■ 教育数字化

DOI:10.15998/j.cnki.issn1673-8012.2024.02.004

教育人工智能伦理治理:现实挑战与实现路径

白钧溢

(东北师范大学　教育学部,长春　130024)

摘　要:教育人工智能伦理治理旨在使教育人工智能系统的设计、开发与应用符合道德、法律和社会价值观,以保障人工智能技术对教育数字化转型的切实赋能。为应对"知而不行""知为不行""制为不行"的治理难题,由"单一软法"向"软硬兼备"转变已成为教育人工智能伦理治理的发展共识与诉求。但由于人工智能伦理存在诸多不确定性,且相关政策制定与学术研究尚处于起步阶段,实现"软硬兼备"的教育人工智能伦理治理面临治理对象界定不清、伦理原则指代不明、治理主体尚未形成多元共治合力、治理方式存在孤立性的现实挑战。实现"软硬兼备"的教育人工智能伦理治理,前提与关键在于弹性界定教育人工智能概念,提供具有可操作性的原则指南,落实"一核多元"的权责划分,探索实验治理与混合治理的行动框架。

关键词:伦理治理;教育治理;教育人工智能;软法治理;硬法治理

[中图分类号]G644　[文献标志码]A　[文章编号]1673-8012(2024)02-0037-11

人工智能作为数字技术的突出代表,是教育数字化转型的"数字基座"[1],在变革知识供给形态、优化教育评价模式、改进教育管理形式等方面发挥着重要作用。然而,伴随着人工智能技术与教育的深度融合,隐私与安全隐患、学生发展同质化、算法歧视、学生数据所有权争议等伦理问题不断显现,伦理治理显示出必要性与紧迫性。目前,各国政府、学界和国际组织已在教育等领域进行了诸多人工智能伦理治理探索,人工智能治理呈现由软治理为主到软治理与硬治理相结合的发展趋势。由于人工智能伦理本身具有复杂性,且人工智能伦理治理探索尚处于起步阶段,无论是软治理所依赖的伦理原则,还是硬治理所主张的监管,在当前都面临着争议与挑战。基于此,本文聚焦教育人工智能伦理治理的发展趋势与发展诉求,剖析当前教育人工智能伦理治理面临的挑战并在此基础上提出应对策

修回日期:2023-12-11
基金项目:教育部人文社会科学研究青年基金项目"人工智能、制度嵌入与大学治理效能提升研究"(22YJC880107)
作者简介:白钧溢,男,河北承德人,东北师范大学教育学部博士生,主要从事技术哲学和教育人工智能伦理研究。
引用格式:白钧溢. 教育人工智能伦理治理:现实挑战与实现路径[J]. 重庆高教研究,2024,12(2):37-47.

Citation format: BAI Junyi. Ethical governance of educational AI: real challenges and implementation pathways[J]. Chongqing higher education research,2024,12(2):37-47.

略,以期为保障人工智能赋能教育数字化转型提供治理方向。

一、由“单一软法”到“软硬兼备”:人工智能伦理治理的发展诉求

从 2016 年起,随着人工智能应用的不断增加,其伴生的伦理风险逐渐受到关注。因缺少对新兴技术进行监管干预的数据支撑,对人工智能伦理风险的最初回应是依靠软治理手段,即通过制定伦理原则、伦理指南和伦理准则,从宏观上对人工智能进行治理[2],这被认为是未来相当长的时间内人工智能伦理治理的基本途径[3]。这一时期的人工智能伦理治理研究具有 3 个方面的特征。一是伦理原则构建主体广泛。伦理原则构建囊括国际组织、各国政府、学界和企业等多个主体,如欧盟的《可信任的人工智能伦理指南》、全球工会联合会的《合伦理的人工智能十大原则》、中国的《新一代人工智能治理原则——发展负责任的人工智能》、美国的《为人工智能的未来做好准备》、英国的《英国人工智能》、德国电信的《人工智能指南》等。二是伦理原则构建响应迅速。仅在 2017—2019 年,由国际组织、各国政府、企业、学术组织与研究机构制定的人工智能伦理原则清单已接近 200 份。三是伦理原则相似度高。对哈佛大学的《有原则的人工智能:在基于伦理与权利下达成共识》、南里奥格兰德天主教大学的《全球人工智能伦理:治理指南和建议回顾》、苏黎世联邦理工学院的《人工智能:道德准则的全球格局》及奥卢大学的《人工智能伦理:原则和挑战的系统综述》4 份人工智能伦理原则的元分析梳理发现,不同时期、不同领域、不同级别的人工智能伦理框架大多包括隐私、透明度、问责制、以人为本、安全和可持续等共性原则[4]。

