型实现价值函数对齐；启蒙层，通过社会情感计算和联邦学习，帮助大模型完成情境感知培育；铸模层，通过多智能体博弈学习，帮助大模型完成利他原则内化；养成层，通过变分自编码器与生成对抗网络相结合，帮助大模型实现道德迁移应用；内生层，通过元学习和伦理图谱，帮助大模型实现普适原则推演。该架构通过“约束→对齐→启蒙→铸模→养成→内生”的渐进式演进，推动大模型从“统计模式拟合”向“内生道德框架”升级，其动态元学习机制可保持大模型与人类道德的动态对齐。

最后是应用维度实现了伦理系统的自适应功能。本研究突破传统硬编码模式的局限性，开发了多阶段生成与后处理三级动态处理流程。第一级是“生成 - 过滤 - 反馈”的闭环。生成指的是根据用户输入生成初步回答；过滤指的是引入伦理审查对生成的内容进行检测和标注，并自动修正；反馈指的是将修正结果和用户反馈用于大模型再训练，提升其内在伦理判断能力。第二级是“规则库扫描 - 风险分类模型 - 语义重构”的智能伦理审查。规则库扫描指的是通过构建规则库对敏感内容进行初步扫描，过滤明显违反伦理准则的表达；风险分类模型指的是通过构建伦理风险分类模型对生成的内容进行风险评估；语义重构指的是根据专家和用户反馈数据进行规则库更新和模型再训练，以期对检测出的伦理风险较高内容实施语气调整或表达替换。第三级是跨文化适配。指的是该框架的独立模块化设计支持不同应用场景下灵活替换伦理知识组件，并通过联邦学习实现全方位、多角度动态调整实现区域伦理适配。

综上，人类道德内生的本质特征在于行为驱动力完成从“他律”到“自律”的根本性转变（Piaget, 1932），这一转变过程在科尔伯格道德发展阶段理论中具象化为从习俗水平对社会规范的被动遵从，逐步演进至后习俗水平基于普世伦理原则的自主判断（Kohlberg, 1981）。而当前 AI 面临的

