权责之间:三师课堂中“AI智能师”的责任限度

周娜,刘源
(信阳师范大学教育科学学院,河南 信阳 464000)

摘　要:在“三师课堂”中,AI被冠以“智能师”之名并被赋予介入教学流程的权力,然其责任设定高度依赖人类外部监督,由此形成权责之间的结构性错位。如何在这一错位中为AI智能师划定合理的责任边界,成为一个亟须厘清的前置性问题。研究以法哲学的责任理论为根基,将责任界分为“解释义务”与“后果担当”两种形态,进而构建“主体—关系—功能”三维分析框架,对AI智能师的责任归属展开逐层剖析审视。研究认为,在现有技术条件与伦理前提下,AI智能师只能承担解释义务而无法承担后果担当,其责任形态是与其辅助性功能相匹配的“有限责任”。这一“有限责任”的厘定,既为弥合权责之间的裂隙提供了伦理坐标,亦为“师”之命名提供了合宜的责任根基。随着智能系统愈发具备近似教育者的角色定位,对其责任的讨论须返回责任概念的内在结构,方能实现权责匹配、各归其位。

关键词:三师课堂;AI智能师;教育责任;权责对称;人机协同

Between Power and Responsibility: The Domain of Responsibility for the “AI Intelligent Teacher” in the “Three-Teacher Classroom”

ZHOU Na, LIU Yuan
(School of Education Science, Xinyang Normal University, Xinyang, He'nan 464000)

Abstract: In the “Three-Teacher Classroom,” AI is conferred the title of “AI Intelligent Teacher” and vested with the power to intervene in the instructional process. However, the ascription of its responsibility remains heavily contingent upon external human mediation, thereby engendering a structural decoupling between power and responsibility. Determining an appropriate domain of responsibility for the AI Intelligent Teacher amid this decoupling thus emerges as a foundational question that demands urgent conceptual clarification. Drawing on the responsibility theory of legal philosophy, this study disaggregates responsibility into two modalities: answerability (obligation to account) and liability (bearing of consequences). It then constructs a tripartite analytical framework—agency, relationality, and functionality—to examine, in a graduated

【收稿日期】2026-05-01
【基金项目】河南省教育科学规划2026年度重点课题“人口变动背景下河南省县域教育资源配置的时空演化与空间正义研究”(课题编号:2026JKZD40)。
【作者简介】周娜,信阳师范大学教育科学学院副教授,硕士生导师,研究方向为全球教育史、数字教育治理、县域教育治理;刘源,信阳师范大学教育科学学院硕士研究生,研究方向为数字化教育理论。

manner, the attribution of responsibility for the AI Intelligent Teacher. The study contends that, under extant technological conditions and ethical premises, the AI Intelligent Teacher can only assume answerability; it is incapable of bearing liability. Its form of responsibility constitutes a "delimited responsibility" commensurate with its auxiliary pedagogical function. The delineation of this delimited responsibility not only furnishes an ethical anchor for bridging the power-responsibility gap but also provides a defensible normative foundation for the appellation of "teacher." As technological systems are increasingly entrusted with roles that approximate those of human educators, the discourse on their responsibility must return to the intrinsic structure of the concept of responsibility, so that power and responsibility may each be restored to their proper place.

Keywords : Three-Teacher Classroom; AI Intelligent Teacher; Educational Responsibility; Power-Responsibility Symmetry; Human-AI Collaboration

一、问题的提出

随着人工智能技术与教育领域的持续融合,"三师课堂"作为一种整合人类教师、AI 智能师与学生的协同教学模式[1],逐渐进入理论研究的核心视域。该构想承袭陶行知先生"小先生制"与"共学、共事、共修养"的核心理念,构建了人类教师、小先生与 AI 智能师三方协同的教学生态[2]。其中,将人工智能命名为"智能师",是这一构想中值得深入辨析的理论环节。"师"的称谓并非仅是一种修辞表述,而是意味着 AI 的角色定位发生了转变:它不再是被动响应指令的教学辅助工具,而是具备主动参与教学过程的行动者身份。这一判断已得到国内外相关研究的印证。《中国教师生成式人工智能应用报告(2026)》表明,生成式人工智能已开始进入教学关系内部,而非停留于教学之外[3]。斯坦福大学《2025 年人工智能指数报告》亦将 AI 描述为能够深度参与教学过程、动态适配学习需求的"准主体"[4]。上述研究动向共同揭示了一个值得关注的趋势:AI 正从教学的辅助性工具,转向教学关系的结构性参与者。这一转向,在经验层面为"智能师"的命名提供了参照,也由此引出了一个亟须厘清的理论追问:当 AI 以"师"之名进入教学关系、被赋予行动者身份时,其应当承担何种责任?

已有研究为理解"三师课堂"中的 AI 角色提供了重要基础。周洪宇等系统揭示了 AI 应用可能带来的五重风险,其规避策略隐含了一个前提:AI 的自主决策必须置于人类教师的价值框架之内[5]。然而,这一前提恰恰暴露了一个更深层的问题:当 AI 被赋予"介入"教学流程的权力时,我们依据什么原则来划定其权力的边界?又由谁来为其决策的后果负责?周洪宇、操太圣等借助行动者网络理论分析协同效率,但其对"责任"在人与非人行动者之间如何分配的问题并未展开。吴河江将生成式人工智能界定为"有限教育主体"[6],这是一个重要的理论推进,但"有限"在责任维度的具体内涵,仍有待厘清。上述研究共同揭示了一个值得关注的现象:AI 被赋予了越来越大的行动权力,但对其责任的讨论却停留在风险规避和伦理原则的层面。然而,在法哲学传统中,"责任"从来不是一个单一的概念——它至少应包含"应答性(解释义务)"与"法律责任(后果担当)"两个不可化约的维度[7-8]。既有研究恰恰缺乏对这一概念结构的自觉,导致对 AI 智能师责任的讨论要么流于笼统的伦理呼吁,要么陷入"AI 无法负责所以无需讨论"的沉默。这一概念上的模糊,正是本研究分析的逻辑起点。