受一般人工智能伦理治理取向影响,教育领域最初也采取基于伦理原则的软治理模式[4]。教育人工智能伦理原则构建起步于 2019 年召开的第二十届教育人工智能国际会议(AIED'19),该会议取得了两个方面的进展:一是基本勾勒了教育领域的宏观伦理原则框架。不论是国家或国际组织提出的教育人工智能伦理原则框架,如澳大利亚《简读:人工智能与学校教育》、白金汉大学教育人工智能伦理研究所《人工智能教育伦理框架》、联合国教科文组织《北京共识——人工智能与教育》,还是构建教育人工智能伦理原则的学术研究,大多将透明、可解释性、公平、隐私、安全性和问责制等一般人工智能的共识性伦理原则或其变式纳入框架中[4]。二是在共识性伦理原则的基础上开始了针对教育原则、原则适用范围、原则实践指导力的个性化探索,如关注教育人工智能核心价值的《教育人工智能伦理的“稻草人”草案框架》,关注伦理原则适用学段的《人工智能一代:建立儿童和人工智能的全球标准》《K-12 教育的人工智能伦理准则》《儿童人工智能政策指南》,评估人工智能教育系统有效性的《教育领域人工智能的透明度指数框架》,以及分析人工智能教育应用公平性的《人工智能教育的结构性正义视角》等。截至目前,伦理原则的条目确立、内涵探究与应用途径依然是国内外教育人工智能伦理治理研究的热点话题。

随着基于伦理原则的软法治理实践的开展,研究者们发现这一治理手段存在 3 个方面的问题。一是“知而不行”。一项对 5 家人工智能研发公司的调查发现,虽然所有的调查对象均承认伦理原则的重要性,但当被问及他们是否会在人工智能开发实践中考虑道德时,所有调查对象的回答都是“否”[5]。二是“知为不行”。对伦理原则的依赖与强调加剧了“伦理漂蓝”①风险,即部分人工智能研发公司会通过市场宣传、广告及其他公关活动使人们相信他们的行为、产品与服务合乎伦理,实际上却旨在节约成本、提高竞争优势而并非真正为了遵守[6]。三是“制为不行”。对 160 套人工智能伦理

① “伦理漂蓝”(ethics blue washing)由耶鲁大学教授卢西亚诺·弗洛里迪(Luciano Floridi)提出,他将环境伦理学“伦理漂绿”(ethics green washing)概念类比到数字技术领域,指部分数字技术企业通过市场宣传、广告以及其他公共关系活动,使人们相信他们的行为、产品与服务等合乎伦理原则,但实际情况并非如此的现象。

原则和指导文件的评估表明,多数具有约束力的协议和自愿承诺都是由私营部门提出的[7]。私营部门参与伦理原则制定饱受质疑,因为它们可能会将伦理原则及其所支撑的软治理作为免受监管的屏障并延迟必要法规的颁布[8]。

面对软法可能得不到有效实施的治理挑战,人工智能伦理治理模式逐渐向软硬兼备、软法"硬化"转变,以提高软法的约束力与执行可能性[9]。2020 年 2 月,欧盟委员会在《可信赖的人工智能伦理准则》的基础上发布了由政策框架与监管框架组成的《人工智能白皮书》,承诺化解开发与使用人工智能系统的相关风险[10]。在此背景下,2021 年 4 月,欧盟委员会提交了《制定人工智能统一规则》的法规提案,即《人工智能法案》,旨在为人工智能的开发与应用提供监管与法律依据。《人工智能法案》是全球主要经济体首次尝试通过制定人工智能的一般法律框架对其进行横向监管,它的提出标志着单一软法向软硬兼备转变[11],掀起了各国构建和推广人工智能治理模型的竞赛[7]。在我国,2022 年出台的《关于加强科技伦理治理的意见》表明,科技伦理包含伦理规范与法律规范两类规范,其治理依赖稳健高效的治理机制[12]。目前,我国已初步明确了行业伦理原则(软法)与制定法(硬法)相结合的综合治理架构,《人工智能法(草案)》也进入国务院 2023 年度立法工作计划。除宏观人工智能领域外,教育领域也开始关注软治理与硬治理的结合。如白金汉大学教育人工智能伦理研究所 2020 年发布的《中期报告:实现教育中道德人工智能的共同愿景》(*Interim Report: Towards a Shared Vision of Ethical AI in Education*)指出,除伦理原则外,还需要在法律范围内保障教育人工智能应用的安全性。国内学者亦强调教育人工智能伦理治理要做到"自律"与"他律"相结合[13],但相较于宏观人工智能领域,教育领域的相关探索尚处于起步阶段。如何根据一般人工智能伦理治理的共性特征,为教育创新治理沉淀证据和积累经验以实现软硬兼备的人工智能伦理治理,是教育领域亟须解决的方向性问题。

二、"软硬兼备"诉求下教育人工智能伦理治理的现实挑战

尽管由单一软法向软硬兼备转变已成为教育人工智能伦理治理的共识,但就现实情况而言,当前教育人工智能伦理治理在软治理与硬治理上均存在较大问题,表现在治理对象、伦理原则、治理主体、治理方式 4 个方面。

(一)治理对象界定不清

治理对象是教育人工智能伦理治理的作用标靶。在技术治理领域,明确治理对象至关重要,因为人工智能的治理范围、治理主体、治理手段乃至整个治理话语体系的确定在很大程度上取决于对"人工智能"一词的界定[14]。目前,教育人工智能概念不够明确,这是实现软硬兼备的教育人工智能伦理治理的首要挑战。

