Advances in Education 教育进展, 2026, 16(3), 246-252
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.163477

Hans汉斯

面向负责任创新的人工智能专业人才伦理素养培养体系研究

李欣盈，张　蕾*，魏楚元*

北京建筑大学智能科学与技术学院，北京

收稿日期：2026年1月25日；录用日期：2026年2月24日；发布日期：2026年3月3日

摘　要

人工智能在高风险领域应用引发的伦理挑战，凸显专业人才“伦理素养”与“技术能力”融合培养的紧迫性。针对高校人工智能专业教育“重术轻道”、伦理与技术培养割裂的症结，本研究通过文献分析与对北京市建筑类高校师生调研，系统诊断现有教育模式在体系性、实践性、评价机制及师资支撑方面的结构性缺陷，创新性提出“三维四柱”一体化培养模型，通过“目标－内容－方法”三维教学框架与“资源－师资－平台－评价”四维支撑系统的深度耦合，构建递进式伦理素养培养路径。以《机器学习》课程“算法公平性”单元为例，设计融合教学方案实现伦理原则向技术实践的转化。研究为人工智能专业伦理教育系统化改革提供了兼具理论创新性与实践可操作性的解决方案，其核心目标是培育兼具技术能力与伦理自觉的“负责任人工智能架构师”，为人工智能技术的健康发展筑牢人才基础。

关键词

人工智能，伦理素养，负责任创新，培养体系

Research on an Ethical Competence Development System for AI Professionals Committed to Responsible Innovation

Xinying Li, Lei Zhang*, Chuyuan Wei*

School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing

Received: January 25, 2026; accepted: February 24, 2026; published: March 3, 2026

*通讯作者。

文章引用: 李欣盈, 张蕾, 魏楚元. 面向负责任创新的人工智能专业人才伦理素养培养体系研究[J]. 教育进展, 2026, 16(3): 246-252. DOI: 10.12677/ae.2026.163477

Abstract

The ethical challenges arising from artificial intelligence applications in high-risk domains underscore the urgency of integrating ethical literacy with technical proficiency in professional training. Addressing the core issues within university AI education—namely an overemphasis on technical skills at the expense of ethical principles, coupled with a disconnect between ethical and technical cultivation—this study employs literature analysis and surveys of faculty and students at Beijing-based architecture universities to systematically diagnose structural deficiencies in existing educational models concerning systemic coherence, practical application, assessment mechanisms, and faculty support. It innovatively proposes a "three-dimensional, four-pillar" integrated cultivation model. This model establishes a progressive ethical literacy development pathway through the deep integration of a "goal-content-method" three-dimensional teaching framework with a "resources-faculty-platform-evaluation" four-dimensional support system. Taking the "Algorithmic Fairness" module within the Machine Learning course as a case study, an integrated teaching scheme is designed to translate ethical principles into technical practice. This research offers a solution for the systematic reform of ethics education in artificial intelligence programmes, combining theoretical innovation with practical applicability. Its core objective is to cultivate "Responsible AI Architects" who possess both technical competence and ethical awareness, thereby laying a robust talent foundation for the healthy development of artificial intelligence technology.

Keywords

Artificial Intelligence, Ethical Literacy, Responsible Innovation, Training System

