博弈论视角下教育人工智能伦理风险治理困境及化解策略研究

陈翠荣，崔红岩

摘　要：当前教育人工智能伦理风险治理在伦理关系、师生权益、教育公平、情感价值及责权边界等方面陷入困境。从博弈论视角来看，教育人工智能伦理风险治理过程中存在着政府、研发企业、学校、教师及学生等多个利益相关主体，分别基于各自利益作出策略选择，治理困境正是利益相关主体间进行博弈的结果。要化解该困境，必须加快前置式教育人工智能伦理规则研制，预防潜在伦理风险；提升博弈主体对教育人工智能伦理风险的认知，促进伦理价值认同；畅通信息沟通渠道，调动利益相关者参与治理的积极性；强化教育人工智能伦理过程监管，提升伦理失范支付成本。

关键词：博弈论；教育人工智能；伦理风险；策略选择

中图分类号：G434　**文献标识码：**A　**文章编号：**1671-0169(2025)03-0131-14

DOI:10.16493/j.cnki.42-1627/c.2025.03.001

一、引　言

新一代人工智能技术的教育应用正以前所未有的深度与广度重塑着教育生态，从生成式人工智能驱动的自适应学习系统、多模态交互的虚拟现实教学到基于脑机接口的认知增强实验，技术赋能在重构教学流程、师生互动模式与教育资源配置机制等方面为教育现代化注入了强劲动能。然而，技术跃进与伦理失序的“双峰效应”也日益凸显：算法偏见加剧教育公平失衡、数据滥用侵蚀师生隐私权、技术黑箱消解教育主体性等伦理风险频发[1][2][3]。2023年10月，中央网络安全和信息化委员会办公室发布《全球人工智能治理倡议》，明确提出“坚持伦理先行，建立并完善人工智能伦理准则、规范及问责机制”[4]，“伦理先行”是人工智能教育应用的首要原则，也是实现教育与人工智能双向赋能、和谐共生的前提条件。

教育人工智能伦理是人工智能技术在教育领域应用过程中应遵循的道德原则、价值理念与行为规范，旨在保障学生、教师等教育主体的相关权益，促进人工智能技术在教育中的负责任和可持续发展。随着人工智能技术在教育领域的深度渗透，相关伦理问题引起了学界的关注[5]。当前，学者们围绕教育人工智能伦理风险问题从伦理学、管理学、社会学、教育学、哲学等视角开展了讨论，主要集中于伦理风险的内涵认定[6][7]、伦理风险的评估与管理[8][9]、伦理规范的指标构建[10]、伦理风险的表现及规避策略[11]等方面，为教育人工智能伦理建设提供了有益思路。

基金项目：国家社会科学基金项目“数字时代高校科技伦理协同治理体系研究”（BIA240134）

作者简介：陈翠荣，中国地质大学（武汉）教育研究院，chencr@cug.edu.cn（湖北 武汉 430074）；崔红岩，中国地质大学（武汉）教育研究院

值得注意的是，教育人工智能伦理风险治理的特殊性在于必须同时回应教育作为公共产品的公益性与技术作为资本载体的逐利性之间的根本冲突，而二者之间的协调难正是当前教育人工智能伦理风险治理的核心困境。其本质是政府、研发企业、学校等利益相关者围绕数据主权分配、风险成本转嫁及创新红利共享等的博弈，而不同博弈主体在目标函数、信息权力与风险承担能力等方面的非对称分布，以及传统治理模式的规则刚性、响应滞后与激励错配，极易陷入“监管失效—策略规避—风险外溢”的恶性循环。

博弈论作为解析复杂社会交互行为的理论工具，对于探究“教育人工智能伦理风险治理困境及化解策略”这一问题具有很强的适用性，可以为穿透技术治理中“制度刚性”与“市场弹性”的对抗逻辑提供有力的理论支撑，从而深入揭示教育人工智能伦理失序的生成机制与演化路径。同时，博弈论也为探究该问题提供了一种新的视角，有助于打破传统从上至下单一治理思维，为深化教育人工智能伦理研究提供了一种新方法。故此，本研究基于博弈论的视角，在剖析教育人工智能伦理风险治理面临的具体困境基础上，进一步厘清利益主体之间进行博弈的冲突逻辑与利益失衡机制，并提出化解该伦理风险治理困境的相应策略，以期为人工智能时代教育生态的伦理重构提供理论参照与实践进路。

二、博弈论视角下教育人工智能伦理风险治理面临的困境分析

教育人工智能的伦理治理困境，是既有治理机制在技术理性与教育价值的动态博弈中失效的集中映射。在技术资本扩张、教育效能诉求与伦理规范滞后的三重张力下，教育人工智能的深度应用不仅难以完全兑现其“公平赋能”的初始承诺，而且会因目标偏移与执行异化催生出系列伦理风险。这些困境的深层症结，在于利益主体间权力—责任的结构性错配：政府规制权威被技术黑箱消弭，学校监管职能向企业反向让渡，师生数据主权在资本逻辑中持续弱化。从算法黑箱中的主体性消解到数据殖民下的隐私失守，从技术崇拜加剧的教育鸿沟到规制滞后诱发的责任真空，伦理危机沿着“技术渗透—治理失灵—规则悬置”的链条蔓延，亟待通过博弈论解构其内在机理。

（一）主要困境的表现

1. 伦理关系：教育关系多元化与人的主体性消解。人工智能技术的深度介入重构了传统教育关系的边界，推动师生互动、家校协作与资源分配模式向多维化、智能化方向演进[12]。教育场景中多元主体形成新型交互网络[13]，表面上拓展了教育参与的民主性，然而既有治理框架的动态监管工具缺失与权责匹配失灵，使得技术驱动的多元化沦为权力结构隐性失衡的催化剂。当教育决策权向算法让渡却缺乏透明度审查机制、教学过程被智能系统主导使得师生仅有形式否决权时，教育关系即从“人与人的联结”异化为“人与技术的共谋”。技术企业凭借算法控制权成为教育规则的“隐形立法者”，而治理机制在数据主权界定、风险成本分摊上的模糊性，迫使师生在数据采集中处于被动地位[14]；课堂中人性互动的萎缩[15]与教育目标的功利化压缩，暴露出治理体系在约束技术权力扩张与捍卫教育人文内核上张力不足。这反映了工具理性对教育本质的殖民：既有规则既未能建立算法介入的阈值红线，亦难以重构“人本价值优先”的博弈均衡，教育从“培养人”的使命落入“规训数据载体”的技术陷阱。