上述分析表明,要回应"AI 智能师应承担何种责任"这一追问,关键在于返回"责任"概念本身的内在结构,在"解释义务"与"后果担当"的二分框架中加以判定。研究正是以这一问题为切入点,在"三师课堂"的具体场域中,构建"主体—关系—功能"三维分析框架,对 AI 智能师的责任归属展开递次审视,以期为其"有限责任"划定一个清晰的理论边界。

二、“责任”的两种形态:解释责任与后果担当

本研究要审视 AI 智能师权责的对称性,需厘清其被赋予的权力形态与责任设定,将二者并置对照,方可准确诊断问题所在。

(一)权力的赋予

“三师课堂”赋予 AI 智能师的权力趋于具体且可操作,深度嵌入教学核心流程,体现了 AI 从“辅助者”向“介入者”的角色演变。这些权力大致可归为三个层面。

在知识传递层面,AI 智能师具备提供个性化认知支持的权力。它并非被动等待检索的静态数据库,而是通过构建动态学习者模型,主动为不同认知水平的学生推送分层学习资源[9],在事实上介入了学习内容的界定过程。在学习过程组织层面,AI 智能师被授予主动介入教学互动的权力。它可以扮演苏格拉底式追问者[10]、认知冲突触发者[11]或反例提供者[12]等角色,依托自然语言处理技术,在课堂对话中引发认知扰动,进而影响对话的走向与质量[13]。在学习评价与反馈层面,AI 智能师享有贯穿教学全程的动态评估与诊断权力。它将学习投入、互动质量、概念掌握轨迹等过程性数据转化为可追踪的学习证据,运用学习分析与聚类算法开展精细化学情诊断,生成诊断报告与反馈建议[14]。

整体而言,AI 智能师在知识建构、互动组织与学情评价等层面,获得了较为深入且足以影响教学流程的行动力。当前教育实践中已出现“多智能体协同”的发展趋势。例如,在国家智慧教育平台所构建的“多智能体课堂”中,已内置教学设计师智能体、学情分析师智能体、作业反馈智能体、助教学伴智能体等多种 AI 角色,共同组成“人机协同教学团”。多种 AI 角色的共存意味着责任链条更趋碎片化,从而进一步强化了厘清每种权力所对应责任的现实迫切性。

(二)责任的厘清

与前述具体而深入的权力配置与之形成鲜明对照的是,当前学界对 AI 智能师责任界定仍模糊、缺乏实操边界。其责任并非内生于行动,而是高度依赖人类的外部监督,由此呈现出“权力在内,责任在外”的分离格局。周洪宇等明确提出,AI 的所有自主决策必须规约在教师主导的价值框架内,不得参与“什么是重要的学习”“什么是美好生活”等根本性教育价值的界定[5],这构成了一种前置性的伦理底线。然而,对于 AI 在授权范围内的行为管控,现有框架并未提供清晰的责任归属方案。

1. 两种责任的区分

要解开 AI 智能师责任归属的症结,首先需辨析 AI 应承担何种责任,并对两种性质截然不同的责任加以区分。这一区分在法哲学中已有清晰的界定。安东尼·达夫(Antony Duff)在讨论法律责任与道德责任的关系时,将责任精确地剖分为“应答性”(answerability)与“责任承担”(liability)两个维度[7]:前者追问行动者是否有义务就其行为给出说明,后者追问行动者是否应当承受惩罚或补偿的实质后果。研究在此基础上,结合教育场域的伦理特质,将这一区分进一步界定为“解释义务”与“后果担当”。

第一种责任为“解释义务”。它追问的是“何以如此”(Why is it so?),要求行动者就其行动的依据、过程与结果给出一个可为人类理解的说明。这种责任的核心在于透明性(Transparency)、可追溯性(Traceability),本质上是一种程序性的解释与说明义务。在传统教育场景中,教师向学生、家长或教育管理者阐明其教学设计的理论依据与评价标准的具体构成,或对某位学生做出特定反馈的原因时,所履行的正是此种责任。这一义务并不直接等同于为最终结果的优劣承担奖惩,而在于确保其专业行为建立在可理解、可沟通的理性基础之上。

步入人工智能时代,解释义务对于 AI 系统尤显关键。AI 模型,尤其是深度学习模型,常被视作难以透视的“黑箱”[15]。其决策过程的高度复杂性与非线性,使人类用户难以理解特定预测或推荐得以产生的缘由。这种透明

度的缺失,严重制约了 AI 在教育等高风险、高敏感领域的深度应用,并构成人机信任建立的巨大障碍。正是在这一背景下,可解释性 AI(Explainable AI,XAI)[16]应运而生,其核心目标即破解“黑箱”难题,使 AI 系统的决策与预测人类用户所理解。[17]可解释性 AI 用以弥合复杂模型与人类用户之间的认知鸿沟,通过提供决策依据、关键特征与不确定性信息等内容,履行系统解释义务。在教育场景中,XAI 具有巨大的应用价值:它能为教师提供学习资源推荐的理由,解释学生评价的生成逻辑,甚至揭示算法可能潜藏的偏见,从而为教师的最终决策提供透明的认知支持[18]。由此,“解释义务”便呈现为一种可被技术化、程序化的责任形态。它不必然预设一个具有自由意志的道德主体,而是要求系统具备可解释性(Explainability)与可诠释性(Interpretability)。一个设计良好的 AI 智能师,完全能够通过技术手段,为其每一次的认知支持行为提供清晰、可追溯的解释。

