人工智能伦理与社会责任教育在“人工智能概论”课程中的融合路径研究

陈芦园，李　春

（南京林业大学信息科学技术学院&人工智能学院，南京 210037）

摘　要：随着人工智能技术的迅猛发展，其引发的伦理与社会问题日益突出，高校人工智能教育亟须从“技术导向”向“价值驱动”转型。文章聚焦“人工智能概论”课程，系统分析当前伦理教育在理念引导、内容设计、教学方法与评价机制等方面存在的不足，指出伦理教育多处于“附属地位”且实践性不足的问题。为此，提出“四位一体”的融合路径：(1)重构课程目标，强化价值引领；(2)推动技术与伦理内容协同；(3)创新教学方法，强化学生参与和批判性思维；(4)完善多元评价机制，推动伦理素养全过程培养。研究旨在为高校构建“以人为本、技术与人文融合”的人工智能人才培养体系提供理论支持和实践路径。

关键词：人工智能伦理；人工智能概论；课程改革

中图分类号：G642　　**文献标识码：**A

Integration pathway of artificial intelligence ethics and social responsibility education in “introduction to artificial intelligence” course

CHEN Luyuan，LI Chun

（College of Information Science and Technology & College of Artificial Intelligence，Nanjing Forestry University，Nanjing 210037，China）

Abstract：With the rapid development of artificial intelligence technology，the ethical and social issues arising from it are becoming increasingly prominent，and a transformation of artificial intelligence education in higher education from a “technology-oriented” approach to a “value-driven” approach is urgently required. The “introduction to artificial intelligence” course is focused on，and the current deficiencies in ethics education in terms of conceptual guidance，content design，teaching methods，and evaluation mechanisms are systematically analyzed. It is pointed out that ethics education is often placed in an “auxiliary position” and is characterized by insufficient practical engagement. Accordingly，a “four-in-one” integration pathway is proposed：(1) curriculum objectives are reconstructed to strengthen value guidance，(2) technical and ethical content is integrated in a coordinated manner，(3) teaching methods are innovated to enhance student participation and critical thinking，and (4) diversified evaluation mechanisms are improved to promote the whole-process cultivation of ethical literacy. The study is intended to provide theoretical support and practical pathways for the development of a people-oriented artificial intelligence talent cultivation system that integrates technology with the humanities in higher education.

Key words：artificial intelligence ethics，introduction to artificial intelligence，course reform

1　引言

近年来，人工智能技术迅猛发展，广泛应用于医疗健康、金融科技、司法审判、教育辅助、军事决策等领域，深刻改变了人类社会的运行方式。同时，人工智能发展过程中引发的伦理与社会问题日益显著，成为公众与政策制定者关注的焦点。例如，深度伪造(deepfake)技术带来的信息可信度危机[1]，削弱了公众对媒体与事实的

基金项目：中国高等教育协会“2024 年高等教育科学研究规划课题”重点项目：面向新工科的农林特色人工智能交叉人才培养模式探索与实践(24NL0203)；江苏省本科高校“2024 年高等教育数字化转型与教育现代化实践研究”专项重点课题：信息技术基础课与思政教育的数字化融合探索与实践研究(2024CXJG024)

信任；自动驾驶系统在突发情况下所面临的道德困境，对既有责任归属机制提出了挑战；人工智能对传统职业岗位的替代，引发了就业结构调整；人工智能模型在训练和应用过程中对个人数据的广泛使用，也可能引发严重的隐私侵犯问题。在此背景下，国际社会和科技界纷纷呼吁将伦理价值观念系统融入人工智能技术的设计、开发与应用中，强调在追求技术进步的同时守住伦理底线，确保人工智能服务于人类福祉。2021 年，联合国教科文组织发布了《人工智能伦理问题建议书》，明确提出应发展“以人为本、具有包容性、可持续性”的人工智能技术；欧盟委员会提出“值得信赖的人工智能”原则，要求人工智能系统在合法性、道德性和技术稳健性方面达到统一标准；我国通过《新一代人工智能发展规划》《科技伦理审查办法（试行）》等政策文件，将人工智能伦理建设提升至国家战略高度，强调在自主可控、安全可用的基础上兼顾公平正义、隐私保护和社会责任等关键维度[2]。