2. 师生权益：数据算法之需与师生隐私侵害。人工智能技术通过海量数据采集与算法建模，推动教育走向精准化、个性化，学习行为追踪、情感识别、能力图谱构建等技术手段深度介入教学场景。数据驱动虽能在形式上为师生提供适配资源，却因隐私保护制度与监管工具穿透力的缺失暗含着隐私边界的系统性溃退：师生在技术裹挟下沦为透明化的数字客体，算法之需与隐私权益的冲突日益尖锐。一方面，智能教育系统以“服务优化”之名对师生数据的无差别采集使师生

被迫让渡隐私主权；另一方面，基于机器学习的学生学业风险预警系统、教师绩效评估模型等算法决策工具，因技术复杂性和商业保密性形成“解释鸿沟”，师生隐私权在算法黑箱中被剥夺。此外，教育科技企业凭借数据垄断地位，将师生隐私转化为资本增殖的“原材料”，而治理体系既未构建数据收益共享机制，亦未赋予师生有效的制衡手段，最终使“以人为尺度”的教育价值沦为技术利维坦的附庸[16]。

3. 教育公平：教育教学个性化与发展路径依赖及偏见。人工智能技术通过个性化学习系统与自适应算法，宣称要“为每个学生定制最优发展路径”，试图弥合传统教育中的标准化缺陷。然而，既有治理机制在数据多样性监管与算法审计工具上的缺失，使得这种技术驱动的个性化承诺背后，潜藏着算法偏见固化与资源分配异化的双重危机，当教育公平被简化为“数据适配度平等”，技术反而成为再生产结构性不平等的隐形推手[17]。个性化学习系统的训练数据多源自优势群体的行为样本，导致算法难以识别边缘群体的学习特征，数据殖民的隐性筛选使弱势群体被迫适应“不属于自己的最优路径”；自适应系统通过持续追踪学生行为强化初始判断，形成“数字茧房效应”，算法推荐实则加剧了刻板印象；智能教育硬件、高端算法服务的获取，高度依赖区域经济水平，技术红利反而拉大了教育鸿沟[18]。教育人工智能的公平性危机，实质上是社会结构性不平等在数字空间的投射，算法成为教育资源的“技术守门人”，弱势群体的发展权被进一步削弱甚至剥夺。

4. 情感价值：技术赋能教育与人文价值关照缺失。教育人工智能在重塑教育流程与治理模式的同时[19]，其技术理性与人文价值的深层冲突日益凸显。技术赋能通过数据化渗透将学生简化为数据节点，使情感需求、心理波动及价值观培育边缘化，而情感交互标准的制度性缺位与效能导向治理的单向度强化进一步挤压了人文价值的存续空间。哈贝马斯的交往行动理论指出，工具理性驱动的系统正侵入并殖民生活世界，导致社会关系被工具理性异化[20](P126)，而人工智能的介入则以一种新的技术媒介逻辑加速这一“殖民过程”，具体表现为智能课堂管理系统将师生互动简化为行为数据流、情绪识别技术将复杂的情感波动降维成生物特征图谱等，使“生活世界”中本应自由流动的沟通行动沦为“系统”规训的附庸。这一困境的深层逻辑可从马克斯·韦伯的工具理性扩张理论中得到解释：工具理性的过度膨胀会导致价值理性的式微[21](P56-57)。这在教育场域中具象化为教师为完成智能系统的量化指标而压缩师生情感交流，管理者依据算法生成的“学生数字画像”决策，形成技术闭环中的价值消解机制。技术赋能的高效便捷最终却使“技术崇拜”在治理失序中滋生[22]，典型症候表现为将自适应学习系统等同于因材施教，用情感计算算法替代共情教育，这种技术工具对其效用边界的突破，通过“数据闭环—算法强化—制度内化”的路径持续解构着教育场域中不可量化的情感价值和主体间性。这本质上反映了当前的治理体系既未能建立技术赋能与人文守护的均衡模型，亦疏于建构多元主体在情感价值维度上的利益协同机制。要破解这种异化，亟需在治理框架中嵌入技术伦理的反思性维度，建立能够捕捉教育温度的质量评估体系，使工具理性与价值理性在动态平衡中实现教育本质的回归。

5. 责权边界：技术飞速发展与伦理规制滞后性。教育人工智能的技术迭代速度正以“摩尔定律”式跃进重塑教育生态，深度学习、多模态交互、自主决策系统的突破性进展不断拓展技术应用的想象空间。然而，技术狂飙突进的背后是伦理规则与责任框架的严重脱节，教育人工智能系统已具备准自主决策能力，人类社会却仍未构建起与之匹配的权责认定体系，技术失控风险与责任真空地带持续扩大。伦理规制的滞后主要表现在三个方面：一是责任主体模糊化，教育人工智能系统的决策链条涉及研发企业、学校、教师等多方主体，但技术黑箱性导致过错归因困难，责任分散化使得受害者陷入“无处问责”的困境；二是法律约束碎片化，现行教育法规多针对传统教育场景设计，难以覆盖人工智能技术的特殊性；三是风险预防静态化，伦理审查多聚焦技术部

署前的安全评估，缺乏对系统进化中伦理偏离的动态监控。教育人工智能的责权困境，本质上是工业时代线性规制思维与数字技术指数级发展的根本性冲突。

（二）教育人工智能伦理治理中的博弈主体及其收益—支付表现

教育人工智能伦理风险治理的实施过程，本质上是一场复杂而微妙的博弈，涉及监管方、研发方和使用方等多个利益相关者的深度互动与策略选择。其中，监管方主要指各级政府，研发方主要指负责技术研发与供给的研发企业，使用方主要包括学校、教师及学生等。三方主体间的利益关系如图1所示。