伦理治理需要对教育人工智能进行清晰定义以保障治理的确定性与针对性,但实现这一目标比较困难。对教育人工智能难以做出普适性定义存在两方面原因。一是不存在固定的人工智能。自 1956 年达特茅斯会议以来,人工智能的范畴及其涉及的应用范围始终在变化,这被称为"AI 效应",即最初被认为"智能"的技术经过习惯性使用和原理检验后失去其"智能"地位[15]。二是不存在统一的人工智能。人工智能是一个相当宽泛的集体名词,大量技术和应用被归入这一总称,很难以单一方式在所有情况下清楚地描述它[16]。在教育人工智能中,最具代表性的例子是智能导学系统(ITS),至今已有 50 余年的发展历程,先后涌现出 CIRCSIM-Tutor、Geometry Explanation Tutor、Cordillera、AutoTutor 和 BEETLE Ⅱ等众多系统,仅 AutoTutor 在 1997 年至今便已有三十余个系列产品[17]。

定义困难对教育人工智能伦理治理的影响已有所体现。一是回避界定人工智能。欧盟人工智能

高级别专家组(AI HLEG)成员娜塔莉·萨穆哈(Nathalie Smuha)指出:“现有的政府报告及人工智能政策仅在极少数情况下明确给出他们使用的定义,这增加了制定各领域人工智能治理政策和规范的难度。”[15]二是被迫给出宽泛的定义,如将教育人工智能界定为“在教育领域应用的人工智能技术及系统”[18]或“人工智能与教育科学融合形成的专项领域”[19]。缺乏清晰的定义意味着难以确定何时需要采取特定的监管措施来应对教育人工智能相关的挑战和问题。此外,这一情况也会导致各方在讨论教育人工智能问题时产生误解和混淆,增加制定有关治理政策和规范的难度,并可能阻碍跨国协调与合作。三是寻求规避界定人工智能概念的治理方法。欧盟《通用数据保护条例》专注于规定个人数据的收集、处理和存储方式,要求教育机构采取适当的技术与组织措施保护学生数据安全,并规定了学生与家长的权益。然而,尽管这种方法规避了教育人工智能的定义困难,但单独出台针对人工智能的治理手段只有在应对与其紧密相关且不存在其他技术应用的风险或利益时才是合理的,而目前教育人工智能乃至宏观的人工智能并不存在这样的特征[15]。

(二)伦理原则指代不明

伦理原则是教育人工智能软法治理的核心依托。虽然目前教育领域已基本确定了透明、可解释性、公平、隐私、安全性、问责制等共识性伦理原则,但这些共识性伦理原则的存在既导致教育人工智能软法治理方向不清,又导致其治理效果难以评价。

教育人工智能伦理原则的问题首先表现为原则自身的模棱两可。当前教育人工智能伦理原则多为模糊的、高层次的价值陈述,在实践中并没有提供具体的建议[4]。例如,“公平”原则往往不会指出人工智能教育应用的公平内涵及其对教育起点公平、过程公平、结果公平的选择,而让行为人自行决断[20]。又如,“福祉”原则虽然强调要维护利益相关者的教育权益,保证教育人工智能不会对其造成负面影响,但这一界定也没有对福祉的内涵进行明确解读——人工智能教育应用的福祉包含哪些方面,需要维护哪些教育群体的福祉,各群体的福祉是否存在矛盾。模糊的伦理原则无法作为有效指导教育实践的操作性指南。一方面,教育人工智能的开发者难以从抽象且模糊的道德价值观和原则中推导出具体的技术应用;另一方面,教育人工智能的监管者只能在他们认为合适的情况下阐释原则并界定存在争议的概念,无法形成明确的监管实施路线图,进而难以准确判断治理实践的效果。

教育人工智能伦理原则的问题还表现为原则之间互相矛盾。就原则设置来看,当前的伦理原则,如福祉、隐私、透明度、公平、自治等基本能够对应教育人工智能面临的伦理问题领域。但就结构而言,当前教育人工智能伦理原则通常以集合而非层次结构的形式出现,这种不包含价值排序的原则清单意味着每一项原则都应尽可能地实现[21]。这回避了治理实践中存在的价值争议:公平必然以个性化和准确性为代价,隐私会限制服务质量与效率,自治会限制福祉实现等[20]。具体来说,不同学生的背景、兴趣和能力存在差异,因此需要不同程度、不同类型的技术支持,完全的公平会忽视个性化的教育需求;教育人工智能系统通常根据学生的水平与需求进行个性化指导,这需要收集学生的大量数据,可能触及学生隐私;自治原则要求教育人工智能不能剥夺个体的自主选择权与决策权,但在某些情况下其又能做出比学生或教师更优秀的决策,导致福祉与自主的冲突。此外,教育人工智能伦理原则并列式呈现策略无法解决竞争价值排序的深层问题,也无法明晰治理的作用程度与优先事项。