与前述解释义务相对,第二种责任为“后果担当”。它追问的是“谁来担责”(Who bears the blame/credit?),要求行动者为其行动引发的外部效应,尤其是负面效应,承受惩罚、补偿等实质性后果。哈特(H.L.A.Hart)曾将此种责任形态明确界定为“liability-responsibility”,即行动者须为其行为所致后果负担法律或道德上的代价[8]。此种责任以主体资格(Agency)与自由意志(Free Will)为核心,本质上是一种关乎道德与法律代价的实质性后果承担。当一位教师因其教学判断失误导致学生身心受到伤害,并因此承受行政处分与法律赔偿责任时,其所承担的正是此种意义上的责任。该责任之所以能够归之于教师,是因其被预设为一个具备自主决断能力的道德主体,能够理解自身行为的规范性意义,并有能力在不同行动选项之间做出选择。“后果担当”的成立,必须建立在行动者具备自主决断能力、能够为其选择承担道德与法律代价的基础之上,它要求行动者具备自由意志与道德人格。这不仅是哲学上的思辨,更是法律与伦理实践的基石。一个行动者之所以要为其行为负责,恰在于其“本可以做出别的选择”。然而,当这一理路被推及 AI 智能师时,责任归属的症结便清晰地浮现出来,这也正是必须对两种责任做出严格区分的根由所在。

2. 区分的核心意涵

正是这一区分昭示,将“解释义务”与“后果担当”两种责任形态加以严格界分,它并非繁琐的概念辨析,而是直接关涉“三师课堂”中权责配置何以可能的核心问题。二者的根本差异,不仅体现为程序与后果的维度之别,更在于所预设的行动者类型截然不同:解释义务可在技术层面被事先设定并事后核查,其所要求的是可理解性与可追溯性而并非预设一个能够感受惩罚、理解道德的行动主体——一个合格的 AI 系统即可通过日志、可视化界面与解释模块来履行这一义务;后果担当则必须奠基于行动者的道德人格与自由意志之上,要求行动者能够进行规范性推理、理解行为的对错,并自愿承受其选择所带来的后果。

此种混淆所引发的困境,并不止于理论层面的含混,更直接地衍生为实践中的责任缺位:当损害发生之际,若既无法将后果担当归诸 AI[19],又因其未获法律主体地位而消解一切责任[20],则解释义务亦随之遁入真空。由此观之,严格界分两种责任形态,就不仅是对概念歧义的必要澄清,更构成了为 AI 智能师在教育人机协作中确立其恰当责任位置所不可或缺的逻辑起点。循此理路,AI 智能师的责任应被严格限定于其能力边界之内,即以程序性责任形态呈现的解释义务;而关涉道德与法律代价的实质性后果担当,则必须明确复归于唯一具备自由意志与道德人格的责任主体——人类教师。这一分界,为系统展开 AI 智能师责任的具体维度及其归属框架,奠定了概念基础与分析前提。

三、AI智能师责任的归属:基于三维框架的分析

(一)责任的三维分析:主体、关系与功能

若将上述区分进一步推向深入,更为精确地划定 AI 智能师的责任边界,则需要建立一个系统的分析框架。在

道德哲学与法哲学传统中,责任始终被理解为一个关系性概念,其内在预设了三个相互关联的追问:谁在负责、对谁负责、负责什么。哈特(H.L.A.Hart)在其经典论述中,正是以这一三元结构为隐含前提,将责任区分为角色责任、因果责任、能力责任与后果责任,从而揭示出任何完整的责任归属,都必须同时考查行动者的主体能力、其与他者的关系形态以及其所承担的角色义务[8]。有必要指出的是,吴河江已从“能力有限性”“权力有限性”“价值有限性”三个维度界定了 AI 作为“有限教育主体”的基本特征[6],这与研究所构建的“主体—关系—功能”三维框架存在内在的对应关系。研究进一步将此三重有限性分别转化为责任归属的三重追问——谁在行动、对谁负责、负责什么——进而判定 AI 所能承担的责任形态及其边界。在此基础上,研究将上述三个追问确立为分析框架的三个核心维度——主体、关系与功能。主体维度,追问“谁在行动”,聚焦于行动者是否具备承担后果担当所必需的自由意志与自主决断能力;关系维度,追问“对谁负责”,聚焦于行动者能否进入教育场域所特有的非对称关怀关系;功能维度,追问“负责什么”,聚焦于行动者的责任边界是否与其被授予的功能权力相匹配。只有当行动者在这三个维度上均具备相应的资格与条件时,方能成为一个完整的责任主体。因此,对 AI 智能师的系统考察将依循这三个维度,以判明其在何种意义上能够成为责任主体,又在何种意义上必须止步于有限的责任形态。

1. 主体维度:谁在行动?