1. 引言

人工智能技术在医疗、司法、雇佣等重要领域里的深入应用产生了算法偏见、隐私侵害、责任不清等一系列的伦理问题，从而引发出“可信赖”“负责任创新”的成为全球科技治理的重要议题。人才是技术创新的源头，人的伦理素养决定着人工智能技术的发展方向以及它所具有的社会价值。但是目前我国高校人工智能专业教育存在着“重术轻道”的现象，培养方案大多重视技能训练和算法优化，对于技术的社会影响、价值判断以及伦理选择等素养的培养不够重视[1]，造成人才供给和社会需要之间的差距明显。由于存在价值盲区，使一部分毕业生虽然有展示技术功底的能力，但是由于伦理认识不足、批判性思维缺乏、责任感不强等原因，不能很好地解决工作中出现的伦理难题，反而会加大技术应用的社会风险。因此，摆脱传统附加式伦理说教，创建起一个贯穿于人工智能专业人才培养全过程、深度融合伦理考量和技术实践的系统化素养培育体系，已经成为新工科建设背景之下落实国家科技伦理治理战略、推进人工智能专业教育改革的根本任务。本文以此命题为基础，通过对人工智能伦理素养核心内涵的界定、对目前培养模式深层缺陷的剖析、提出一体化培养框架并给出可操作的实践路径来为高校实现由“技术训练”向“价值引领”的教育转型提供理论依据和实践方案。

2. 人工智能伦理素养的内涵与培养现状审视

2.1. 人工智能伦理素养的多维内涵解析

为了建立严密的理论框架，本文首先对两个主要的概念做出界定。其一就是负责任的创新，它是研

发过程中合法性的体现，也是一种前瞻性、反思性、包容性的创新方式。重视在技术研究与开发的开始阶段以及整个生命周期内，把广泛的社会伦理考量、社会价值和社会利益主动地、系统地融入到设计、评价和决策当中，使创新成果更加符合社会期待并为长远福祉服务，避免产生潜在的负面效应。第二，“人工智能伦理素养”是人工智能专业人才应该具备的工作社会影响所需要的复合型品质，包括意识维度、知识维度、能力维度和行为维度这四个互相联系的部分。在意识方面要具有伦理敏感性和责任感，也就是能提前发现技术活动中可能存在的伦理风险点；在知识上需要系统地掌握公平、透明、问责、隐私、安全等主要的伦理原则以及相关的法律法规、行业规范；能力方面则集中在伦理分析、决策权衡、沟通辩护等方面的能力上；最后一切都要内化为行为层面的自觉遵守伦理准则，并且有进行“伦理化设计”的能力。这四个维度互相联系、层层推进，构成起“负责任创新”的行为素养基础。

2.2. 当前培养模式的核心问题诊断

基于上述多维内涵，反观当前培养模式，可以发现其存在若干深层困境。首要问题是体系性缺失，伦理教育多以孤立的选修课或零星讲座形式存在，未能与数据结构、机器学习等核心课程体系有机融合，更未形成贯穿本科四年的连贯培养主线，呈现“碎片化”特征。其次是知行脱节，教学内容聚焦抽象伦理理论与哲学讨论，与学生日常接触的算法设计、模型训练、工程项目等技术实践严重脱节；教学方法以单向灌输为主，缺乏基于真实产业伦理困境的深度研讨、角色扮演等互动式教学形式。第三是评价机制空白，尚未建立科学的伦理素养评估体系，缺乏衡量学生伦理意识觉醒、伦理决策能力提升的有效工具与标准，导致教学效果无法量化、教学改进缺乏方向。最后是支撑体系薄弱，一方面，许多学生在智能化环境中伦理认知薄弱，面对这些复杂的伦理问题很难展开深刻剖析并做出高效回应[2]；另一方面，师资存在跨学科协同困境，技术类教师普遍缺乏伦理教学训练，而伦理学者又对技术细节把握不深，难以实现伦理与技术的深度融合教学。这些问题共同导致了当前伦理教育效果不彰，难以支撑“负责任创新者”批判思维能力的培养目标。

3. “三维四柱”伦理素养培养体系的整体构建

针对当前培养模式中存在的体系性缺失与实践脱节等核心问题，本研究提出一个系统化的解决方案——“三维四柱”式伦理素养培养体系[3]。该体系旨在超越将伦理教育作为附加模块的传统思路，转而将其深度嵌入人工智能专业教育的基因之中。