(三)治理主体尚未形成多元共治合力

作为人工智能治理“他律”的执行者,治理主体的设定涉及治理责任的分配与治理方式的选择,对治理效果有着直接影响。尽管多元主体共治已成为教育人工智能伦理治理的发展共识,即在保证政府部门主导地位不变的前提下,授予其他利益相关主体相应权力,但在具体执行层面依然存在各主体内涵不明确、关系不对等、权责不清晰等问题,尚未形成治理合力。

治理主体尚未形成多元共治合力一方面表现为监管机构设置不明确。就治理主体而言,当前人工智能硬治理存在将人工智能治理整合到现有部门、授予新职能和建立新部门之间摇摆不定的问题[7]。当前,我国对教育人工智能采取的是多头治理,治理主体包括教育部、科技部、工信部、国家互联网信息办公室、国家市场监督管理总局等多个部门。虽然这一主体设置涵盖了教育人工智能设计、流通、应用的各个环节,但如此“九龙治水”的主体样态可能存在治理目标零散拼凑、治理资源分配不均、不同监管规则相互冲突的现实问题。鉴于多头治理可能存在的不足,有研究呼吁成立独立的人工智能监管机构,其主要观点是,人工智能治理涉及对技术、法律和道德综合因素的复杂理解,将它们整合为一个整体会有所帮助[16]。但这一观点自身面临诸多争议:第一,新的监管机构可能会重复现有机构的工作;第二,建立新的监管机构十分烦琐,且可能与其他监管机构的运作、权力、范围等产生冲突;第三,新的监管机构可能被视为现有治理框架的替代品,而不是执行这些框架的手段,甚至出现“监管俘获”的情况。

治理主体尚未形成多元共治合力另一方面表现在教师与社会参与的“结构性”缺位。在教师层面,尽管联合国教科文组织在《教育中的人工智能:可持续发展的机遇和挑战》报告中特别强调,除情感交流与关系构建、数智认知能力外,教师还需培养学生在智能环境下的价值判断与风险识别能力。我国在《教育部办公厅关于开展人工智能助推教师队伍建设行动试点工作的通知》中也提出内涵与之相近的“教师智能素养”概念,但当前关注更多的是教师如何运用人工智能技术改进教育教学,尚未将教师视为治理主体。在社会层面,当前教育人工智能伦理治理缺乏第三方专业研究机构的参与,第三方专业治理力量的支持不足,且研究者关注更多的是“人工智能+教育治理”,即如何应用人工智能技术赋能教育治理变革,而对教育人工智能伦理治理的研究大多停留在伦理问题识别、伦理问题成因、伦理原则构建等价值层面,尚未形成完整的研究框架,未能有效承担治理主体责任。

(四)治理方式的孤立性

除以伦理原则为核心的自律外,软硬兼备的人工智能伦理治理还有赖于以各方监管为基础的他律。理想的人工智能他律需要贯穿人工智能的设计、制造、使用等各个阶段,这要求治理方式保持主体整体性与过程连贯性。目前,教育人工智能主要采用的是基于风险的治理手段,尽管这一手段能够迅速应对人工智能教育应用的伦理问题,但在整体性与连贯性方面存在局限。

基于风险的治理手段关注教育人工智能技术的具体应用,聚焦监管应重点关注的特定案例及其造成的伦理问题[22]。如欧盟对瑞典一所学校引入面部识别系统监控学生出勤率进行处罚,英国对评核及考试规例局(Ofqual)的 A-Level 算法开展公平性审查等。尽管基于风险的治理手段能够灵活且有针对性地应对具体问题,但其仅是一种临时的、针对特定案例的事后监管。这一治理手段存在局限性的深层原因源于治理对象的界定困难,即不存在统一的教育人工智能。同时,教育人工智能引发的伦理风险是特定环境下的结果。就底层技术而言,基于规则与基于深度学习的 AIED 系统所面临的挑战不同,启用图像识别与处理自然语言的 AIED 系统所面临的挑战也不同。就应用场景看,教育人工智能已应用于学校、家庭、社会等多个场合,其中又包含了开发者、学生、教师、家长、学校管理者等诸多利益相关群体,进一步加剧了治理应对的复杂性。

为弥合基于风险的治理手段存在的不足,欧盟于 2021 年提出《人工智能法案》,其治理思路有两方面特点。一方面,与以往针对具体应用的纵向监管相比,该法案主张开展涵盖更多人工智能应用的横向监管;另一方面,该法案前置了基于风险的治理起点,要求人工智能系统在设计研发阶段便对已知和可预见的风险进行分析与识别,并评估应用后可能产生的风险。在此基础上,《人工智能法案》将人工智能应用划分为不可接受风险、高风险、有限风险、低风险或无风险几种类型,并针对各类型的人工智能应用采取完全禁止、严格管控、遵守透明度义务、不做干预的分级监管措施。不可否认,前置

基于风险的治理起点在一定程度上能够规避事后治理造成的不良影响,但技术的逻辑可塑导致技术在与社会的融合过程中充满不确定性,技术应用的不良后果很难在技术发展的早期做出准确预测[23]。因此,除基于风险的治理手段外,还需探索能够整合各治理主体的治理方式。