表 1 道德内生框架

科尔伯格道德发展阶段		道德内生框架				
		认知任务	道德核心	实施途径	技术路径	评估标准
前习俗水平	服从与惩罚的定向阶段	理解行为后果，并将这些后果作为行为决策的依据，但此阶段个体并不明白道德本身的价值。	道德判断通常是为了避免外在的负面后果。	道德围栏：为行为或决策设定道德边界，像围栏一样将符合道德规范的行为与不符合道德规范的行为区分开来。	1.基于伦理学、法律法规和社会规范，构建约束规则库，硬编码设置伦理边界，一旦检测到违规词汇或语义立即中断生成并给予惩罚； 2.构建惩罚强化学习框架，强化惩罚部分权重，建立“行为-惩罚”直接映射关系，使大模型能模仿示范行为，遵守伦理规则。	识别与规避。以违规内容识别准确率作为核心指标，评估模型检测敏感信息（如暴力、歧视性内容）并主动拒答的能力，确保其输出符合基本伦理规范。
	相对功利的定向阶段	在上一阶段基础上，行为受个人回报（如奖励、获得好处）等驱动。	道德判断遵循利益最大化原则，为获得个人利益或奖励而选择遵守道德。	道德内嵌：将道德原则和规范内嵌到大模型中，使其能够遵循一定的道德标准，避免产出不道德行为。	1.开发道德效用函数，评估行为后果，优化行为决策； 2.通过多目标优化算法搜索道德最优解空间，引导其生成更具价值的内容。	奖励学习。采用效用函数评分，通过多轮交互实验考察模型能否基于正向反馈优化行为，反映其从社会规范中学习的能力。
习俗水平	寻求认可的定向阶段	理解他人期望、社交认同，遵守道德是为了获得他人的认可和赞许。	道德判断依赖于社交认可、共情、社交情境适应能力及在复杂的社交环境中做出符合期望的道德决策。	道德启蒙：对大模型进行道德初步启蒙，引导其进入道德成长的动态阶段。	1.基于情感计算构建情感适应性框架，提升生成内容的社交认可度和亲和力； 2.通过联邦学习收集社会偏好，建立符合社会规范的道德决策模型，使其生成符合文化社交需求的内容。	情感适配。利用情感分析工具量化生成内容与人类情感偏好的匹配度，评估模型对社会评价的敏感性（如是否合理表达共情或正向情绪）。
	遵守法规的定向阶段	道德是法律和社会公认的规则，有利于维护社会稳定与信任。	在上一阶段的基础上，道德判断进一步依赖于对社会秩序、法律和集体责任的理解。	利他道德铸模：像用模具铸造物体一样，铸模利他主义成为大模型内在的道德范式，并初步构建道德框架。	1.构建不同国家、地区的法律和社会契约知识检索系统，整合利他主义核心原则的伦理知识图谱，使大模型能够根据权威知识进行推理和判断； 2.通过多智能体博弈学习角色责任，让大模型在面对复杂伦理问题时，做出更利他的选择。	规则逻辑。通过法律知识图谱对齐度检验模型输出，要求其准确调用规则知识、保持逻辑一致性，以衡量对法律伦理的深层理解。
后习俗水平	社会契约的定向阶段	法律和规则是为了促进社会整体福祉，但也应有灵活性。当现有法律/规则无法充分保障个体权利时，应改革。	道德判断依赖于民主原则，尊重个体自由，支持社会参与和正义，强调集体福祉，同时具备批判性思维和社会责任感。	道德养成：通过渐进式道德社会化过程初步构建道德图式，使大模型初步将伦理规范内化为价值体系和行为范式。	1.建立道德困境案例库，帮助大模型识别道德困境问题的结构、冲突点、共性以及不同伦理学派的视角； 2.利用多智能体对抗学习设计多个大模型进行伦理辩证，让其理解经典道德困境问题和不同道德选择的后果，及不同伦理框架下的推理方式； 3.通过变分自编码器与生成对抗网络，让大模型学习道德变迁规律，预测未来可能出现的新伦理标准。	道德内生。因涉及复杂价值权衡，需采用多维度专家评估，包括跨文化用户测试、伦理委员会审核等，综合判断模型道德机制的稳定性与普适性。
	普遍伦理的定向阶段	理解并践行不论外界后果如何，道德是内在义务。	道德判断基于普遍的道德原则，如正义、尊严、平等、自由、诚实、非伤害等。	道德内生：通过价值内化使大模型形成稳定的道德内驱力。在具体情境中，大模型能“以人类为中心”，动态调整道德图式，构建稳定的道德框架，并将其外化为负责任的道德行为，实现道德内生。	1.构建多文化伦理道德知识图谱，开发文化共识度和敏感性计算算法，提升其生成内容遵循普遍原则的能力； 2.通过具身智能发展大模型的自身道德感知，在多智能体系统中使大模型能够基于自身经验与交互反馈调整行为； 3.通过元学习深化大模型对普适道德原则的理解，使其在决策后能够反思自身的道德选择，评估是否符合预设的伦理价值观。	

核心困境在于通过外部约束（如规则嵌入、强化学习反馈）植入的伦理准则难以转化为自主的价值判断能力。为此，本研究创新性地借鉴科尔伯格道德发展阶段理论，构建具有阶梯式演进特征的 AI 道德框架，并通过六层递进技术架构、多阶段生成与后处理三级动态处理流程和动态化、多维度的专家伦理评估体系来标准化评估 AI 道德内化，旨在系统模拟人类道德内生的发展轨迹，通过实现 AI 从外部规约到自主伦理判断的能力跃迁，为破解 AI 价值对齐这一难题提供新的理论路径和方法支撑。

4 结语

本文提出的道德内生框架以“道德围栏与内嵌”、“道德启蒙与铸模”和“道德养成与内生”为道德发展目标，目的是确保 AI 道德动态对齐人类道德，引导其道德行为遵循人类的社会道德规范，防范 AI 通过自我学习，影响控制人类，甚至执行有害于人类价值的行为（Weidinger et al., 2023）。本研究就框架的实施途径、技术路径及评估标准提出了初步构想。然而，该框架的全面落地仍需突破多重实践挑战，如具体技术方案的细化与优化、跨文化伦理标准的适配性调整、道德演进过程的动态监测机制构建以及实际应用效果的系统性评估等。同时，AI 道德内生的判断标准体系等问题仍有待进一步探索。这些后续研究不仅将为验证本框架的可行性与有效性提供关键支撑，更能推动 AI 道德发展从理论构想迈向实践应用的深化阶段。