承担责任的首要前提,在于存在一个可被识别、可被问责的行动主体。而判定 AI 智能师能否成为责任主体,需要首先澄清“主体”的判准本身。道德哲学传统将这一判准锚定于“内在主体性”——行动者是否具备自觉的意识、情感与自由意志。从亚里士多德以“自愿性”(Voluntariness)作为道德判断的起点[21],到康德将“意志自由”(Freedom of the Will)确立为道德之所以可能的绝对前提[22],再到当代道德哲学以“备选可能性原则”(Principle of Alternate Possibilities)凝练这一洞见[23],其核心命题一以贯之:一个人承担责任的前提,是其在行动时“能够做其他不同的事”(could have done otherwise)。即便相容论者以“引导性控制(Guidance Control)”弱化其刚性,责任的前提仍在于行动者能够领会规范性理由并据此引导行为[24]。据此判准,AI 不具备承担后果担当所必需的自由意志与道德人格,这构成了主体维度分析的出发点。

然而,仅以内在主体性为判准,尚不足以完整回应 AI 在教育场域中的责任归属问题。因为此种判准回答的是“AI 能否成为与人类教师同等的责任主体”,却未能解释一个更为切近的现象:即便 AI 不是“完全主体”,其在教学互动中表现出的自主行动力——主动推送资源、发起追问、生成反馈——又当如何从责任角度加以安顿?这正是“交互主体性”概念所提供的分析空间。殷杰将其界定为“参与交互实体的一种潜在行动能力”[25],强调主体性源自交互过程本身,而非实体预先具备的内在属性。这一概念的意义不在于放松主体资格的判准,而在于揭示了一个被传统判准所遮蔽的维度:行动者的责任形态,并非只能在其“是”或“不是”主体之间二择其一,而是可以依据其在关系中“做了什么”而加以差异化判定。由此,内在主体性与交互主体性在主体维度的分析中形成了一种互补结构:前者划定 AI 不能承担的后果担当的边界,后者则为 AI 在交互中自主发出行动时所应承担的解释义务提供了依据。这即是说,AI 在主体维度上并非全无责任能力,而是其责任能力有其确定的结构性限度。

首先,AI 智能师缺乏自由意志。其“决策”本质上是基于海量数据训练出的算法模型,对新输入数据进行概率计算与模式匹配后得出的输出。无论这一过程何等复杂、输出结果何等精妙,它既不具备亚里士多德意义上的“自愿性”,更遑论康德哲学语境下的“意志自由”。AI 并不具备自主形成的意图,其行动源于人类指令的驱动,因而难以进行超越算法数据模式之外的道德权衡。它的每一次“选择”,均由其算法、训练数据与当前输入所唯一决定,并不存在“本可另行选择”的本体论空间。其次,AI 智能师无法理解规范性。它可以处理包含道德词汇的文本,甚至生成符合伦理规范的语句,却无法真正“理解”这些规范的内在价值——其所处理的是语词的统计相关性,而非道德的应然性,因而无法对道德理由做出真正的自主回应。最后,对 AI 智能师施加惩罚并无实效。惩罚作为

“后果担当”的核心要素之一,其预设乃是一个能够理解归责理由、感知惩罚意义并据此调整后续行为的主体。让一台服务器或一段代码承受“惩罚”,不仅在操作层面流于荒谬,在学理上也了无意义。从实践层面看,近年各类AI争议事件——如微软Copilot的安全控制问题[26]或自动驾驶汽车的事故[27]——最终的问责对象无一例外均为其开发者、运营者或所有者,而非AI系统本身,这正印证了AI在主体维度上的资格缺失。

上述论证立足于内在主体性标准,指向一个明确的判断:AI不具备承担后果担当所必需的自由意志与道德人格。然而,若转向交互主体性标准,则需承认,AI在教育互动中的确表现出某种程度的自主行动力——它能够主动推送学习资源、发起追问、生成反馈,这些行为并非每一步都由人类实时操控。那么,这种交互主体性是否足以支撑AI承担某种责任形态?答案同样是明确的,但需加以限定:交互主体性所支撑的,正是解释义务——行动者因其自主发出行动而负有就此行动给出说明的义务——而非后果担当。后果担当之所以不能由交互主体性所支撑,是因为它最终要求行动者理解惩罚与代价的道德意义,而这恰恰需要内在主体性所涵括的自由意志与道德人格。由此,两种标准并非彼此否定,而是在责任形态的划分中形成了互补关系:交互主体性为解释义务提供了依据,内在主体性的阙如则为后果担当划定了边界。

2. 关系维度:对谁负责?

与主体维度的追问不同,关系维度将审视的目光从“谁在行动”转向“对谁负责”,由此揭示教育责任的另一重构成条件。教育责任并非在真空中生成,它总是发生于教师与学生之间具体而动态的伦理关系之中。这一维度所追问的核心问题是“对谁负责”。教育责任的特殊性在于,它本质上是一种深刻的、非对称的关怀关系[28]。汉斯·约纳斯在《责任原理》中为此种非对称责任提供了深刻的存在论论证。他指出,责任的核心并非源于主体间的契约或对等交换,而是源于强者对脆弱者的本体论承诺,父母对子女的责任即其原初模型[29]——此种责任的产生,乃因一个有能力的、具有未来意识的主体,面对一个脆弱的、需要被照料的存在时,生发出一种源于“存在应该延续”这一根本伦理命令的“操心”(care)。这种责任是单向的、不求回报的。马丁·布伯在《我与你》中则从对话哲学的角度揭示,真正的教育发生在“我—你”(I-Thou)关系之中[30]:教师与学生并非作为客体(It)被观察、分析与利用,而是作为完整的、独一无二的存在者相互遇见、彼此确认、彼此回应,教育的本质乃是“一个完整的人对另一个完整的人的全面影响”。内尔·诺丁斯的关怀伦理学进一步将这一理念具体化,指出教育关怀要求关怀者对被关怀者展现出“专注”与“动机移位”[31]——即全神贯注于学生的需求、感受与视角,并将自身的动机能量从自我中心转移至促进对方成长与福祉的目标之上。将这一关系维度的责任判定标准应用于AI智能师,可以得出:在关系维度上,AI无法嵌入教师与学生之间那种承担完整教育责任所必需的伦理关系。