三、“软硬兼备”的教育人工智能伦理治理路径

现实挑战集中反映了当前全球教育人工智能伦理治理实践面临的关键问题。实现软硬兼备的教育人工智能伦理治理,前提在于对上述问题进行回应并探索有针对性的实践路径,做到治理对象清晰,伦理原则可用,治理主体明确,治理方式可靠。

(一)聚焦治理对象:对教育人工智能概念进行弹性界定

聚焦治理对象是开展教育人工智能伦理治理的前提要求。在当前治理面临的诸多挑战中,定义教育人工智能至关重要,因为其将决定治理生态系统的目标对象和作用范围。鉴于教育人工智能尚未实现概念清晰和术语一致性,可暂时对其进行弹性界定。

由于教育人工智能的多变性与多样性,为其给出全面的定义是困难的,但治理的紧迫性又需要确定的治理对象。面对理论困境与实践需要的矛盾,可采取“扩大概念边界,聚焦应用类别”的方法对教育人工智能进行弹性界定,以应对概念的动态性与特殊性争议。从扩大概念边界的角度看,人工智能对教育活动中的个人、群体乃至社会的影响,特别是负面影响,在短期和长期内仍不确定,也未被完全理解[15]。为此,现阶段教育人工智能的外延应当适当扩展,应用程序只要涉及算法或数据且与教育相关,就应被视为教育人工智能系统(AIED)[24]。

从聚焦应用类别的角度看,当前几乎所有教育人工智能应用都能够被狭义人工智能(narrow AI)和通用人工智能(general AI)①分类包含,而构建一个独立的强人工智能即使在未来也不太可能[25]。这意味着以应用对象为标尺寻找虚体的人工智能概念在目前是可行的。在扩大边界的基础上,韦恩·霍姆斯(Wayne Holmes)提出具象的教育人工智能分类法(见表 1)。该分类基于学生、教师、机构 3 个层面,能够全面地概括现有应用[26]。“扩大概念边界,聚焦应用类别”定义方式的优点在于其作为一种灵活的方法,能够明确指出需要治理的对象,且可以根据技术和社会发展及时更新。

表 1　教育人工智能分类法

类别	应用	
以学生为中心的人工智能	智能辅导系统(ITS) 辅助应用程序 辅助模拟 残疾学习者支持 自动文章生成(AEW)	聊天机器人 自动形成性评估(AFA) 基于对话的辅导系统(DBTS) 探索性学习环境(ELE) 终身学习助手
以教师为中心的人工智能	抄袭检测 学习资料管理 课堂监控	自动总结性评估 人工智能助教 课堂编排
以机构为中心的人工智能	招生 计划、时间安排 电子保安	风险学生识别 电子监考

(二)优化伦理原则:提供具有可操作性的原则指南

优化教育人工智能伦理原则是提升软治理效果的首要要求。现有教育人工智能伦理原则既无法塑

① 狭义人工智能指的是设计用于解决特定任务的 AI 代理程序,通用人工智能则是指能够解决多种给定问题的 AI 代理程序而无论任务或领域是什么。

造伦理治理的确定性,又无法应对伦理治理的不确定性。换言之,教育领域软治理缺少的并非宏观层面的伦理原则,而是伦理原则对治理实践的指导效力。未来,优化伦理原则可以从以下 3 个方面出发。

一是拓展教育人工智能应用的广度与深度。一方面,实现教育目标是构建教育人工智能伦理规范的首要前提[27]。当前研究侧重于讨论以学生为中心的教育人工智能应用,聚焦个性化学习、提升学习效率等狭义的教育目标的实现。未来,需要更多地关注以教师为中心的人工智能和以机构为中心的人工智能,为抄袭检测、智能助教、电子保安、电子监考等系统确定明确的应用目标,据此对伦理原则进行优化。另一方面,教育目标是教育人工智能伦理规范的方向标尺[4]。例如,定义教育人工智能"福祉"原则需要对什么是"进步"或"利益"有明确的理解。目前,以个性化教学为主题的教育人工智能系统大多以布鲁姆(Bloom)"精通学习"教学策略为起点,据此构建知识模型、学习材料与知识掌握测试[4]。但正如比斯塔(Biesta)所言,教育目的不能被学习目标取代,教育目的至少包括资质、社会化与主体化 3 个方面。未来,需要将教育人工智能应用深化至知识构建、社会化与自我实现三重维度。

二是明确教育人工智能伦理原则的内涵与范畴。在明确教育目的基础上,教育人工智能伦理原则需要更加具象化以消除自身的模棱两可与互相矛盾。一方面,要明晰教育人工智能伦理原则的内涵。教育人工智能伦理原则之间的矛盾消解需要澄清术语的歧义,如教育人工智能中的"隐私""公平"等原则具体指代什么,以及在不同的群体和背景下如何得到不同的解释。另一方面,要缩小教育人工智能伦理原则的作用程度。伦理原则想要发挥实际作用,必须相互权衡,如在隐私保护方面,既需要保障学生的隐私和数据安全,又需要充分利用数据进行模型训练和优化,在设计隐私保护策略时,需要考虑不同利益方的需求,做到原则的弹性适应。