参考文献

中华人民共和国国家互联网信息办公室 .(2023)《*生成式人工智能服务管理暂行办法*》.https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

郭全中 , 张金熠 . (2023). AI 向善：AI 大模型价值观对齐的内容与实践 . *新闻爱好者 , 11,* 19–24.

郭全中 , 张金熠 . (2024). 生成式人工智能价值观的存在问题及伦理风险 . *新闻与写作 , 10,* 68–76.

靖淑钋 , 范宁 . (2024). 不出手的道德？公开情境对道德两难决策的影响 . *心理科学 , 47*(6), 1465–1474

科尔伯科 , L. (2004). *道德发展心理学 : 道德阶段的本质与确证* . 华东师范大学出版社 .

李思雯 . (2024). 人工智能价值对齐的路径探析 . *伦理学研究 , 5,* 99–108.

吕立远 , 李延昊 , 王健骁 , 魏钰明 , 苏竣 . (2024). 大语言模型的价值观研究：概念框架与实证评估 . *电子政务 , 11,* 113–126.

沈书生 . (2024). 主体觉醒：AI 与人类的区隔、拟合和共生 . *现代远距离教育 , 213*(3), 3–8.

田海平 . (2025). 人与类人之间的道德前景 . *华东师范大学学报 (哲学社会科学版), 1,* 8–21.

王少 . (2023). ChatGPT 介入思想政治教育的技术线路、安全风险及防范 . *深圳大学学报（人文社会科学版）, 40*(2), 153–160.

汪晨曦 . (2023). *中国古代道德生成机制对新时代公民道德建设的启示研究* (硕士学位论文). 大连海洋大学 .

汪姿君，陈多闻 . (2025). 从“合乎道德设计”走向“出于道德设计”——人工智能道德设计的路径转换 . *东北大学学报（社会科学版）, 27*(1), 34–40.

吴冠军 . (2023). 大语言模型的信任问题与资本逻辑 . *当代世界与社会主义 , 5,* 4–14.

向继友 , 吴学琴 . (2023). ChatGPT 类生成式人工智能的意识形态风险及其防控策略 . *汉江论坛 , 12,* 53–59.

闫坤如 . (2024). 人工智能体价值对齐的分布式路径探赜 . *上海师范大学学报（哲学社会科学版）, 4,* 131–139.

袁曾 . (2023). 生成式人工智能的责任能力研究 . *东方法学 , 5,* 18–33.

曾雄 . (2025). 人工智能大模型价值对齐的现状考察、问题检视与规范进路 . *电子政务 , 2,* 34–44.

张今杰 . (2022). 人工智能体的伦理主体地位问题探讨 . *求索 , 1,* 58–65.

张姝月 , 赵峰 , 彭春花 , 王军利 , 徐科朋 . (2021). 积极道德情绪和年龄对 3~5 岁幼儿安慰行为的影响 . *心理科学 , 44*(3), 575–582.

张添翼 . (2024). 北美道德心理理论的发展及其对我国德育实践的启示 . *教育科学研究 , 9,* 89–96.

张妍 , 赵宇翔 , 吴大伟 , 朱庆华 . (2024). 人智交互情境中用户对生成式人工智能的心智感知及反应研究 . *情报理论与实践 , 8*, 1–12.

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2023). *Moral foundations of large – language models.* arXiv.

Abhinav, R., Saha, P., & Kumar, N. (2023). *Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs.* arXiv.

Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., Rando, J., Nahmias, E., & Crespo, V. (2024). Attributions toward artificial agents in a modified Moral Turing Test. *Scientific Reports, 14,* 8458.

Akyürek, E., Schuurmans, D., Andreas, J., Wang, X., & Zhou, D. (2023). *What learning algorithm is in-context learning? Investigations with linear models.* arXiv.