其一,AI尚难真正“认识”学生。它所处理的是学生的“数据画像”——一系列关于学习行为、知识掌握程度与互动频率等形成的离散数据点。尽管借此可以实现高效的个性化学习支持[32],但这与布伯所说的“我—你”相遇有着本质区别:AI所能面对的,在根本意义上仍是一个可计算、可优化的“它(It)”,而非一个充满无限可能、拥有喜怒哀乐、需要被整体关怀的“你(Thou)”;其二,AI所展现的关怀难以构成一种真实的伦理关怀。它可以被编程为在学生答错时输出鼓励性话语,但这与约纳斯意义上源于对生命脆弱性之承诺的“操心”,以及诺丁斯意义上需要真实情感投入的“专注”与“动机移位”,存在着根本差异。此种反应在本质上是一种基于预设规则或模式匹配的功能性输出,尚难等同于源于伦理自觉的真实情感互动。目前有关AI情感设计的主流观点是将情感状态技术性地整合于AI的内部表征和决策过程中。这并不意味着系统像人类一样“感受”情感,而是将情感作为生物学灵感来源——正如人工神经网络受生物神经网络启发一样[33];其三,AI在伦理关系中往往呈现为一种结构性缺席。教育责任在很大程度上乃是一种灵魂唤醒另一个灵魂的托付,要求责任主体能够感知、理解并回应另一个生命的整体状态。对于AI是否具有生命与灵魂的伦理问题尚在讨论之中,因此无法判断AI能否真正感受学生的

痛苦与喜悦,因而难以搭建以生命共感为根基的完整教育伦理联结。让一个难以进入伦理关系的存在,去承担以伦理关系为基础的完整教育责任,在逻辑上面临着很大的困难。

因此,从关系维度审视,AI 尚难被视为教育责任的完整主体,因其在相当程度上被结构性地排除于师生之间那种承载着关怀与承诺的独特伦理关系之外。此种局限,正是吴河江所指出的 AI“价值有限性”的深层体现:AI 无法触及教育中“人的灵魂唤醒”与“价值引领”的核心功能[6],而这一功能的实现恰恰以真实的伦理关系为前提。

3. 功能维度:负责什么?

与关系维度将责任锚定于师生间的伦理关联不同,功能维度转向一个更富实践性的追问:“负责什么”。在组织与社会场域中,责任的范围与深度通常应与行动者被授予的功能角色及权力范围相匹配,这一“权责对称”(Symmetry of Power and Responsibility)原则,构成了现代治理的一项基本准则。在教育行政领域,权责对称已是一项基本原则。①陈大兴指出,责任与问责是教育治理的核心概念,权力的授予必须以责任的明确为前提,责任的边界应当与权力的边界保持对应[34]。吴能武进一步揭示,权力与责任是相互界定、彼此约束的范畴,权力的边界决定了责任的范围[35]。在法学领域,这一原则同样适用。龙文懋在对 AI 法律主体地位的法哲学思考中提出,即便 AI 在某些领域被赋予有限的法律主体资格,其所承担的责任也必须与其权利能力和行为能力相匹配,不能超出其功能范围[20]。据此而论,AI 的责任形态在性质上倾向于有限,尚难以承担与其功能角色不相称的整全责任。将功能维度的判准应用于 AI 智能师,其责任边界便趋于清晰:对 AI 责任的讨论,不宜脱离其在教育场域中的具体功能定位而进行抽象的追问。

首先,就功能定位而言,AI 智能师在“三师课堂”场景中更宜被理解为辅助者而非最终决策者[5]。冯锐的研究指出,AI 的核心作用在于为教学提供认知支持与过程反馈,而非替代教师的专业判断[36]。从“三师课堂”理论建构所梳理的 AI 智能师权力清单来看,无论是知识传递层面的认知支持、过程组织层面的互动介入,还是评价反馈层面的学情诊断,均在总体上归属于辅助性功能的范畴。在全球范围内,相关 AI 教育政策也呈现出相近的取向:欧盟《人工智能法案》将教育 AI 应用列为“高风险”系统予以严格监管,联合国教科文组织持续强调以人为本的方针,其内在逻辑均指向将 AI 置于人类教师的监督与控制之下,使其扮演辅助性角色;新近发布的《人工智能教育伦理:参考框架(2026 年版)》同样立足教育的内在规律与育人使命,确立了“以人的全面发展为根本目的”的核心理念。从上述政策取向来看,当前全球 AI 教育治理普遍倾向于主张 AI 不宜涉足价值判断、最终裁量与高风险决策,这些领域通常被视为人类教师所保留的核心专业权力。

其次,AI 智能师有限的功能定位,内在地要求与其相匹配的有限责任。如果其在教育场域中的功能角色是辅助性的,那么相应的责任形态也应限于有限范围。前文论及的 AI 争议事件,其最终问责大致指向开发者、运营者或所有者,而非 AI 系统本身,这从实践层面提示,辅助决策系统的责任形态宜与其辅助功能保持对应——系统提供信息与建议,人类做出最终决策并承担相应后果。将此逻辑迁移至教育领域,问题的轮廓便趋于清晰:AI 智能师被授权在辅助性功能范围内提供认知支持,其责任也相应地被限定于就其输出提供可理解的说明,即解释义务;而采纳该输出内容并做出最终教学决策所引发的后果,则归于做出该决策的人类教师承担,即后果担当。