其构建遵循四大核心理念：一是价值引领，把“负责任创新者”作为人才培养的终极画像，伦理素养由软性要求变成硬性标准；二是全程贯穿，保证伦理教育不是某个学期单独进行的课程，而应该深深地融入到每一个学科之中，在学生心中留下伦理问题的重要印记；三是领域融合，伦理议题要同机器学习、自然语言处理、计算机视觉等具体的科学技术领域深度结合，构成有领域特色的教学内容，防止出现伦理和科技“两张皮”的情况；四是有能力导向，教学设计的最终目的就是培养出具备解决实际复杂伦理问题能力的人才，而不是只是记住一些伦理教条。

“三维”教学框架和四柱支撑系统之间形成的是一个深度耦合的闭环，即目标维度的阶段化设计要依靠评价反馈柱来动态校准，在内容维度上课程融合需要依靠课程资源柱来提供案例、工具支持，在方法维度上四阶教学法的实施需要师资发展柱和实践平台柱共同保证，三者互相配合，保证培养体系的可落地性和持续性。

3.1. “三维”教学框架的系统化设计：目标、内容与方法的融合

“三维”教学框架是按照教育学核心要素来对教学目标、教学内容和教学方法做一体化、系统化的

规划，目的是解决培养什么、怎样培养以及怎样用以培养的问题。

3.1.1. 目标维度：递进式阶段化设计

根据学生的认知发展规律和专业学习进度，把伦理素养培养总目标分解成可观测、可评估的四个阶段子目标，实现本科四年伦理素养递进式提高。具体来说，大学一年级主要是伦理感知和意识启蒙，目的是用震撼性的案例和行业事件来引起学生对于技术社会影响的初始关注，培养起最初的伦理责任感；二年级是伦理知识体系化的建构，目的就是使学生系统地了解公平、透明、问责、隐私等主要的伦理原则，熟悉相关的法律规范和行业标准，形成伦理判断的知识结构；三年级重在伦理分析决策能力的提升，目的在于让学生能够在模拟或者真实的项目情境(比如人脸识别系统的开发)里，用伦理分析的方法来进行技术方案风险评价、多样的利益平衡，并提出有针对性的改进意见；四年级重视的是伦理综合实践和价值塑造，在毕业设计、企业实习等综合性实践中，能够独立完成技术方案的伦理影响评价和伦理化设计，促使负责任创新的价值观内化为职业信仰。该递进式目标体系给课程内容的设计和教学效果的评价提供了一条清晰的途径。

3.1.2. 内容维度：技术和伦理深度融合

其核心任务是打破伦理内容与技术课程的壁垒，将伦理与技术相融合，在每一门人工智能专业课中加入相应伦理的内容知识。例如，在《机器学习》课程中，数据采集与预处理章节自然融入数据隐私与知情同意议题；模型训练与评估章节重点引入算法公平性与偏见检测内容；模型解释章节则对应可解释AI 与透明性原则。在《计算机视觉》课程中，目标检测与人脸识别技术必须结合监控伦理、隐私权与社会公平的讨论。为支持上述融合教学，可开发动态更新的“领域伦理案例库”，每个案例都包含技术背景、伦理冲突、相关原则、讨论问题与延展阅读，为教师提供即取即用的教学素材。

3.1.3. 方法维度：四阶教学实现能力跃迁

我们设计出一条“启发、探究、协同、实战”的四阶教学法链[4]，来实现由被动接受向主动建构能力的飞跃。第一阶段：案例启发。用现实中有伦理冲突的真实案例来激发学生对道德直觉和批判性质疑。第二阶段：课题探究。通过小组形式，让学生就某个具体伦理技术课题(比如设计出一个更公平的信贷评价算法)进行文献综述、技术调研和伦理分析，最后形成初稿的研究报告。第三阶段：协同辩论。组织课堂辩论或者角色扮演，模拟出一个算法伦理审查委员会，学生分别扮演技术工程师、伦理学家、产品经理、用户代表等角色，对某一技术方案的伦理可行性展开辩论和辩护，锻炼学生的多角度思考能力和伦理论证能力。第四阶段：项目实战。在课程设计或者创新项目中，要让学生明白伦理考虑是设计的一部分，在完成需求分析、方案设计、开发实现和伦理评价之后，还要提交包含技术文档和独立的伦理评估报告的完整作品。这四种方法相互联系、互相配合，一起构成能力本位的教学目标。