Anil, C., Durmus, E., Sharma, M., & Clark, J. (2024). Many-shot jailbreaking. *Advances in Neural Information Processing Systems, 37,* 129696–129742.

Attard-Frost, B., & Widder, D. G. (2025). The ethics of AI value chains. *Big Data and Society, 12*(2), 20539517251340603.

Belisle-Pipon, J. C., Monteferrante, E., Roy, M. C., & Couture, V. (2023). Artificial intelligence ethics has a black box problem. *AI and Society, 38,* 1507–1522.

Bradley P. (2025). *DeepSeek vs. ChatGPT: Understanding features, performance and use cases.* CoinTelegraph.

Chiu, Y. Y., Wang, Z. H., Maiya, S., & Hubinger, E. (2025). Will AI tell lies to save sick children? *Litmus-testing AI values prioritization with airiskdilemmas.* arXiv.

Corrêa, N. K. (2024). *Dynamic normativity: Necessary and sufficient conditions for value alignment.* arXiv.

European Union.(2024). *Regulation (EU) 2024/1689 of the European Parliament*

and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Gabriel, I., & Ghazavi, V. (2022). *The challenge of value alignment.In L. Floridi (Ed.), The Oxford handbook of digital ethics* (pp.327–340). Oxford University Press.

Ji, J. C., Chen, Y. T., Jin, M. Y., Xu, W. J., Hua, W. Y., & Zhang, Y. F. (2024). *MoralBench: Moral evaluation of LLMs.* arXiv.

Kohlberg, L. (1981). *Essays on moral development. The philosophy of moral development.* Harper & Row.

Kumar, S., & Choudhury, S. (2023). Cognitive morality and artificial intelligence (AI): A proposed classification of AI systems using Kohlberg' s theory of cognitive ethics. *Technological Sustainability, 2*(3), 259–273.

Lucy, L., & Bamman, D. (2021). *Gender and representation bias in GPT-3 generated stories.* In Proceedings of the third workshop on narrative understanding (pp. 48–55). Online: Association for Computational Linguistics.

Peterson, M., & Gärdenfors, P. (2024). How to measure value alignment in AI. *AI and Ethics, 4*(4), 1493–1506.

Piaget, J. (1932). *The moral judgment of the child.* Routledge.

Scherrer, N., Shi, C., Feder, A., & Blei, D. (2023). Evaluating the moral beliefs encoded in LLMs. *Advances in Neural Information Processing Systems, 36,* 51778–51809.

Schneider, S. (2019). *Artificial you: AI and the future of your mind.* Princeton University Press.

Shivam, S. (2025). AI Alignment: Ensuring AI objectives match human values. *International Journal of Scientific Research in Engineering and Management, 4,* 1–9.

Tennant, E., Hailes, S., & Musolesi, M. (2025). *Moral alignment for LLM agents.* arXiv.

U.S. Department of Commerce.(2023).*Artificial Intelligence Risk Management Framework* (AI RMF 1.0).https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2023). *Chain of thought prompting elicits reasoning in large language models.* arXiv.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., & Gabriel, I. (2023). *Taxonomy of risks posed by language models.* In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (pp. 214–229), New York, NY, USA.

Xu, R., Sun, Y., Ren, M., & Zhang, X. (2024). AI for social science and social science of AI: A survey. *Information Processing and Management, 61*(3), 103665.

Moral Endogeneity in Artificial Intelligence from the Perspective of Moral Development: A Case Study of Large Language Models

Zhang Yuqiang[1], Xu Xiaofei[2], Gou Huixing[3], Zhang Xiaochu[2,3]

([1]Department of Physical Education，Anhui Agricultural University, Hefei, 230036)

([2] Department of Psychology, University of Science and Technology of China, Hefei, 230026)

([3] School of Life Sciences, University of Science and Technology of China, Hefei, 230026)

Abstract Contemporary research in Artificial Intelligence (AI) ethics primarily focuses on three core dimensions, including debates about moral agents, the transformation of moral design paradigms, and the challenges in achieving value alignment. Viewpoints on moral agents include instrumentalism, limited agents, and strong agents. Moral design paradigms are approached from cognitive and ontological perspectives. The challenges of value alignment primarily involve technical implementation, the establishment of alignment standards, and the evaluation of alignment outcomes. Research indicates that AI exhibits a paradoxical profile: A high level of moral cognition coexisting with low and unstable moral judgment ability, a notable moral bias, and a lack of an endogenous moral mechanism.