高校作为人工智能专业人才培养的重要阵地，其课程设置和价值导向不仅影响学生对技术的理解与应用方式，也会潜移默化地塑造其价值判断、伦理意识和社会责任感。然而，目前多数人工智能概论类课程仍以技术逻辑为主导，伦理教育缺乏系统性和实践性，在教学评价中的权重较低，难以引起足够重视。这种“重技术、轻伦理”的模式，容易导致学生对人工智能的社会影响缺乏深入认识，也难以有效应对技术发展带来的复杂挑战[3]。

因此，将人工智能伦理与社会责任教育有机融入“人工智能概论”课程，是人工智能教育顺应时代发展的必然选择，也是构建以人为本的科技育人体系的重要环节。本文从理念引导、教学内容重构、教学方法创新与评价机制完善等方面入手，推动课程从“技术导向”向“价值驱动”转型，提升学生的伦理判断能力和社会责任意识，为高校构建技术与人文融合的人才培养体系提供理论依据及实践路径。

2 人工智能伦理与社会责任教育现存的问题

围绕理念引导、教学内容、教学方法与评价机制 4 个方面，当前高校人工智能伦理与社会责任教育仍存在以下问题。

2.1 缺乏以价值导向为核心的理念引导

虽然国家政策层面日益强调“科技向善”与“以人为本”的技术价值观，但在高校教育实践中，这一理念尚未充分转化为课程建设的内在目标与日常教学逻辑。部分院校未能将人工智能伦理教育纳入人工智能课程体系的核心，仍将其视为附属内容，缺乏系统推进意识与长远规划。例如，国内部分高校的人工智能课程教学大纲并未明确将“人工智能伦理素养”设定为教学目标，也未将伦理议题纳入教学评价与课程考核体系，缺乏对学生伦理认知的系统引导与培养机制。以南京林业大学信息科学技术学院为例，其“人工智能概论”课程教学大纲侧重人工智能基础知识、典型算法和应用实践的讲授，未对人工智能伦理素养、社会责任意识等内容作出明确规定，也未在课程目标中充分体现价值导向要求。这表明，伦理理念尚未真正融入部分高校课程建设的顶层设计与执行机制。与之相比，国外部分高校已将人工智能伦理素养纳入课程主线，并在人才培养目标中予以强调，这一做法与国内部分高校伦理教育相对缺位的现象形成一定差距[4]。

总体而言，国内高校在人工智能课程中的伦理教育理念引导仍显不足，存在定位不清、顶层设计不足与实践落实不够等问题。这说明人工智能伦理教育在理念引导层面仍处于相对边缘的位置，尚未形成系统化、制度化的育人逻辑。

2.2 教学内容缺乏结构化与系统性设计

部分高校“人工智能概论”课程在内容设计上仍以技术导向为主，偏重基础算法、模型结构和应用实践讲解，人工智能伦理相关内容呈现“零散化、点状化”特征，缺乏与技术知识体系并行的伦理知识架构。伦理内容往往以“临时插入、随讲随议”的方式出现，难以支撑系统的伦理认知与能力培养。以国内某高校“人工智能概论”课程为例，伦理议题仅作为第 14 讲的选讲内容，既缺乏与核心技术模块的横向关联，也缺少从伦理理念到伦理判断的纵向知识递进，导致课程内部逻辑衔接不足。这种“技术与伦理割裂”的设计缺陷，在部分高校中具有一定普遍性。相比之下，斯坦福大学 CS181“Computers, Ethics, and Public Policy”课程通过构建较为完整的伦理教学结构，系统讲授算法偏见、技术公平等议题，并配套政策文件、伦理分析框架与案例工具包，使学生能够在现实情境中综合评估技术影响，体现了伦理教学的系统性与实践性。