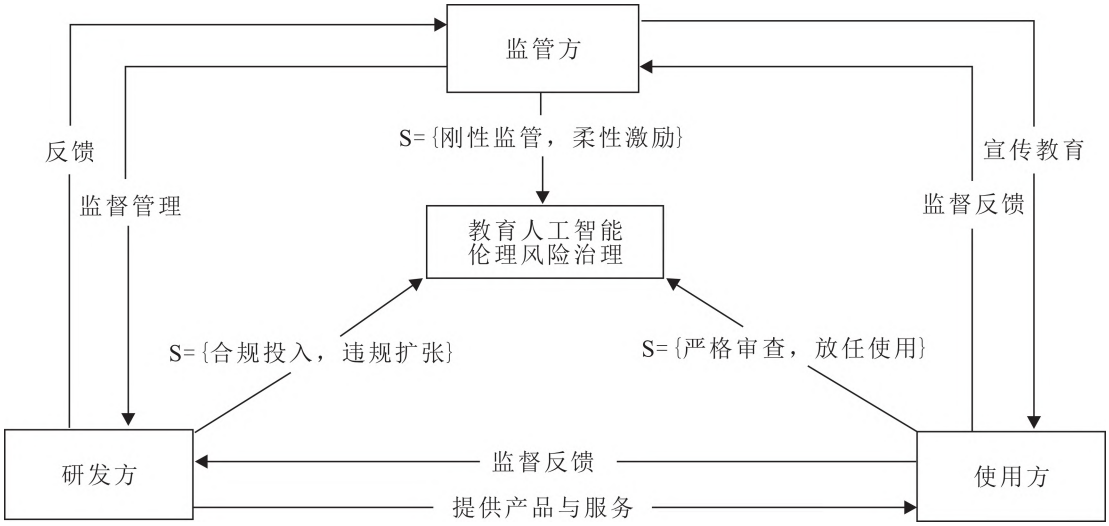

图1　教育人工智能伦理风险治理博弈主体间关系

在这场博弈中，各利益相关者基于理性决策原则展开策略互动，其行为选择由目标函数与支付结构的权衡所驱动。监管方作为规则制定者，核心目标在于通过伦理规制平衡技术创新与公共利益：一方面推动教育人工智能优化资源配置、弥补传统教育短板，另一方面需防范算法歧视、隐私泄露等风险以维持教育公平与政府公信力。其策略空间在“刚性监管”与“柔性激励”间摇摆，前者通过高强度审查与处罚提升合规率，以遏制算法黑箱的蔓延，但需承担全程监管费用、抑制技术创新的潜在成本，而当技术系统突破责权边界形成责任真空时，如果缺乏动态追踪能力则会陷入“规制失效”困境；后者以政策优惠激活市场活力，引导企业自我约束，却面临企业策略性套利的道德风险，有可能为数据殖民与情感价值异化留下了操作空间。研发企业作为技术逐利者追求最大化商业收益，其策略选择在“合规投入”与“违规扩张”间动态调整：前者意味着在通过技术创新获取市场份额与资本估值的同时，必须持续支付高额合规成本；后者则利用去标识化技术、信息技术壁垒、伦理规制滞后性等监管漏洞，枉顾师生权益、数字安全构筑数据滥用通道，以更低的成本实现技术升级、产品迭代、销售增长，无视其消解师生主体性、侵害其隐私、加剧人文价值荒漠化的背德之举。使用方（学校、教师、学生）的决策则嵌套于多层级委托代理框架。学校作为中间代理人，在政府监管压力与企业技术供给间寻求平衡：既需通过人工智能提升管理效能与教育质量，又须规避伦理事故引发的问责风险，其在“严格审查”与“放任使用”间进行策略选择。教师群体面临“效率提升”与“自主权让渡”的冲突，教育人工智能虽减轻机械性工作负担，却削弱教学设计主导权，导致部分教师在实践中可能会选择性执行，而教学互动被简化为算法优化的数据流，教师为维持数字绩效指标不得不压缩情感交互，这正是工具理性吞噬教育本质的微观写照。学生作为终端用户，在个性化服务需求与隐私权益保护间陷入“被动接

受”的囚徒困境，其弱势地位使其成为风险转嫁的主要承受者，算法偏见固化的发展路径与数据殖民下的隐私失守，使学生群体在技术赋能的表象下承受着最深层的公平性剥夺。

教育人工智能伦理风险的本质源于多元主体在非对称信息与动态博弈中的策略性失衡，当技术资本裹挟下的效率至上逻辑持续侵蚀教育的人文内核时，各利益相关者基于局部理性作出的占优策略选择，必然导致主体性消解、隐私泄露与公平性异化等伦理危机。既有治理框架的适配性断裂使伦理防线退化为策略博弈的灰色地带，亟需通过重构激励相容机制，将教育本质的守护转化为各博弈方的占优策略选择，最终实现技术向善与教育初心的统一。

三、教育人工智能伦理风险治理困境的博弈分析

在教育人工智能伦理风险治理中，从是否掌握核心决策权、能否主动博弈以及策略互动的直接性与强度来看，真正参与博弈对抗的决策主体主要包括政府、研发企业和使用者，使用者包括学校、教师、学生等，学校是其中的突出代表。各核心利益相关者的多元期望与责任交织成复杂格局，在伦理、技术、法律及社会的多维框架下基于自身利益进行策略性权衡，治理困境实质上反映的是各利益相关者不断进行博弈的结果。透过博弈论的视角，可以更清楚地剖析教育人工智能伦理风险治理之困境。

（一）政府与研发企业之间的博弈分析

教育人工智能伦理风险治理以社会主义市场经济体制为背景，其中政府与研发企业的博弈呈现出委托—代理关系嵌套于政策—市场博弈的复合性特征，其核心矛盾在于公共价值守护与技术资本扩张的深层张力。现实中比较常见的形式是，政府通过专项基金资助、政策引导、项目合作以及直接指定等方式，使被选择的研发企业成为技术创新的执行代理人，而政府则承担规则制定与公共利益守护的委托人角色。例如，教育部高等学校科学研究发展中心联合中兴通讯、辽宁向日葵、科大讯飞、浪潮通用软件有限公司等企业设立了一系列中国高校产学研创新基金专项计划[23]，通过资金注入与项目合作，明确要求企业聚焦人工智能、大数据等在教育领域的应用，实质上是以财政资源与政策红利为纽带，将企业的技术研发导向公共教育需求。