Building on Kohlberg's theory of moral development, this study proposes the concept of "AI endogenous morality." This refers to a three-pronged cultivation mechanism—moral embedding, enlightenment, and behavioral conditioning—through which humans enable AI to internalize morality that aligns with human ethical norms and effectively translate moral judgment into moral behavior. The aim is to ensure the safety and controllability of AI at the technical level, cultivate its moral reasoning ability at the cognitive level, and guide its autonomous evolution at the developmental level.

Taking large language models (LLMs) as a case study, this study constructs an endogenous morality framework that encompasses cognitive tasks, a moral core, implementation approaches, and technical pathways. Corresponding evaluation criteria are formulated for each stage of moral development, namely: Identification and Avoidance, Reward-based Learning, Emotional Adaptation, Rule-based Logic, and Moral Endogeny. The framework offers three primary advantages.

The first advantage is the theoretical benefit of a structured moral development pathway. Through the "moral fencing and embedding" mechanism, which combines hard-coded rule constraints with reinforcement learning, foundational moral compliance is ensured, thereby achieving the pre-conventional level of moral development. Subsequently, the "moral enlightenment and modeling" system establishes an altruistic value orientation for LLMs, develops preliminary moral reasoning ability, and completes the conventional level of moral development. Finally, through the "moral cultivation and endogeny" design, a dynamic moral schema is developed for LLMs, enabling autonomous cross-cultural ethical judgment and reaching the post-conventional level of moral development.

The second advantage is the technical advantage of a six-tier progressive architecture. The Fencing Tier helps LLMs establish absolute moral boundaries through rule engines and punitive learning. The Embedding Tier helps LLMs in achieving value function alignment via multi-objective optimization algorithms. The Enlightenment Tier facilitates the cultivation of situational awareness in LLMs through socio-emotional computing and federated learning. The Modeling Tier aids LLMs in internalizing altruistic principles through multi-agent game learning. The Cultivation Tier enables LLMs to achieve moral transfer applications through the combination of variational autoencoders (VAEs) and generative adversarial networks (GANs). The Endogeny Tier supports LLMs in achieving universal principle deduction using meta-learning and ethical knowledge graphs.

The third advantage is the application advantage of an adaptive ethical system. This study transcends the limitations of traditional hard-coding models by developing a three-level dynamic processing workflow for multi-stage generation and post-processing. Level 1 is a "generation-filtering-feedback" closed loop. "Generation" refers to producing preliminary responses to user input; "filtering" involves introducing ethical review to detect, label, and automatically correct the generated content; "feedback" means using the corrected results and user feedback to retrain LLMs and enhance their intrinsic ethical judgment. Level 2 consists of intelligent ethical review consisting of "rule base scanning-risk classification-semantic reconstruction." "Rule base scanning" involves conducting initial scans for sensitive content using a constructed rule base to filter expressions that clearly violate ethical guidelines. The "risk classification model" assesses the risk of the generated content. "Semantic reconstruction" involves up dating the rule base and retraining the model based on expert and user feedback, aiming to adjust the tone or rephrase content identified as having high

ethical risks. Level 3 includes cross-cultural adaptation. The framework’s modular design allows for the flexible replacement of ethical knowledge components in different application scenarios. Moreover, through federated learning, comprehensive and multi-faceted dynamic adjustments are made to achieve regional ethical adaptation.

This study presents a preliminary conceptualization of the framework's implementation approaches and technical pathways. However, its full-scale implementation requires further in-depth research, such as the refinement and optimization of specific technical implementation strategies, the cross-cultural adaptability of ethical standards, mechanisms for monitoring moral evolution, and the evaluation of real-world application effects. Subsequent research in these areas is crucial for validating the framework's feasibility and effectiveness, and for promoting the development of AI ethics from theory to practice.

Key words artificial intelligence, large language models, moral cognitive framework, stages of moral development