3.2. “四柱”支撑保障系统的协同建设

一个先进的教学框架要有一个坚实的支持系统才能落地生根、持续运行。“四柱”支撑保障系统从资源供给、人力保证、实践场所与评价反馈四个方面给三维教学框架赋予全方位的制度性支撑，两者配合构成闭环式的培养链条。

3.2.1. 课程资源柱：建设开放共享的融合式教学资源库

资源库是教学内容与方法革新的物质基础。教学资源开发通过编制人工智能伦理教育案例库、思辨话题集等教学素材，为教师提供即拿即用的伦理教育工具包，是伦理教育实施的资源支撑。其建设应包括：1) 模块化伦理教学案例库：收录覆盖各技术领域的经典与前沿伦理案例，每个案例配备教学指南、

讨论题、相关代码或数据接口。2) 虚拟仿真实验平台：开发“算法公平性审计沙箱”等在线实验模块，支持学生导入或生成数据集，训练模型，并使用多种公平性指标进行量化评估，直观呈现偏见缓解策略的实施效果。3) 标准化教学工具包：提供伦理影响评估模板、负责任 AI 设计检查清单以及开源伦理分析工具集等实用工具。这些资源应以在线平台形式开放共享，促进校际协作与动态更新。

3.2.2. 师资发展柱：打造跨学科的教学共同体

师资能力是体系实施的核心瓶颈。为破解“教师不会教、不愿教”的困境，核心举措是建立稳定的“人工智能专业教师 + 科技伦理学者 + 行业实践专家”三元教学共同体。教师培训体系通过伦理案例工作坊、教学设计实训等模块，提升教师识别伦理问题、设计伦理教学活动的能力，是伦理教育实施的人才保障。例如，通过“算法偏见识别”工作坊，培训教师如何在编程教学中引导学生识别并修正算法偏见，实现从理论到实践的能力转化[5]。同时，学校应将教师在伦理教学方面的投入与成果，纳入教学绩效考核、职称评聘体系，给予实质性激励。

3.2.3. 实践平台柱：构建多层次、沉浸式的实践场域

伦理素养必须在实践中养成。应构建三个层次的实践平台：1) 校内基础平台：设立“负责任 AI 创新实验室”，配备必要的算力与软硬件，支持学生开展算法公平性、可解释性、隐私计算等方向的探索性实验。2) 校内高阶平台：模拟成立“学生项目伦理审查委员会”，让学生扮演委员，对同学的研究计划或竞赛项目进行伦理审查，体验治理流程。3) 校外协同平台：与科技企业或研究机构共建实践基地，让学生进入真实的产品研发流程，在资深工程师和伦理官的指导下，参与实际项目的伦理风险评估与设计优化，直面产业一线最紧迫的伦理挑战。

3.2.4. 评价反馈柱：创建基于证据的多元化评价体系

科学评价是体系持续改进的导航仪。必须改变单一论文考核方式，建立过程性、表现性、多元化的评价体系。该体系应收集多种证据：过程性证据，如学生在项目中的伦理讨论记录、迭代日志；表现性证据，如其在模拟伦理审查会上的发言质量、在团队中协调伦理冲突的表现；成果性证据，如技术报告中的伦理分析章节深度、设计的伦理缓解方案创新性。可引入“伦理素养成长档案袋”，整合上述证据，并利用图表进行可视化分析，动态描绘学生素养发展轨迹。评价结果不仅用于评分，更应提供个性化反馈，指导学生持续改进。