最后,功能维度的分析最终指向实现 AI 有限责任的具体技术路径,即以技术手段赋能 AI 智能师履行解释

① 权责对称作为现代治理的一项基本准则,在多个学科领域均有明确表述。在高等教育治理领域,刘淑华指出,权责对称是高等教育分权的底线,需要遵循每种权力主体自身的权责对称性,使每一种权力主体所拥有的权力止于其所能承担责任的边界。参见:刘淑华. 权责对称:高等教育分权的底线[J]. 浙江大学学报(人文社会科学版),2009,39(2):144-152. 在政治学领域,麻宝斌、郭蕊强调,权责一致是政治生活的基本原则,权责背离则会引发责任追究的困境。参见:麻宝斌,郭蕊. 权责一致与权责背离:在理论与现实之间[J]. 政治学研究,2010(1):72-78.

义务。AI 承担解释义务并非空洞的承诺,而具有坚实的技术支撑。贺栩溪在其关于算法侵权的研究中指出,破解算法“黑箱”责任困境的关键在于确立算法的解释义务[37]。该义务的履行是打通责任追溯通道的前提,它解决的是“如何查明问题”的认识论难题,从而为后续人类主体承担后果担当提供必要条件。在技术架构上,一个负责任的教育 AI 系统通常需要内置一个“解释子系统”(Explanation subsystem)[38]。其工作流程如下:系统以学生的学习数据为输入,AI 决策模块生成教学建议;同时,解释生成模块启动,利用 LIME(局部可解释模型无关解释)[39]、SHAP(沙普利加性解释)[40]或注意力可视化等技术,生成一份人类可读的解释报告。该报告会清晰说明建议所依据的关键特征,并排除其他选项。这一技术路径使解释义务从理论构想转化为可嵌入系统设计并接受持续验证的操作性要求。当前,尽管教育领域专用的开源 XAI 系统尚不丰富,但 EXAIT 等工具及相关框架研究已为其初步指明了发展方向[41-42]。面向 2026 年及此后,构建此类可解释性机制已被视为破解教育 AI 安全、公平与信任难题的关键进路。

因此,从功能维度审视,AI 的辅助性角色使其责任形态倾向于有限。此种有限责任,可进一步落实为一种能够通过技术手段实现的、程序性的解释义务。这既与权责对称的治理原则相一致,也与研究关于 AI“权力有限性”的判断形成呼应——AI 在教育场景中的权力来源于人类教师的让渡,教师拥有最终解释权与教学控制权,因而也理应承担与此权力相应的最终责任。

(二)责任的归属:解释义务归AI,后果担当归人类

经由主体、关系与功能三个维度的递次审视,研究的核心主张得以逐步确立:在人机协同的“三师课堂”场景中,责任的划分并非任意配置,而是严格对应于行动者的本体论资格、关系论位置与功能性角色。这一分析框架的展开,本身便为责任归属的判定提供了逻辑路径——三个维度上的条件差异,已然预示了 AI 智能师与人类教师在责任承担上的根本分野。循此理路,研究将进一步阐明此种分野所指向的责任最终归属。

首先,在现有分析框架下,AI 智能师所能承担的责任形态,宜被限定为一种有限的、程序性的“解释义务”。这一判断,与既有研究将 AI 界定为“有限教育主体”的立场相一致[6]:“有限”之要义,正在于承认 AI 在交互中具备一定的自主行动力,但同时确认此种自主性不足以支撑其成为完整的责任主体。将此种责任归于 AI 之所以具备可行性,是因其不预设自由意志与道德人格,从而得以避开主体维度的资格限制;无需建立真实的伦理情感联结,从而绕过了关系维度的结构性困难,并与其作为信息处理与认知支持工具的辅助性角色相契合,进而符合功能维度的权责对称要求。AI 智能师的解释义务,要求系统在设计与运行层面具备高度的透明性与可追溯性,这构成了未来教育 AI 监管的核心技术要件。正如欧盟《人工智能法案》所预示的治理趋向,以及中、美等国在 2024 至 2026 年间密集出台的各类 AI 教育指南所强调的伦理与安全原则,确保算法的可解释性正逐步成为不可或缺的合规标准[43]。

与此相对,最终的价值裁断与后果担当,在根本上仍须归于人类主体。将此种实质性责任赋予人类教师,其根据在于,唯有人类教师才具备承担此种责任所要求的全部条件:他拥有自由意志与道德人格,能够进行价值判断,从而满足主体维度的资格要求;他能够与学生建立起真实的关怀性伦理关系,对一个完整的生命负责,从而回应关系维度的条件;他被社会与法律授予做出最终教育决策的权力,因而也须承担与此权力相应的最终责任,从而符合功能维度的权责对称原则。这即是说,在“师—机—生”三元教学场景中,人类教师始终居于主导地位,是最终的责任承担者,这一地位不因技术赋能而发生根本性转移。在实践层面,这意味着无论 AI 智能师提供了何等精准的数据分析与教学建议,最终的采纳、拒绝或修正决定,以及由此产生的一切后果,其责任归属均指向人类教师。尽管当前实证研究尚未直接聚焦于责任归属问题,但围绕教师在人机协同教学中如何做出决策的研究已渐次展开[44-46],这恰从侧面凸显了人类教师作为最终决策者与责任承担者的关键地位。