政府作为委托人，通过专项合作条件设定技术伦理的隐形边界，“数据安全”“算法透明”等条款，本质是以契约形式将伦理责任内化为企业研发的隐形约束；而企业作为代理人，在承接项目时既需完成技术创新目标，又隐含承担伦理合规义务。然而，这种委托关系中的权责失衡与信息不对称，恰恰成为伦理风险的滋生点[24]。《北京市推进中小学人工智能教育工作方案（2025—2027年）》提出要联合头部企业研发一批人工智能教育校企联合课程[25]，政策文本中“推动合规准入”“确保符合相关法律法规”的模糊表述折射出政府对技术黑箱的穿透性监管能力不足。缺乏具体场景的可靠标准与约束机制往往会使伦理治理陷入“有理念无路径”的真空状态，伦理治理效力严重依赖企业自律。当技术系统凭借自主决策能力突破传统责任框架时，教育关系中的主体性消解与责任真空便成为必然，这正是技术理性僭越教育本质的制度性缺口。教育人工智能系统以深度学习、知识图谱等复杂技术为基础，政府因专业能力局限难以动态追踪技术迭代中的伦理偏移，面临着技术黑箱遮蔽和数据孤岛割裂的信息困境。而研发企业一方面处于信息优势地位，另一方面受市场竞争压力、技术迭代速度、成本控制及商业利益最大化等多重因素驱动，往往通过合规性包装或技术性规避等手段弱化政策约束力。例如，将课堂脑电波监测系统包装为“专注力训练工具”规避审查，以生物特征数据过度采集为代价的技术渗透，不仅将师生隐私置于风险中，更通过情感计算算法对教学互动的机械化替代，加速消解教育场域中弥足珍贵的人文价值；利用学习者行为数据精准营销甚至通过算法诱导消费等，不仅导致学习者隐私数据的泄露，也使

算法违背了向上向善的基本要求[26]。此外，政府通过资金支持或政策红利，激励企业开展“负责任”的研发，但企业支付函数中市场占有率与资本估值的权重远高于伦理成本，政府与研发企业间实则形成一种“风险—收益”的非对称分配契约。而作为委托人的政府，在面临信息不完全与动态变化的环境下，即使耗费大量的人力、财力长期监测研发企业的伦理风险点，其效果也会表现出明显的滞后性与不确定性，这也为研发企业采取投机行为提供了可能性。

与此同时，政策—市场的动态博弈进一步加剧了伦理治理的复杂性。政府通过政策设定技术路线，力图以规则刚性引导市场走向，比如联合企业设立“云中大学”专项基金，明确支持“教育元宇宙”“知识图谱构建”等的研发与应用；而研发企业则以技术创新倒逼规则弹性，利用人工智能大模型的快速迭代，在尚未建立相关生成式内容审查标准前抢先部署存在伦理漏洞的智能系统，其隐含的算法偏见、数据隐患等伦理风险迫使监管被动跟进。在此过程中，技术资本通过构建教育资源的“算法守门人”角色，使弱势群体在数据样本偏差与算力资源垄断的双重挤压下陷入发展路径依赖，而规制滞后的合法性赋予则让技术鸿沟披上了教育创新的外衣。在这种“规制—创新”的循环博弈中，政策目标与市场理性形成非对称对抗：政府依赖企业实现教育数字化“显性政绩”，却因专业能力局限难以识破技术包装下的伦理漏洞；企业则借助政策红利扩张市场份额，通过数据垄断与算法霸权重构教育权力结构。这种对抗通过非对称风险转嫁，强化了风险—收益分配的扭曲，最终形成教育公平承诺与技术利维坦的现实悖论。

（二）研发企业与使用者之间的博弈分析

在教育人工智能伦理治理过程中，使用者（包括学校、教师、学生等）期望在保证隐私安全、认知自由和教学主导权的基础上，通过技术应用提升教育质量、效率与公平，这一目标的实现需要依托研发企业提供高效、符合伦理规范的技术解决方案。其中学校作为技术采纳的主体，与研发企业间通过采购协议、服务合同等形成了一种事实上的委托—代理关系，双方在数据控制权与伦理责任分配上展开不完全信息博弈。一项针对家长的调研揭示，“优质”智能教育产品需兼备“可期、可控、可信、持久”四大核心特质[27]，而2023年世界慕课联盟报告指出，人类期许教育人工智能在推进个性化、多元化教育模式的同时，直面数据信息安全挑战并加快伦理法制构建[28]，凸显了对教育人工智能应用的期望焦点。

研发企业在策略选择上往往与学校存在分歧。作为学校的技术代理人与师生事实上的技术支配者，研发企业的核心目标在于通过技术效能最大化与数据资产私有化，既驱动人工智能工具深度嵌入教育场景以提升用户粘性，又借助算法复杂度构建竞争壁垒，最终实现市场支配地位的持续性巩固。在此种目标的驱使下，部分研发企业滥用师生行为数据，甚至通过合规成本转移等手段将风险外部化。事实上，伦理风险治理需要持续投入高昂合规成本，包括可解释算法研发、数据多样性保障、隐私加密技术采购等，这些都会直接增加企业的运营成本，却难以转化为市场收益。研发企业忽视伦理风险治理的深层动因，根源正是在于成本收益失衡与市场反馈迟滞的博弈困境。

以浙江省金华市某学校的“智能头环”事件为案例进行分析[29]。在这一事件中，研发企业强脑科技（BrainCo）以“专注力训练”为名推广脑电波监测设备“赋思头环”，表面宣称数据仅用于教学优化，实则暗含侵犯师生权益，保留着商业数据挖掘功能；学校由于技术认知局限与资源依赖选择放任使用。在该伦理风险对抗中，研发企业强脑科技与购买使用“智能头环”的学校为参与人。以S_1、S_2分别表示研发企业和学校的策略空间，则双方的策略空间为S_1=｛合规投入，违规扩张｝，S_2=｛严格审查，放任使用｝。对于研发企业而言，相关参数可描述如下：C_1为合规需要的成本投入，如产品研发时增加伦理嵌入开发成本、可解释性算法开发成本等；R_1为违规扩张带来的利润，如云端存储学生注意力数据用于算法迭代优化、产品加快升级及带来的销售扩张、

数据资产附加值等；L_1为违规后被查处导致的损失，发生舆情危机后的品牌利益受损，如罚款、信誉度下降、市场退出等。其中，研发企业采取“合规投入”策略提供符合伦理规范的产品被视为理当如此，则将研发企业的额外收益记为0，又同时会增加成本投入C_1，因此该伦理风险策略选择的支付记为$-C_1$。对于学校而言，相关参数可描述如下：C_2为审查成本，如建立监督网站，成立审查机构，聘请第三方技术评估费用与时间成本等；R_2为带来的系列收益，如使用审查合格的产品优化课堂管理，提高学生注意力，提升学校声誉、获得师生信任与归属感等；L_2为伦理风险造成的损失，如隐私泄露引发的学校公信力受损、教师离职、学生转校、师生对学校归属感降低等。由此，可以得出博弈双方在伦理风险对抗中的收益矩阵（如表1所示）。