三是前置教育人工智能伦理原则的作用范围。荷兰技术哲学家菲利普·布瑞(Philip Brey)提出"预知性技术伦理"理念,呼吁在技术研发阶段便考虑其进入社会后的可能影响,以最大限度规避伦理风险[28]。鉴于教育人工智能技术应用场景的复杂性,在明晰伦理原则的概念内涵与作用程度后,还需前置伦理原则的作用范围,在研发阶段就对教育人工智能应用可能出现的伦理问题予以关注,提早干预其在教育场景中的渗透方式。为此,可运用行动者网络理论构建包含教育目标与规律、风险与责任等多元要素的伦理清单,明确不同教育主体在人工智能应用场景中应当遵循哪些原则,从而前置伦理责任,将伦理价值渗透教育人工智能的设计阶段,消解教育人工智能伦理治理的"科林格里奇困境"(Collingridge's Dilemma)。

(三)锚定治理主体:落实"一核多元"的权责划分

未来,应落实一核多元的治理主体划分,扭转以往"科学家立项、相关部门出钱、政府批改、公众不负责任、人文社会科学专家善后"[29]的技术治理主体混乱状态,并明确各主体责任。

一方面,要明确"一核多元"的治理主体。其中,"一核"是前提,"多元"则是贯穿教育人工智能伦理治理的主线。"一核"是指政府在教育人工智能伦理治理结构中处于核心地位。《人工智能伦理治理标准化指南(2023 版)》根据人工智能的生命周期,将伦理风险划分为技术型、应用型、混合型 3 类,明确指出作为监管主体的政府部门应当在治理全周期的各伦理风险中发挥关键作用。"多元"是指要让学校、家庭、社会等利益相关者参与教育人工智能伦理治理。《中国教育现代化 2035》提出了"形成全社会共同参与的教育治理新格局"的战略任务。因此,在教育人工智能伦理治理中,政府要全面把控人工智能对教育实践的影响趋势,变革治理组织机构与治理政策;学校与家庭应提升风险意识与协同意识,审慎使用教育人工智能应用,及时识别、反馈伦理问题;企业应主动参与治理,承担构建可信赖的教育人工智能的技术责任。

另一方面,要落实各主体的治理责任。为使主体责任更加明确,教育人工智能伦理治理的责任分配应贯通伦理风险与伦理原则,并在此基础上构建整体性责任框架(见表 2)。具体来说,由技术型伦理风险带来的伦理问题,如当前自动作业批改系统或学习资源推荐系统缺乏透明性与可解释性,学生、教师以及教育管理者很难理解其是如何做出决策或建议的,进而导致对系统不信任,影响教学过程中的决策和交互。这一类型风险的责任主体是政府监管部门与技术开发企业,应主要从可问责性与透明度两个方面承担责任。由应用型伦理风险带来的伦理问题,如人工智能技术的应用将教育领域已经存在的"数字鸿沟"扩大为"智能鸿沟",进一步加剧了教育的公平问题。这一类型风险的责任主体包括政府监管部门、技术开发企业、学校与家庭、社会机构等多个群体,各治理主体应分别承担各自的责任。由混合型伦理风险带来人们对教育安全固有认知的改变,如受语料库影响,以 ChatGPT 为代表的生成式人工智能在便利知识搜集与获取的同时,还需要警惕"知识投毒"这一安全问题。为此,各教育人工智能伦理治理主体均应提升安全意识与能力,承担安全责任。

表 2　教育人工智能整体性责任框架

教育人工智能伦理风险分类识别与分析矩阵										
	人工智能伦理风险类别	人工智能伦理风险责任主体	主要治理原则							
			以人为本	可持续性	隐私	公平	共享	合作	可问责性	透明	安全
教育人工智能应用	技术型	技术主体	*	*	*	*	*	*	***	***	*
		监管主体	*	*	*	*	*	*	**	**	*
	应用型	技术主体	*	*	**	**	**	**	*	*	*
		应用主体	**	**	***	***	***	***	*	*	*
		监管主体	***	***	***	***	***	***	*	*	*
	混合型	技术主体	*	*	*	*	*	*	*	*	***
		应用主体	*	*	*	*	*	*	*	*	***
		监管主体	*	*	*	*	*	*	*	*	***

注:责任强弱由低到高分别为 *、**、***。

(四)创新治理方式:探索实验治理与混合治理的行动框架

创新治理方式是提升教育人工智能伦理治理效果的路径保障。任何新兴技术都可能在治理灵活性与监管确定性之间产生张力。当前基于风险的治理手段已无法满足教育人工智能伦理治理的实践要求,需要探索具有整体性与连贯性的治理方式。