4. 关键实施路径与教学案例验证

任何宏大的体系构想都需要清晰的实施蓝图与具体的实践验证。本章旨在将第三章的理论框架转化为可操作的行动路线，并通过一个深入的教学设计案例，展示体系如何在实际课堂中落地生根，从而论证其可行性与有效性。

4.1. 分阶段、渐进式的实施路线图

为确保“三维四柱”体系能够平稳、有效地落地，需要一个审慎而务实的三阶段实施路线。1) 试点阶段：由院系统筹，选择《人工智能导论》《机器学习》2 门核心课程作为试点，组建“技术教师 + 伦理学者 + 行业专家”跨学科教学团队，完成课程大纲修订、首批融合教学案例开发，并对授课教师开展专项培训，打造可复制的“样板课程”；2) 推广阶段：将试点课程的成功经验推广至《计算机视觉》《自然语言处理》等主干课程群，建成院系级共享教学资源库，形成稳定的跨学科协作机制，并将伦理实践纳入专业实习考核；3) 成熟阶段：在人工智能专业全面落地该体系，形成制度化的培养方案，同步向学校其他工科专业辐射，并联合兄弟院校共建资源共享平台，打造可推广的教育模式。

4.2. 示范性教学单元深度剖析：以《机器学习》“算法公平性”议题为例

为具象化“三维四柱”体系的实操路径，本研究以《机器学习》课程中至关重要的“算法公平性[6]”专题为例，设计一个可供直接使用的示范性教学单元。

本单元的教学目标是：学生能够理解算法公平性不同定义(如群体公平、个体公平)的社会与数学内涵；掌握至少两种公平性量化指标(如人口统计均等差、均等机会差)的计算方法；并能在给定数据集和简单模型上，完成一次基础的公平性审计，同时能对“公平–性能”权衡进行伦理反思。评估方式则采用多元综合评估法：1) 过程性评估(40%)：课堂研讨参与度、实验操作规范性；2) 成果性评估(60%)：实验报告(含技术实现 + 伦理分析章节)、课堂辩论表现，重点考核公平性认知深度与伦理决策逻辑。

教学过程设计有五个环环相扣的环节。环节一：案例锁定、问题点燃。以招聘算法性别歧视、医疗风险评估模型种族偏见等真实的案例作为切入点。用案例视频、新闻报道等材料来激起学生对伦理的关切和探究的动机。环节二，对概念进行解析并引入相关工具。教师对公平性的核心定义、统计均等、机会均等、预测率均等不同的公平性指标进行系统的讲解，并给出相应的数学表达式以及它们所对应的政治诉求，另外还介绍了用于审计公平性的 Python 工具库。环节三：动手实验与数据对话。学生按照有敏感属性(比如性别、种族编码等)的数据集分成小组，用逻辑回归或者决策树的二分类预测模型来训练该模型。随后他们要分别计算出模型的整体准确率、召回率和选定的公平性指标，在整体上和各个敏感属性子群体上，直观地发现算法偏见的存在形式，尝试通过调整决策阈值、数据预处理等方法来缓解偏见。环节四：深入讨论和价值权衡。这就是教学最高的境界。组织课堂讨论或者小组辩论，针对核心矛盾：当通过阈值调整、算法干预提高模型对弱势群体的公平性不可避免地造成模型整体准确率或者其他性能指标下降的时候，应该怎样做决策？决策的根据就是这些信息。刑事司法、医疗信贷等领域是否需要不同的公平性标准[7]？环节五：综合报告和反思内化。要求学生提交一份详细的实验与反思报告。报告不仅要包含数据描述、模型选择、代码、结果图表等技术部分，还要有单独的伦理分析章节。本章要求学生描述实验过程中发现的公平性问题，分析造成这些问题的社会原因和影响，论证在“公平–性能”权衡上做出的选择理由，并讨论在实际应用中还需要考虑哪些非技术因素。通过完整的教学闭环，学生可以将抽象的伦理原则转化为技术实践中自觉的决定，达到伦理素养和技术能力同步提高的目的。