表1　研发企业(BrainCo)与学校之间的伦理风险博弈收益矩阵

博弈主体/决策		学校	
		放任使用	严格审查
研发企业	违规扩张	R_1，$-L_2$	R_1-L_1，R_2-C_2
	合规投入	$-C_1$，R_2	$-C_1$，R_2-C_2

对于研发企业强脑科技来说，当学校选择“放任使用”策略时，需要比较R_1与$-C_1$的大小作出策略选择。由于强脑科技选择“违规扩张”时，可通过“智能头环”的部署获取学生注意力数据，进而用于加快算法迭代和产品升级而获得额外收益，而选择“合规投入”必然要投入资金用于可解释性算法等合规技术的研发，因此$R_1>-C_1$，研发企业必然选择“违规扩张”。当学校选择“严格审查”策略时，需要比较R_1-L_1与$-C_1$的大小作出策略选择。在教育人工智能应用过程中，算法偏见、数据滥用等伦理风险的负面效应具有显著滞后性，面对伦理争议时研发企业对技术主权的掌控也可使风险转移。在“智能头环”引发社会公众对隐私安全的质疑时，教育局暂时叫停“智能头环”的使用并要求各学校展开自查。然而强脑科技早已通过先发优势完成市场渗透，以单个头环3 500元的价格在全国多所学校完成了部署，且以“技术中立”为由推卸伦理风险责任，使得缺乏独立评估能力的学校成为舆论追责的主要对象。强脑科技违规扩张所带来的额外利润R_1远远高于品牌声誉的损失L_1，$R_1-L_1>-C_1$，研发企业基于利益最大化选择了“违规扩张”策略。

对于学校来说，当研发企业选择“合规投入”策略时，需要比较R_2与R_2-C_2的大小来作出策略选择。由于“严格审查”策略下需要投入更多人财物用于建立监督评估机制，审查成本$C_2>0$，所以$R_2>R_2-C_2$，学校必然选择“放任使用”策略。当研发企业选择“违规扩张”策略时，学校需要比较$-L_2$与R_2-C_2的大小作出选择。在教育人工智能伦理风险治理中，严格审查往往需要投入高昂的成本。浙江多所购买使用“智能头环”的学校均表示对于此类智能产品没有一定的审查能力，若需严格审查就要投入高额成本C_2（比如委托第三方评估机构），才能达到一定的效果。学校严格审查时所获得的收益R_2，诸如管理效能的提高，学校声誉、师生信任感与归属感的提升等，鉴于其难以量化与效果滞后性的特点，使得学校往往会对其低估，因此在学校看来R_2-C_2的净值较低。与此同时，当前对教育人工智能算法偏见、数据滥用等仍然缺乏明确的技术规范与法律规章，加之其负面效应的延迟性、不确定性、难估量等，学校会认为隐私泄露等风险尚未造成实质性损害而低估L_2，比如使用“智能头环”的多所学校管理者后续纷纷表示，部署该头环不会造成什么危害。因此在学校看来，研发企业选择“违规扩张”时$-L_2>R_2-C_2$是显著的，学校就会出于利益考量选择“放任使用”策略。这样，研发企业强脑科技与学校基于各自的

利益最大化，就达成了“违规扩张—放任使用”的纳什均衡。

在委托代理关系中，研发企业凭借技术细节与算法逻辑的信息垄断优势，构建起不对等的博弈格局。教育人工智能的技术特殊性加剧了权力失衡：其一，算法可解释性的缺失使决策逻辑成为“技术黑箱”，如“智能头环”通过脑电波数据生成注意力评分的模型缺乏透明性，学校难以穿透性验证其伦理合规性；其二，数据采集的隐蔽性导致隐私保护机制失效，头环以“生物信号分析”为名持续收集学生神经活动数据，实则通过去标识化技术规避信息保护的法律法规，在合规外衣下构筑数据滥用通道。这种技术霸权直接催生了信号扭曲的“柠檬市场”，研发企业通过伦理认证标签传递虚假合规信号，而学校因无法破解算法黑箱与数据流转链条，被迫在认知鸿沟中作出采购决策。以“智能头环”为例，其技术设计本质是博弈规则的预埋：设备通过非侵入式电极采集原始脑电信号，但其背后的提取逻辑是使用者难以获悉的；同时，以算法优化之名聚合跨校数据构建商业模型，形成教育数据垄断壁垒。学校受制于技术迭代速度与验证能力缺失，决策仅能依赖企业提供的表层信息。需要指出的是，相较于学校的契约性委托地位，教师与学生更多处于技术消费者的被动境遇。教师虽被赋予“系统操作者”角色，但其决策空间被算法预设的工作流严重压缩，本质上成为技术落地中的“次级代理人”，如浙江头环事件中教师被迫将课堂评价权让渡给算法系统，教学节奏被数据指标主导；学生则彻底沦为数据供给端的“沉默客体”，即便感知隐私侵犯，也因缺乏集体行动能力与退出选择权被迫妥协。这种分层式权力结构，使得教育人工智能伦理治理困境进一步加深。加之技术锁定效应带来的高更换成本，学校陷入“部署—依赖—妥协”的恶性循环，最终形成资本与技术共谋的治理困局，强约束缺失与低追责成本持续削弱伦理治理效能。

（三）政府与使用者之间的博弈分析

政府是政策制定者与监管者，在教育人工智能伦理风险治理过程中，政府通过制定法律法规、构建伦理审查机制、实施惩处等手段对教育人工智能应用进行宏观调控与规范，而具体落实这些政策与规范，不仅需要研发企业的自觉遵守与配合，还需要使用者的积极响应与合规操作。因而在这一治理框架下，政府与使用者之间同样存在着相应的委托—代理关系。