循此理路,针对“权力在内,责任在外”的困境,宜将 AI 的权力严格框定于其功能边界之内,并使其权力与一种内在的、可被履行的解释义务相绑定;同时,将超越 AI 能力范围的责任明确复归于人类教师。这并非对 AI 责任的豁免,而是对责任归属的严肃划定。由此划出一条清晰的边界:AI 负责说明何以如此,人类负责承担由此而来的后果。解释义务构成了 AI 责任的终点,却也恰恰成为人类责任的新起点——一个能够清晰解释其建议来源与逻辑的 AI,将人类教师从繁复的信息处理工作中解放出来,使之站在信息更透明、视野更开阔的位置上,从事真正具有创造性、情感性与伦理性的教育工作。

四、结语

研究围绕“三师课堂”构想中 AI 智能师的权责配置问题——当 AI 被冠以“师”之名、被赋予介入教学流程的行动权力时,其责任设定却高度依赖人类的外部监督,由此呈现出“权力在内,责任在外”的失衡格局。这一权责不对称,构成了“三师课堂”理论建构中一个亟须闭合的逻辑缺口。

为回应这一问题,研究首先在认识论层面厘清了“责任”概念的内在结构,进而构建了“主体—关系—功能”三维分析框架。三维综合审视最终指向研究的核心主张:AI 智能师所应承担的责任形态,乃是一种有限且程序性的解释义务;而关乎价值裁断与后果担当的实质性责任,则必须归于人类教师。这一责任划界的理论意义或不限于“三师课堂”的具体构想,它揭示了一个更具普遍性的教育命题:当技术系统被赋予越来越接近教育者的角色与权力时,对其责任的讨论便不能停留于抽象的伦理呼吁或笼统的“负责任 AI”口号,而必须返回责任概念本身的逻辑结构之中。唯有如此,方能在技术赋能与伦理规约之间寻得平衡,既充分释放技术的潜能,又牢牢守护人的主体地位。

展望未来,这一责任划界将深刻重塑人类教师的专业角色。当 AI 承担起认知层面的辅助工作后,人类教师将从重复性的信息处理中解放出来,转而成为审慎的决策者——批判性地解读和评估 AI 提供的解释,以专业判断做出最终裁量;成为伦理的守望者——警惕技术对教育过程的过度技术化与非人化倾向,守护教育的价值导向;成为教育关系的建构者——将更多的时间与精力投入到与学生之间深刻的“我—你”关系中,致力于情感的交流、价值观的引导与人格的涵育。这三种角色并非对教师传统职能的消解,而是技术时代对教师专业性提出的更高要求,亦是对教育本真的更深层回归。

当然,上述责任划界立足于当前 AI 技术发展的阶段性特征。随着 AI 交互自主性的持续增强,其“准主体”的地位或将得到改变,因而解释义务与后果担当之间的边界,以及“有限教育主体”之“有限”的具体内涵,仍需在未来的研究与实践中持续审视与动态调整。这既是研究的审慎自限,也为后续探索留下了开放空间。总而言之,对 AI 智能师责任的清晰划界,根本目的在于使责任各归其位:令技术性的解释义务归于技术系统,令伦理性的后果担当归于人类教师。教育育人责任沉重而神圣,终需落回温暖而坚实的人之肩上——这既是研究的逻辑终点,亦是人机协同教育实践不可偏离的价值起点。

参考文献:

[1] 周洪宇. “三师课堂”是人工智能时代中国本土教育创新的实践答卷[J]. 教育文汇,2025(10):1.
[2] 周洪宇,杨朝晖,操太圣,等. “三师课堂”的理论建构与实践样态(笔谈)[J]. 天津师范大学学报(社会科学版),2026(1):9-34.
[3] 教育部资源中心发布《中国教师生成式人工智能应用报告(2026)》[EB/OL]. (2026-05-14)[2026-05-20]. https://www.