国内部分高校已开始探索伦理模块嵌入，如清华大学“人工智能伦理与治理”选修课引入技术法规、伦理原则及国际比较视角，初步搭建伦理教育的理论基础，但由于属于选修课程，尚未实现对全体学生的系统覆盖，教育效能仍受到一定限制。总体而言，当前人工智能课程伦理内容设计仍存在统筹不足、体系不够完善、课程定位偏弱等问题，导致伦理教学在部分高校中仍处于边缘位置。

2.3 教学方法创新不足

“人工智能概论”课程的教学方法仍以传统讲授式为主，教师侧重知识输出，学生处于相对被动的接受状态，缺乏对伦理议题的主动探究与深度参与。具体表现如下：一方面，教学过程中缺乏引导学生开展跨学科协作、模拟伦理决策的互动设计，如伦理困境辩论、真实场

景推演等实践环节较少，导致伦理教育容易停留在理论讲授层面；另一方面，部分教师受自身伦理素养或跨学科知识背景的限制，难以有效组织案例研讨、情境模拟等创新教学活动，进一步制约了学生伦理思维的培养。

与之相比，国外高校已探索多元化的教学模式。例如，牛津大学“Digital Ethics Lab”课程通过情境再现、角色扮演、伦理沙盒等形式，引导学生在模拟人工智能伦理困境中开展决策推演，强化实践应用能力；斯坦福大学相关课程则采用“案例研讨+政策分析”的教学方式，结合具体技术场景分析伦理冲突，提升学生的批判性思维。国内高校中，复旦大学“人工智能与社会”课程尝试设置模拟人工智能道德审查会、小组剧本讨论等实践项目，中国人民大学也在课程中引入伦理反思报告机制，但这类创新主要集中于少数院校，尚未形成较为成熟、可推广的教学范式，难以覆盖全体学生的伦理能力培养需求。

总体而言，当前教学方法存在单一化和实践性不足的问题，导致伦理教育难以充分激发学生的深层认知与价值认同，成为制约教育效果的重要因素。

2.4　教学评价机制不完善

“人工智能概论”课程的评价体系仍以知识性考核为核心，普遍存在对学生伦理理解能力、责任意识与价值判断等维度评价不足的问题。传统评价方式（如闭卷考试、技术项目报告等）主要考查学生对算法原理、模型结构等技术知识的掌握程度，却难以衡量其在伦理情境中的分析能力与责任意识。部分课程虽尝试引入伦理写作任务，但由于评价标准较为模糊、形成性评价工具不足，难以系统追踪学生伦理素养的发展过程。

在国际教育实践中，部分高校已探索通过多维度记录与评估学生伦理素养发展的方式，将学生价值立场变化、案例分析能力和团队协作表现等作为评价内容，通过纵向记录与多维度评价动态呈现伦理素养的养成过程。此外，牛津大学、麻省理工学院等院校将伦理辩论表现、情境决策模拟等实践环节纳入考核体系，实现“知识掌握—能力应用—价值内化”的评价衔接。

国内高校虽已开始探索多元评价路径，如中国人民大学信息学院在课程中引入伦理反思报告与小组互评机制，复旦大学尝试通过实践项目成果评价学生的伦理应用能力，但这类探索仍处于相对零散的实践阶段，尚未形成较为完善的评价框架。总体而言，当前评价机制的滞后导致伦理教育效果难以得到有效反馈，也难以持续引导学生形成系统的伦理责任认知。

3　人工智能伦理与社会责任教育在“人工智能概论”课程中的融合路径

为实现人工智能伦理与社会责任教育在“人工智能概论”课程中的有效嵌入及深度融合，课程设计需从目标设定、内容构建、教学方法和评价体系等维度进行系统优化。通过在教学全过程中渗透伦理教育，可逐步实现从技术知识传授向伦理思维培养的教学转变，进而培育学生的科技责任感与人文素养。

3.1　课程目标重构：从“知识导向”走向“价值引领”