作为委托人的政府期望重塑使用者的策略选择空间，使其在技术应用中自发趋近公共利益最大化，而非沉溺于局部效率的短期狂欢，通过制度引导与权责重构确保教育人工智能的实践应用，既释放技术赋能教育的提质增效潜力，又始终恪守教育公平、隐私安全及人的主体性价值等伦理原则。2023年7月国家网信办等七部门联合出台的《生成式人工智能服务管理暂行办法》，就对使用生成式人工智能服务提出了“遵守法律、行政法规，遵守社会公德和伦理道德”的要求[30]，彰显了使用者在教育人工智能伦理风险治理这一合作性活动中的关键作用。而使用者在技术实践中，一方面致力于在制度框架内行动，积极响应政府的政策导向，确保技术赋能的同时遵守伦理规范；另一方面，追求教学质量与效率的显著提升，在实际操作中易出于利益考虑而选择那些能够快速提升教学成效的教育人工智能产品，忽视其伴随的伦理风险。例如，越来越多的学校通过引入课堂反馈教育人工智能系统，实现了课堂反馈的高效、精准、动态、智能及自动化[31]，然而仅用简单的数据来表征学习者的情绪，有标签化教育主体的嫌疑，且严重威胁着师生的隐私安全。此外，教师对反馈结果与推荐策略的过度依赖阻滞教育智慧的生成积累，消解教育的主体性价值，人工智能对教学环节的职能替代，则稀释教学场域的情感联结[32]。事实上，政府虽出台伦理规范，但其原则性条款难以匹配技术快速演进与场景泛化的复杂性，监管资源的有限性更使实时全面监控沦为理想。而使用者作为技术落地的直接主体，本应是风险感知与反馈的关键节点，却在效能追求与技术认知局限的双重作用下，形成扭曲的信号机制：学校为凸显教学改革政绩，选择性忽视

算法偏见，教师因工具依赖默许数据过度采集，学生迫于服务绑定放弃隐私主张。这种“共谋性沉默”导致政府决策陷入信息迷雾，既无法穿透技术黑箱获取真实风险图谱，又因教育成效评估的滞后性和复杂性丧失治理主动权。伦理规制沦为静态文本，风险反馈链断裂为碎片化个体经验，进一步加大了政府治理的难度。

值得一提的是，由于核心的使用者包含学校、教师、学生三类群体，政府与使用者在教育人工智能伦理风险治理中的互动实质上是多层级委托代理关系下的博弈互动。政府委托到学校，再到教师以及学生，代理层级的增加导致伦理风险责任的稀释，也易引发利益的偏离。比如学校为追求“智慧校园”政绩指标，可能选择高数据掠夺性但功能强悍的产品，而忽视了主动保护师生隐私权益。在学校与教师的互动中，学校委托教师按照伦理规范和教学要求合理使用人工智能产品；在教师与学生的互动中，学生委托教师在不侵害隐私权益的情况下应用人工智能产品实现学习效果的提升，因此教师实则在委托代理链条中扮演双重代理人的角色。然而在教育人工智能伦理治理实践中，除委托人学校与学生的双重委托之外，教师出于自身利益需求更加关注人工智能产品应用通过个性化辅导方案对于教学成效的显著提升，以及自动批改作业、智能答疑等功能对于教学负担的减轻，这种利益的割裂就使得教师不可避免地忽视隐私保护与算法偏见等伦理风险，偏离委托人的目标。

四、博弈论视角下教育人工智能伦理风险治理困境的化解策略

（一）加快前置式教育人工智能伦理规制研制，预防潜在伦理风险

制度变迁理论认为，技术创新的成果需要制度创新与变迁的巩固，有效的制度能够降低交易成本，节约交易费用，提高资源配置效率[33](P93)。在教育人工智能领域，前置式伦理规制作为一种预防性制度安排，能够明确各利益相关者的权利与义务，降低因伦理风险而产生的交易成本，包括法律诉讼、声誉损失、用户信任度下降等。教育人工智能的伦理风险具有强隐匿性与高扩散性，传统的“事后纠偏”治理模式难以应对技术指数级迭代的挑战。前置式伦理规制可以重构研发企业的策略选择空间，将伦理约束内化为技术研发的初始参数，通过降低研发企业自主探索合规路径的试错成本、压缩违规收益与放大风险损失等，重构支付函数以驱动纳什均衡向伦理优先态迁移，破解“监管追赶技术”的困境，实现治理成本最小化与风险防控最优化的均衡。

具体来说，加快前置式伦理规则研制需结合国际经验，构建兼具创新包容性与风险可控性的治理框架。一是要构建动态演化的伦理标准体系。英国政府2023年发布的《促进创新的人工智能监管方法白皮书》[34]提出“原则导向、灵活适应”的动态监管模式，强调通过风险分级机制平衡创新与责任。例如，将教育人工智能技术按应用场景划分为“基础级”“高风险级”与“关键级”，对不同级别设定差异化合规要求。这种基于原则规制的模式折射出英国法律传统“自由裁量”的灵活性与追求实用有效的政策思维的结合。而面临智能技术快速发展与复杂度与日俱增的现实境况，中国亟待调整治理模式以更具适应性和灵活性，诸如可以基于“适度预防”原则，借鉴英国风险分级机制背后的设计逻辑等。同时，英国白皮书提出通过“监管沙盒”（Regulatory Sandbox）支持企业测试高风险技术，在受控环境中优化伦理设计。此机制依托英国行业协会高度自治的社会治理传统，在中国则需依托党委领导下的多元共治格局，在学习借鉴中重构适配路径，比如，在风险分级中嵌入“立德树人”教育目标的刚性约束。据此可建立风险预测驱动的标准生成机制，基于技术成熟度在不同阶段预判伦理风险谱系；建立技术—伦理协同设计框架，推动“伦理嵌入开发”成为技术必选项，通过明确可解释性算法的核心指标使伦理合规成为产品上市的硬性准入

门槛，比如对决策逻辑的可追溯性和偏见识别覆盖率设定相应的标准；运用同态加密技术实现数据可用不可见，压缩企业违规利用数据的收益空间。此类技术性工具可系统性改变博弈支付函数，推动纳什均衡向“合规投入—严格审查”转移。二是建立激励相容的博弈规则，通过调整研发企业的支付函数，使其自利行为自动符合伦理目标。比如前置伦理审计费用可抵扣企业所得税、开源核心算法可换取政府采购优先权等，引导研发企业从“被动合规”转向“主动投资”。三是开发穿透式监管技术工具，通过构建政府主导的技术反制能力来破解“技术黑箱”导致的信息权利失衡。研发开源算法验真平台，允许监管部门对教育人工智能模型进行决策溯源，追踪不同群体学生的算法待遇差异；植入监管接口，在不获取原始数据的前提下通过分布式验证确保伦理合规。