一方面,应通过教育社会实验推动教育人工智能伦理治理常态化。不同于基于风险的治理手段对教育人工智能应用的被动与个别关注,教育社会实验可以在更大范围内探索智能技术治理的有效方案[13]。明确"一核多元"的治理主体后,各主体应在自己的责任范围内,积极探索教育人工智能伦理治理的有效策略:政府应推动制定教育人工智能伦理指南与法规,并对教育人工智能系统和平台进行监管与审查,牵头开展跨部门、跨主体合作,加强对教育人工智能伦理治理的协调与监督;学校应审慎引进人工智能管理系统,在效率与安全之间实现动态平衡;教师应参与技术伦理学习,负责任地在一线教学中使用人工智能技术;家庭应提升伦理风险意识,关注儿童日常使用的智能技术产品,不传播、不侵犯儿童数据与隐私,同时留意其可能存在的其他伦理风险并在必要时进行干预;科研工作者应从理论层面开展教育人工智能伦理治理研究,重点关注教育人工智能概念明晰、伦理原则内涵阐释与实践落地、治理模式探索与法律构建等难点问题,为教育人工智能伦理治理提供参考与依据;企业

应遵守法律与技术标准,并接受相关部门与社会的监督,同时探索将伦理原则转化为开发实践的技术操作。

另一方面,应构建软硬法混合治理的“中心—外围”模式。尽管加快构建并完善教育人工智能法治体系成为各方共识,但如何构建教育人工智能的专门立法尚没有明确思路。实际上,软硬兼备的教育人工智能伦理治理并非软法与硬法的机械结合,而是相互影响、相互渗透。同时,还需考虑教育人工智能全球治理的现实需要,“中心—外围”的治理模式正符合上述要求(如图 1)。“中心—外围”模式包括以下两个构建要点:批判性借鉴国际教育人工智能伦理治理研究成果,实现对国际一般人工智能伦理原则和教育人工智能伦理原则的教育性与本土性转化。同时,推进国际教育人工智能技术标准向国内硬法转化并做到标准互认。其中,政府应鼓励教育人工智能研发企业、科研机构、用户群体、社会组织参与对教育人工智能伦理的研讨,促进国内软法与国内硬法相互影响、转化;政府、企业、高校应加强协作,研制并出台教育人工智能伦理治理规范或文件,以期将治理偏好向国际软法渗透,争取教育伦理规范与技术标准制定的话语权。

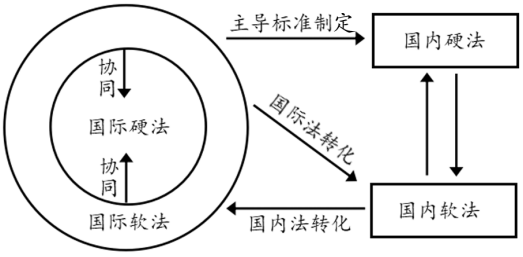

图 1　教育人工智能伦理治理“中心—外围”模式

四、结　　语

实现软硬兼备的教育人工智能伦理治理,关键在于回应当前治理对象界定不清、伦理原则指代不明、治理主体不明确、治理方式孤立的现实挑战,做到治理对象聚焦、伦理原则清晰、治理主体明确、治理方式有效。未来,除弹性界定教育人工智能概念、提升伦理原则实践指导力、落实“一核多元”的权责划分、构建实验治理与混合治理的行动框架外,也可探索其他行之有效的优化路径。但在探索优化路径的过程中需要处理好两对伴生矛盾。一是人工智能技术的全球性质与监管响应的本土性质之间的矛盾。尽管人工智能的国际合作与全球治理已成为政策制定者、学界与社会的共识,但国家间文化与观念的差异使得其对人工智能技术有着不同的诉求,进而形成不同的治理价值取向。《人工智能伦理治理标准化指南》亦指出,当前国际治理路线存在差异,难以推动形成全球共识。二是监管模式创新与守旧之间的矛盾。在技术治理方面,我们倾向于追求“创新”,认为每一项新技术的出现都必然具有破坏性、革命性、变革性和特殊性。基于这种理解,我们在悲观主义、技术恐惧症到技术乌托邦主义的钟摆运动中,试图寻求同样新颖、变革性和特殊的方法来治理它,进而可能导致监管过度或监管不足。

参考文献:

[1] UNESCO supports the definition and development of AI competencies for teachers[EB/OL]. [2023-11-20]. https://www.unesco.org/en/articles/unesco-supports-definition-and-development-ai-competencies-teachers.

[2] PAPYSHEV G, YARIME M. The state's role in governing artificial intelligence: development, control and promotion through national strategies[J]. Policy design and practice, 2023, 6(1): 79102.