在课程目标设定层面，应从传统以知识掌握和技术能力为主的导向，转向“技术+伦理”双维育人的价值导向。教师在制定课程目标时，不应仅限于传授人工智能基本原理、主流算法与应用框架，还应明确提出“理解人工智能的伦理挑战”“形成初步社会责任意识”“具备在技术方案中融入伦理考量的能力”等目标要求。这种转向不仅回应了人工智能技术快速发展带来的社会挑战，也契合新时代对高素质复合型人才的培养需求，有助于增强课程目标设计的前瞻性和系统性。在这一背景下，课程目标的重构不仅强化了人工智能教育的社会责任属性，也为教学内容与方法改革奠定了基础，确保课程培养的学生不仅会“造工具”，也能够理解“如何负责任地使用工具”。

具体而言，应将人工智能伦理素养作为课程核心能力指标之一，明确其在教学大纲、课程标准和能力评价中的地位。目标设计应体现“以学生为中心”的理念，引导学生主动思考技术与社会的关系，培养其辨析现实伦理问题、表达多元价值立场以及在复杂情境中进行伦理判断的综合素养。例如，教师可以借鉴布鲁姆教育目标分类法，将“伦理认知、伦理理解、伦理应用、伦理分析、伦理评价、伦理创造”等层级融入课程目标体系，形成由浅入深的进阶路径。在实际操作中，可通过设计如下目标声明具体化教学要求：“学生应能识别人工智能技术中的伦理风险与社会影响；分析人工智能系统中的价值冲突与责任归属；撰写伦理评估文档，提出具有可行性的伦理改进建议；在小组项目中融入伦理原则以进行人工智能应用设计。”基于此，课程目标不仅具有较强的可操作性，也为学生学习成果评价提供了清晰框架。

3.2　教学内容设计：技术与伦理的结构性协同

实现伦理素养与技术能力的协同发展，需要在教学内容设计上构建系统化结构。教师可以采用“嵌入式+独立模块”的混合方式：一方面，将伦理问题自然融入技术知识点教学；另一方面，设置专门的伦理专题教学单元，保障伦理教学的系统性与相对独立性。在嵌入式设计中，教师可在讲授关键技术内容时同步引入相关伦理议题。例如，在人脸识别技术教学中讨论隐私侵犯与人权保障；在推荐算法模块中引导学生思考“信息茧房”现象；在生成式人工智能应用案例中分析深度伪造可能带来的虚假信息传播问题。这有助于学生将伦理思考融入技术认知，避免陷入“技术中立”的认识误区。同时，教师可以设置2—3学时的独立模块，系统讲授人工智能伦理的核心概念（如责任归属、公正性、透明性等），以及

伦理评估框架(如ACM伦理准则、IEEE相关建议),并介绍中外法规政策(如欧盟人工智能法案、《科技伦理审查办法(试行)》等)。通过知识体系化设计,帮助学生建立从伦理理念到实践判断的认知逻辑链。

此外,课程应引入国内外典型案例进行深度剖析(如“微软Tay机器人失控事件”“人工智能招聘系统性别偏见案例”),并结合小组合作撰写“伦理评估报告”或“项目社会影响分析”,以任务驱动方式促进学生在真实语境中形成伦理判断。在案例遴选上,应兼顾不同领域与多元背景(如无人机在冲突地区使用引发的人道主义辩论等),使学生能够在跨文化、跨情境的比较分析中提升伦理敏感度。同时,教师可以配合多模态资源(如视频访谈、政策白皮书与交互式数据可视化),帮助学生从数据证据和利益相关者视角分析技术影响,形成对伦理问题的整体理解。

3.3 教学方法创新:参与式与探究式教学的有机结合

教学方法创新应以提升学生的伦理思辨能力、协作意识与现实责任感为核心。相较于传统的灌输式讲授,教师应更加注重启发式、互动式与探究式教学策略,营造学生主动参与的教学氛围。可重点采用以下方法:首先,教师可以通过案例研讨法,引导学生围绕真实争议性事件开展小组调研与辩论,如“人工智能算法是否应参与司法量刑”“人工智能是否应具备情感识别能力”等,让学生在多元观点中辨析问题、形成自己的判断。其次,教师可以设计情境模拟法,构建“算法决策引发歧视投诉”“自动驾驶紧急避险伦理难题”等角色扮演场景,训练学生在不确定条件下进行价值判断。同时,建立跨学科协同机制,邀请法学、社会学、心理学等领域教师共同参与课程设计与教学,开展“技术+伦理+社会”三位一体的项目式教学。教师可指导学生以跨专业团队形式完成“伦理可解释性系统设计”或“人工智能应用伦理合规审查报告”等实践任务。