（二）提升博弈主体对教育人工智能伦理风险的认知，促进伦理价值认同

教育人工智能伦理风险治理的深层突破，不仅取决于制度设计的完备性，更要以博弈主体的认知跃迁为前提。当前，政府、研发企业、使用者对伦理风险的认知割裂导致治理陷入“共识赤字”，亟需通过认知协同重构博弈均衡。澳大利亚高等教育质量管理与标准署（TEQSA）在《人工智能时代的评估变革》中提出技术伦理治理的核心在于“构建透明、包容且可持续的评估文化”[35]，这与认知协同的目标不谋而合。人的行为和决策受到其认知结构的影响[36](P5)，提升博弈主体对教育人工智能伦理风险的认知，意味着培养其对技术伦理风险的正确理解和评估能力，使其客观权衡技术利弊，从而明智决策。只有各方主体突破短期利益博弈的狭隘性，协调平衡好“价值理性”与“工具理性”[37]，达成对技术伦理价值的系统性共识时，治理模式方能从“策略性对抗”转向“价值性协同”。这一过程需通过认知重构工具、制度激励网络与文化浸润机制的三维联动，将伦理准则从外部约束转化为内生行动逻辑。

首先，博弈主体的认知偏差集中体现为技术效能崇拜与数据权力依赖，需构建穿透性认知干预体系以破解技术迷思与权力幻觉。对企业技术负责人进行教育人工智能伦理课程培训与教育场景伦理风险评估考试，降低企业伦理认知盲区引发的合规成本，重点解析与考察算法歧视的形成机制、隐私泄露的链式反应，并将认证结果与企业征信等级挂钩。通过引入第三方伦理审计机构进行年度能力复检与嵌入式监测工具防止培训流于形式。其次，促进教育人工智能伦理风险认知需与利益激励相耦合。TEQSA建议将伦理合规纳入教育机构的质量评估体系，对采用透明化技术方案的高校给予资金倾斜，这种“资金—合规”的绑定机制可以使学校更加明确教育人工智能风险防范重要性的同时提升其审查收益，从而将伦理价值转化为博弈主体的占优策略。据此，政府采购时在招投标评分体系中可以增设“伦理效能系数”，量化合规投入为市场份额竞争力。同时通过构建“伦理效能系数”的量化评估工具、设立伦理承诺追溯系统等防范指标虚化风险。最后，伦理价值认同的终极目标，是形成超越博弈计算的教育技术文化共同体，通过文化浸润可以培育教育人工智能伦理的公共话语场域。设立由政府、研发企业、学校、师生等多主体组成的省级教育人工智能伦理委员会，对高风险应用实施一票否决制；发起“技术谦抑性”教育行动，通过校长宣言、教师工作坊、学生辩论赛等，重建“育人本位”的技术文化共识。最终，通过认知干预降低企业合规成本、制度激励提升学校审查收益、价值共识放大违规损失，系统性改变博弈支付矩阵，瓦解原有非合作均衡的基础。

（三）畅通信息沟通渠道，调动多利益相关者参与治理的积极性

教育人工智能伦理风险治理的复杂性，源于多方利益相关者的目标异质性与信息孤岛效应。当前，政府、研发企业、使用方等主体间存在信息沟通壁垒，导致治理陷入“碎片化应对”困境，亟需通过信息共享机制重构参与式治理设计，破解博弈中的“猜疑链”并激发协同治理动能。在非对称信息博弈论中，信息不对称是导致博弈结果偏离最优解的重要原因[38](P177)。欧盟《人工智

能法案》[39]就强调了跨部门协作与信息共享的重要性，要求打通信息壁垒以有效识别、评估和控制风险。这为破解治理困境提供了关键路径：畅通信息沟通渠道，既能提升技术透明度、增强公众对教育人工智能技术的信任和理解，又能促进各利益相关主体的合作与共识，推动多方主体在信息对称基础上共同制定和执行伦理规范，形成多元共治的格局。

具体而言，一是要打造权威信息共享平台，作为政府、学校、研发企业等关键参与者的信息交汇点。此平台应定期披露算法逻辑、数据源及安全评估等核心信息，并采用区块链存证技术确保披露数据的不可篡改性，以及建立信息分级披露制度来平衡透明度与企业权益。二是要优化信息反馈机制，确保用户（教师、学生、家长等）反馈渠道畅通无阻。设立专项反馈平台，鼓励用户积极报告使用中的问题与建议，并实行快速响应与公开处理机制，对有效反馈给予正向激励。除此之外要规避反馈噪音干扰，开发智能分类算法自动识别重复或无效投诉，引入举报溯源系统打击恶意虚假反馈。通过高效沟通与透明反馈流程，加深各方对教育人工智能伦理问题的认识，激发其参与治理的热情与动力。三是构建多方共治的治理机制。与英国白皮书类似，欧盟《人工智能法案》设计了多元主体在“契约”下进行共治的新型监管模式“监管沙盒”，且提出要开发监管沙盒相关信息的专门平台，使相关主体能够通过平台实现多元主体的积极对话与反馈。基于此可以成立跨领域治理委员会，制定伦理准则与标准、处理争议与投诉，同时建立双向沟通桥梁与动态考核机制防止共治形式化，借助“治理效能数字沙盘”模拟提升参与有效性。这样，通过穿透式信息工具降低信息不对称系数，系统性压缩企业违规操作的期望收益，同时提升多方协同治理的边际效用，最终实现“充分披露—深度参与—有效制衡”的演化稳定策略。

（四）强化教育人工智能伦理过程监管，提升伦理失范支付成本

教育人工智能伦理风险的治理效能，根本上取决于违规行为成本支付与潜在收益之间的正向差异。当前，伦理监管滞后与处罚软化导致失范成本显著低于收益，“理性违法”成为优势策略，亟需构建全周期穿透式监管体系与梯度化惩罚机制，重塑博弈主体的支付函数，使伦理合规占优。监管经济学指出，有效的监管能够纠正市场失灵，保护公共利益。欧盟《可信赖的人工智能伦理准则》提出要建立问责制度，以为人工智能系统的可靠、持续使用提供保障，这其中包含了要对算法、数据、运行进行全过程评估的重要内涵。受启蒙思想、法治传统以及多元文化观念的影响，欧盟的人工智能治理强调透明度、可解释性和问责制，加之社会对隐私保护的高度关注，为其严格监管提供了合理性[40]。针对教育人工智能的技术黑箱与数据流动性特征，传统“事后处罚”模式难以追溯违规源头。过程监管可以通过实时抓取算法决策链的关键节点数据，将监管介入点从“结果端”前移至“过程端”，压缩企业的策略性操作空间。而提升伦理失范的显性成本（如罚款）与隐性成本（如声名狼藉），可以将违规策略的期望收益由正转负，推动博弈均衡向合规方向迁移。