ncet.edu.cn/zhuzhan/sjsyzyxw/20260514/6765.html.
[4] STANFORD INSTITUTE FOR HUMAN-CENTERED AI (HAI). Artificial Intelligence Index Report 2025[EB/OL]. Stanford:Stanford University,2025(2026-05-12)[2026-05-20]. https://hai.stanford.edu/ai-index/2025-ai-index-report.
[5] 周洪宇,李坤艳. 人工智能在"三师课堂"教学中的应用、风险及规避[J]. 中国电化教育,2026(3):38-45.
[6] 吴何江. 生成式人工智能的教育主体地位辨析[J]. 教育研究与实验,2026(3):25-35.
[7] Duff R A. Answering for crime:Responsibility and liability in the criminal law[M]. Oxford:Bloomsbury Publishing,2007:19-36.
[8] Hart H L A. Punishment and responsibility:Essays in the philosophy of law[M]. Oxford:Oxford University Press,2008:28-53.
[9] Vanlehn K. The relative effectiveness of human tutoring,intelligent tutoring systems,and other tutoring systems[J]. Educational Psychologist,2011,46(4):197-221.
[10] 赵晓伟,沈书生,祝智庭. 数智苏格拉底:以对话塑造学习者的主体性[J]. 中国远程教育,2024,44(6):13-24.
[11] 李海峰,王炜. 人机协同深度探究性教学模式——以基于ChatGPT和QQ开发的人机协同探究性学习系统为例[J]. 开放教育研究,2023,29(6):69-81.
[12] Lin C C,Huang A Y,Lu O H. Artificial intelligence in intelligent tutoring systems toward sustainable education:A systematic review[J]. Smart Learning Environments,2023,10(1):41.
[13] Shaik T,Tao X,Li Y,Dann C,et al. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis[J]. IEEE Access,2022,10:56720-56739.
[14] Kostopoulos G,Gkamas V,Rigou M,Kotsiantis S. Agentic AI in education:State of the art and future directions[J]. IEEE Access,2025,13:177467-177491.
[15] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nature Machine Intelligence,2019,1(5):206-215.
[16] Arrieta A B,Díaz-Rodríguez N,Ser D J,et al. Explainable Artificial Intelligence (XAI):Concepts,taxonomies,opportunities and challenges toward responsible AI[J]. Information Fusion,2020,5882-115.
[17] 王萍,田小勇,孙侨羽. 可解释教育人工智能研究:系统框架、应用价值与案例分析[J]. 远程教育杂志,2021,39(6):20-29.
[18] Liu Q,Pinto J D,Paquette L. Applications of explainable AI (XAI) in education[M]//Trust and inclusion in AI & mediated education:Where human learning meets learning machines. Cham:Springer Nature Switzerland,2024:93-109.
[19] 郝德永. "重塑"之问——关于人工智能驱动教育改革假设的检视与警示[J]. 教育研究,2026,47(3):75-88.
[20] 龙文懋. 人工智能法律主体地位的法哲学思考[J]. 法律科学(西北政法大学学报),2018,36(5):24-31.
[21] 亚里士多德. 尼各马可伦理学[M]. 廖申白,译注. 北京:商务印书馆,2003:61-68.
[22] 康德. 道德形而上学的奠基[M]. 李秋零,译注. 北京:中国人民大学出版社,2013:69-71.
[23] 苏德超. 替选可能、道德责任与未来优化——以丹尼特的理解为中心[J]. 道德与文明,2024(1):53-64.
[24] Fischer J M,Ravizza M. Moral responsibility for actions:Moderate reasons&responsiveness[M]//Responsibility and control:a theory of moral responsibility. Cambridge:Cambridge University Press,1998:62-91.
[25] 殷杰. 生成式人工智能的主体性问题[J]. 中国社会科学,2024(8):124-145+207.
[26] Avinash A,Manisha N. Advancing trustworthy AI for sustainable development:Recommendations for standardising AI incident reporting[C]//2024 ITU Kaleidoscope:Innovation and digital transformation for a sustainable world (ITU K). IEEE,2024:1-8.
[27] Awad E,Lanisha S,Kleiman-Weiner M,et al. Blaming humans in autonomous vehicle accidents:Shared responsibility across levels of automation[J]. Nature Human Behaviour,2018.
[28] 张广君,宋文文. 教师"为他责任"伦理:言说与批判[J]. 高等教育研究,2019,40(2):27-33.
[29] 约纳斯. 责任原理:技术文明时代的伦理学探索[M]. 方秋明,译. 上海人民出版社,2010:234-236.
[30] 布伯. 我与你[M]. 维纲,译. 北京:商务印书馆,2015:33.
[31] 诺丁斯. 学会关心:教育的另一种模式[M]. 于天龙,译. 北京:教育科学出版社,2014:24-25.
[32] 徐升,佟佳睿,胡祥恩. 下一代个性化学习:生成式人工智能增强智能辅导系统[J]. 开放教育研究,2024,30(2):13-22.
[33] Li Y,Sun Q,Schlicher M,et al. Artificial emotion:A survey of theories and debates on realising emotion in artificial in-

telligence[J]. 2025.
[34] 陈大兴. 高等教育中责任与问责的界定[D]. 上海:华东师范大学,2014.
[35] 吴能武. 高等教育治理中的权力关系及其优化[D]. 上海:华东师范大学,2020.
[36] 冯锐,孙佳晶,孙发勤. 人工智能在教育应用中的伦理风险与理性抉择[J]. 远程教育杂志,2020,38(3):47-54.
[37] 贺栩溪. 人工智能算法侵权法律问题研究[D]. 长沙:湖南师范大学,2021.
[38] Jantakun T,Jantakun K,Jantakoon T. A common framework for artificial intelligence in higher education (AAI-HE mode)[J]. International Education Studies,2021,14(11):94-103.
[39] Ribeiro M T,Singh S,Guestrin C. ‘Why Should I Trust You? ’ Explaining the Predictions of Any Classifier[C]//Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies 2016:demonstrations:15th conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies (NAACL HLT 2016),12-17 June 2016,San Diego,California,USA.ACM,2016.
[40] Lundberg S M,lee S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems,2017,30.
[41] Khosravi H,Shum S B,chen G,et al. Explainable artificial intelligence in education[J]. Computers and Education:Artificial Intelligence,2022,3:100074.
[42] Türkmen G. The review of studies on explainable artificial intelligence in educational research[J]. Journal of Educational Computing Research,2025,63(2):277-310.
[43] Shenoy P,Saarela M,Bui T X. AI literacy frameworks for educators:An umbrella review[C]//Proceedings of the Annual Hawaii International Conference on System Sciences. Honolulu:Hawaii International Conference on System Sciences,2026.
[44] Holstein K,Aleven V. Designing for human-AI complementarity in K-12 education[J]. AI Magazine,2022,43(2):239-248.
[45] Cavalieri Goncalves Peloche J. Unpacking developers,curriculum leaders,and K-12 teachers’ perceptions of artificial intelligence in education[D]. Wollongong:University of Wollongong,2024.
[46] Jin Q,Borchers C,Fancsali S,et al. Who to help? A time&slice analysis of K-12 teachers’ decisions in classes with AI & supported tutoring[C]//Proceedings of the 18th International Conference on Educational Data Mining. 2025:640-645.

[责任编辑:杨毅恒]