4.3. 预期成效与挑战应对

4.3.1. 预期成效

预期实施“三维四柱”培养体系将实现三重核心成效：其一，学生层面，伦理敏感性与技术伦理决策能力显著提升，能在技术实践中主动融入伦理考量，形成“技术 + 伦理”的复合型思维；其二，教学层面，产出一批可复制、可推广的融合式教学案例与资源库，推动人工智能专业课程从“技术导向”向“价值引领”转型；其三，行业层面，为社会输送兼具技术功底与伦理自觉的“负责任人工智能架构师”，助力人工智能技术健康可持续发展。

4.3.2. 核心挑战与应对策略

然而，体系落地仍面临三大核心挑战，需针对性制定应对策略：其一，针对融合式课程设计、案例开发等带来的教师工作负荷增加、改革动力不足的问题，可通过建立伦理教学工作量专项核算机制、搭建校级共享教学资源库、推行跨学科团队协作教学等举措，减少重复劳动、分摊工作压力；其二，面对技术类教师、伦理学者与行业专家跨学科协作松散、易出现“临时拼凑”的困境，需成立校级人工智能伦理教育教学中心作为常设协作机构，建立定期研讨与项目绑定机制，以联合申报教改项目、编写教材等方式强化协作黏性；其三，针对学生伦理素养长期成效难以量化的问题，应构建“短期评估 + 长期跟

踪”的立体化体系，通过课程报告、实践项目评分完成短期考核，依托企业合作建立毕业生职场伦理决策随访机制，结合伦理素养成长档案袋与第三方独立测评[8]，动态优化培养方案。

5. 结语

面对人工智能发展带来的深刻伦理挑战，高校的人才培养范式必须进行系统性变革。本文所构建的“三维四柱”培养体系，试图将伦理素养从边缘化的“附加选项”转化为人工智能专业教育的“核心维度”。通过目标、内容、方法的系统重构与资源、师资、平台、评价的协同支撑，该体系为培育能够驾驭技术复杂性、进行负责任创新的新一代人才提供了可行的框架。在未来的深化方面，可以考虑如何使该体系动态适应技术的快速迭代，如何开发更科学的素养评价工具，以及如何构建校企社政协同的育人生态。通过教育层面的价值重塑，为人工智能的健康可持续发展筑牢人才与伦理的根基。

基金项目

教育部人文社会科学研究规划一般项目(22YJAZH110)、北京市教育科学十四五规划重点课题(CHAA22061)、北京建筑大学校级重点教研项目(Y2118)。

参考文献

[1] 刘宇, 钟葉, 贾棋, 王维民. 生成式人工智能背景下科技伦理课程思政[J]. 实验室科学, 2025, 28(4): 193-198.

[2] 冯界山. 人工智能时代大学生数智伦理素养培育研究[J]. 行车指南, 2025(2): 241-243.

[3] 莫宏伟, 樊赵兵. 研究生课程“人工智能原理与方法”思政教学方法研究[J]. 大学(教学与教育), 2024(9): 108-111.

[4] 刘淑艳. 基于布鲁姆教育目标分类驱动的“四元互动”教学模式构建[J]. 北京教育(高教), 2025(10): 59-61.

[5] 陈宏斌. 人工智能伦理教育的实施路径与评估体系构建[J]. 中国信息技术教育, 2025(19): 12-15.

[6] 胡子睿. 面向敏感信息缺失场景的机器学习算法公平性研究[D]: [硕士学位论文]. 北京: 中国科学技术大学, 2024.

[7] 严顺. 算法公平问题及其价值敏感设计的解法[J]. 伦理学研究, 2024(2): 101-109.

[8] 孙菊芬, 范彬. 记录成长足迹实施立体评价——基于“学生成长·数字档案袋”的综合素养立体评价方法[J]. 浙江考试, 2022(6): 20-23.