此外,教师可以引入教学辅助工具(如伦理沙盒平台、虚拟模拟审议系统等)增强学习体验;鼓励学生参与“伦理播客”“视频公民讲堂”等公开表达项目,增强传播能力和社会参与意识。同时,可在课程后期引入“逆向设计”环节,让学生基于已有人工智能项目主动寻找潜在伦理风险,并提出可行的改进方案,从而强化批判性分析能力和实践责任意识。此外,教师可设计“跨年级导师制”小组,由高年级学生指导低年级学生完成伦理设计任务,通过同伴教学促进知识内化与经验传递,构建持续性的学习共同体。通过多种教学方法的有机融合,可进一步促进学生深度思考与伦理认知内化。

3.4 教学评价机制建设:从结果考核到过程育人的转型

评价机制不仅决定学生学习成果的衡量方式,也会反向影响其学习行为和价值取向。因此,构建符合人工智能伦理教育特点的评价体系,应注重“知识—能力—态度”三维融合,实现从“知识测验”向“伦理素养成长”的转变。在知识层面,教师可以设计以问题为导向的开放式测验,如伦理困境选择题、短文分析题等,考查学生对伦理原则、法规政策的理解与应用。在能力维度,应重点评估学生能否在实际项目中识别伦理风险、提出改进建议并撰写伦理评估文档等,可通过“技术伦理分析报告”“小组共建伦理设计方案”等成果进行评价。在态度与反思层面,可以引入自评与互评机制,通过“学习反思日志”“价值冲突日记”“伦理自省问卷”等方式,促使学生表达个人观点、审视自身成长。为保证评价的公平性与多元性,应建立“教师评价—学生互评—自我评价”三位一体的评价结构,并给予阶段性评价意见和形成性反馈。同时,评价应保持开放性与发展性,不仅关注学生当前的表现结果,还应记录其在课程周期内的持续改进轨迹和思维变化过程,从而促进其形成长期的伦理反思与自我提升意识。

具体而言,教学评价体系包括5个核心维度。(1)知识理解:主要通过开卷或闭卷测试,系统考查学生对人工智能伦理基本理论、核心概念、相关法律法规及政策框架的掌握情况,重点评价其基础认知水平与概念准确性,占总成绩的30%,作为课程学习的知识基础。(2)项目实践:通过学生提交伦理分析报告,或在技术项目中融入伦理设计的完整方案,评价学生的技术理解能力以及将伦理原则应用于实际问题解决的能力,占总成绩的20%,体现应用导向。(3)课堂表现:包括学生在课堂讨论中的参与积极性、观点表达的深度与逻辑性,以及与同伴的互动情况,占总成绩的20%,强调参与过程与协同意识。(4)价值反思:通过学习反思日志、自我评估报告等方式,引导学生表达伦理立场并深化自我认知,占总成绩的20%,是伦理教育内化的重要体现。(5)同伴评价:通过互评记录与团队协作反馈,体现学生的合作精神以及对他人观点的尊重,占总成绩的10%,有助于培养学生的协作意识与公民责任感。这一五维评价体系覆盖知识掌握、能力应用与价值判断等层面,能够较为全面地反映学生伦理素养的发展情况,为课程目标达成与持续改进提供评价依据。

4 结束语

在人工智能加速发展的时代背景下,培养既掌握核心技术又具备伦理责任意识的综合型人才,已成为高等教育的重要任务。人工智能伦理与社会责任教育不应被视为技术教学的附属内容,而应成为人才培养体系的有机组成部分。本文围绕理念引导、教学内容、教学方法与评价机制4个维度,系统提出人工智能伦理教育融入“人工智能概论”课程的路径与策略,旨在推

(下转第81页)