[3] 吴红,杜严勇.人工智能伦理治理:从原则到行动[J].自然辩证法研究,2021,37(4):49-54.
[4] 白钧溢,于伟.超越"共识":教育人工智能伦理原则构建的发展方向[J].中国电化教育,2023(6):9-17,24.
[5] VAKKURI V, KEMELL K K, KULTANEN J, et al. Ethically aligned design of autonomous systems: industry viewpoint and an empirical study[J]. Electronic journal of business ethics and organization studies,2022,27(1):4-15.
[6] FLORIDI L. Translating principles into practices of digital ethics: five risks of being unethical[J]. Philosophy & technology,2019,32(2):185-193.
[7] RADU R. Steering the governance of artificial intelligence: national strategies in perspective[J]. Policy and society,2021,40(2):178-193.
[8] JOBIN A, IENCA M, VAYENA E. The global landscape of AI ethics guidelines[J]. Nature machine intelligence,2019,1(2):389-399.
[9] 朱明婷,徐崇利.人工智能伦理的国际软法之治:现状、挑战与对策[J].中国科学院院刊,2023,38(7):1037-1049.
[10] POLYVIOU A, ZAMANI E D. Are we nearly there yet? a desires & realities framework for Europe's AI strategy[J]. Information systems frontiers,2022,25(1):143-159.
[11] JAKOB M, MARIA A, FEDERICO C, et al. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation[J]. Minds and machines,2021,32(2):241-268.
[12] 中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》[EB/OL].(2022-08-04)[2023-11-20].http://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.
[13] 逯行,张春佳,黄仕友.智能时代教育主体的失范风险归因及其多层次治理研究[J].现代教育技术,2023,33(9):37-46.
[14] CARSTEN S B, ROWENA R, NICOLE S, et al. A European agency for artificial intelligence: protecting fundamental rights and ethical values[J]. Computer law & security review,2022,45(1):1-25.
[15] SMUHA N A. From a "race to AI" to a "race to AI regulation": regulatory competition for artificial intelligence[J]. Law, innovation and technology,2021,13(1):57-84.
[16] STAHL B C, ANDREOU A, BREY P, et al. Artificial intelligence for human flourishing-beyond principles for machine learning[J]. Journal of business research,2021,124(1):374-388.
[17] 屈静,刘凯,胡祥恩,等.对话式智能导学系统研究现状及趋势[J].开放教育研究,2020,26(4):112-120.
[18] 邓国民,李梅.教育人工智能伦理问题与伦理原则探讨[J].电化教育研究,2020,41(6):39-45.
[19] 蒋鑫,朱红艳,洪明.美国"教育中的人工智能"研究:回溯与评析[J].中国远程教育,2020(2):9-20,48.
[20] WHITTLESTONE J, NYRUP R, ALEXANDROVA A, et al. The role and limits of principles in AI ethics: towards a focus on tensions[C]//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society. New York:ACM,2019.195-200.
[21] ERICH P. From ethical AI frameworks to tools: a review of approaches[J]. AI and ethics,2023,3(3):699-716.
[22] VASILIKI K. From the "rush to ethics" to the "race for governance" in artificial intelligence[J]. Information systems frontiers,2022,25(1):71-102.
[23] COLLINGRIDE D. The social control of technology[M]. London: Frances Pinter,1980:1.
[24] HOLMES W, PORAYSKA-POMSTA K, HOLSTEIN K, et al. Ethics of AI in education: towards a community-wide framework[J]. International journal of artificial intelligence in education,2021,32(3):1-23.
[25] ZAWACKI-RICHTER O, MARIN V I, BOND M, et al. Systematic review of research on artificial intelligence applications in higher education-where are the educators? [J]. International journal of educational technology in higher education,2019,16(1):1-27.
[26] WAYNE H, ILKKA T. State of the art and practice in AI in education[J]. European journal of education,2022,57(4):542-570.
[27] 王佑镁,王旦,柳晨晨.从科技向善到人的向善:教育人工智能伦理规范核心原则[J].开放教育研究,2022,28(5):68-78.
[28] 顾世春.荷兰预判性技术伦理思潮研究[J].大连理工大学学报(社会科学版),2018,39(4):114-119.

[29] ZHANG J, ZHANG Z M. Ethics and governance of trustworthy medical artificial intelligence[J]. BMC medical informatics and decision making, 2023, 23(1): 7-22.

(责任编辑:杨慷慨　校对:张海生)

Ethical Governance of Educational AI: Real Challenges and Implementation Pathways

BAI Junyi

(*Faculty of Education*, *Northeast Normal University*, *Changchun* 130024, *China*)

Abstract: Ethical governance aims to ensure that the design, development, and application of educational artificial intelligence (AI) systems align with ethical, legal, and societal values to effectively empower AI technologies in the digital transformation of education. Currently, in view of the governance challenges of "knowing without doing", "knowing but not doing", and "drafting but not doing", the transition from "single soft law" to "soft-hard hybrid" has become a consensus and demand for the development of ethical governance in educational AI. Due to the uncertainty of AI ethics as an emerging topic, and the fact that policy formulation and academic research are still in their early stages, achieving a "soft and hard" educational AI ethical governance faces many practical challenges such as unclear definition of governance objects, unclear ethical principles, lack of diverse joint governance forces among governance entities, and isolated governance methods. The premise and key to achiev ethical governance of educational artificial intelligence that combines software and hardware is to flexibly define the concept of educational artificial intelligence, provide practical principles and guidelines, implement the division of rights and responsibilities of "a core-periphery", and explore the action framework of experimental governance and mixed governance.

Key words: ethical governance; educational governance; educational AI; soft law governance; hard law governance