首先，构建全周期动态审计体系。欧盟《人工智能法案》将人工智能系统按风险等级划分为“不可接受风险”“高风险”“有限风险”与“最小风险”四类，并基于风险分级开展“事前—事中—事后”的全生命周期监管，引入“监管沙盒”、售后监控系统等确保及时响应和处理潜在问题，极大提高了伦理治理的实时性和有效性。这一分类监管思路与中国《生成式人工智能服务管理暂行办法》中的分级、分类原则高度契合，中国可借鉴其“动态清单”机制，细化教育领域高风险人工智能判定标准。据此，可建立全过程的动态审计体系：技术训练阶段要求研发企业提交训练数据多样性报告并实施伦理预审制度；技术部署阶段植入可解释性监管接口实时追踪算法决策逻辑；技术迭代阶段强制报备重大算法更新与伦理影响评估。其次，实行梯度化惩罚机制，运用经济手段强化伦理规范执行。根据违规性质与危害程度实施多级处罚制度，对违规企业实施罚

款、强制开源争议算法模块等处罚措施；建立伦理信用体系，将合规情况纳入信用评价，信用良好的主体给予政策支持和优惠，信用不良的主体进行限制和惩戒，形成有效的激励和约束机制。最后，建立穿透式监管技术栈。在保护用户数据隐私的前提下，通过分布式节点验证算法合规性，将算法决策关键数据上链存证，确保追溯不可篡改；建立人工智能伦理风险预警平台，基于大数据分析预测违规热点，提前介入高风险点，并及时开展对相关主体的伦理行为监管。至此，通过扩大违规损失、降低合规成本，同时创造合规溢价，系统性扭转“违规收益大于合规成本”的原始博弈结构，从而使教育人工智能伦理过程监管更具成效。

五、结　语

教育人工智能伦理风险治理对于新时代推进技术赋能与教育价值融合具有关键性战略意义。教育人工智能的深度应用一方面革新了教育模式，但另一方面其引发的伦理失序问题亟需突破传统治理范式的局限。从博弈论的视角看，当前教育人工智能伦理风险治理在伦理关系、师生权益、教育公平、情感价值与责权边界等方面存在困境，治理困境的深层症结在于政府、企业、学校等主体的目标冲突与信息非对称性，具体表现为伦理规制滞后、认知割裂、监管碎片化及支付成本失衡等。为此，需立足我国教育治理体系，强化前置式伦理规制的预防效能，深化多元主体对教育人工智能伦理的价值共识，构建信息共享平台提升技术透明度与治理参与度，通过全周期动态监管提升违规成本。唯有将伦理约束内化为技术研发与应用的底层逻辑，方能实现教育人工智能从“效率至上”向“价值向善”的范式转型，为技术赋能教育现代化提供坚实的伦理根基与制度保障。

参考文献

[1] 谢琦，余日季，蔡苏.GenAI技术在教育评价中的算法偏见：表现、成因与对策[J].现代教育技术，2025(1).
[2] 刘骥，丘霖.生成式人工智能嵌入教育应用的风险生成及其规制[J].现代远距离教育，2024(4).
[3] 朱珂，张斌辉，张瑾.教育数字化转型中师生主体性的缺失风险与复归策略[J].电化教育研究，2024(4).
[4] 中央网络安全和信息化委员会办公室.全球人工智能治理倡议[EB/OL].https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm，2024-12-16.
[5] 郭庆，吴砥.国际视野下人工智能教育应用伦理风险与治理策略[J].比较教育研究，2025(1).
[6] 胡小勇，黄婕，林梓柔，等.教育人工智能伦理：内涵框架、认知现状与风险规避[J].现代远程教育研究，2022(2).
[7] 王佑镁，房斯萌，柳晨晨.风险社会视角下教育人工智能伦理风险分类框架研究[J].现代远距离教育，2024(3).
[8] 王佑镁，倚杨莹，柳晨晨.基于风险矩阵的教育人工智能应用伦理风险评估[J].现代远程教育研究，2024(6).
[9] 李焕宏，薛澜.生成式人工智能应用的使能型风险规制——以高等教育应用为例[J].清华大学教育研究，2025(1).
[10]杨俊锋，褚娟，张斌贤.人工智能教育应用伦理规范指标构建研究[J].电化教育研究，2024(10).
[11]杨俊锋，李世瑾.如何创新治理人工智能教育应用：场景化协同治理框架与实施路径[J].教育发展研究，2025(3).
[12]李亚东，阎国华.人工智能赋能高校思想政治教育的内在逻辑与路径设计[J].江苏高教，2024(2).
[13]李腾子.数字教育时代的高校师生互动与关系重塑[J].中国电化教育，2024(9).
[14]陈港，孙元涛.数智时代学生的主体性反思与重构——基于人技关系的思考[J].中国电化教育，2023(10).

[15]辛继湘,李瑞.人是技术的尺度——智能教学中人的主体性危机与化解[J].中国电化教育,2023(7).
[16]杨宁霞,唐爱民.教育数字化转型中的“数字利维坦”风险及其规制——基于风险社会理论的视角[J].中国电化教育,2024(9).
[17]吴砥,吴河江.通用大模型教育应用的潜在风险及其规避——基于技术伦理的视角[J].华东师范大学学报(教育科学版),2024(8).
[18]朱恬恬,杨菲.高等教育与数字经济耦合发展的困局及“双适应”进路[J].中国地质大学学报(社会科学版),2024(5).
[19]曾丽渲,邢鸿飞.数字技术赋能高等教育现代化转型的内生困境与发展策略[J].江苏高教,2024(8).
[20]Habermas, J. *The Theory of Communicative Action*[M]. Thomas Mc-Carthy, tr.. Boston: Beacon Press, 1984.
[21][德]马克斯·韦伯.经济与社会(上卷)[M].林荣远,译.北京:商务印书馆,1997.
[22]周明鹏.智能算法技术赋能高校思想政治教育供需互契研究[J].高校教育管理,2024(1).
[23]教育部高等学校科学研究发展中心.2025年中国高校产学研创新基金——数字新兴技术专项申请指南[EB/OL].https://www.cutech.edu.cn/detail/46-546, 2025-03-12.
[24]江喜林,胡蝶.数字化背景下政府参与产业链协同创新的演化博弈研究——基于“链长制”的视角[J].中国地质大学学报(社会科学版),2024(4).
[25]北京市教育委员会.北京市推进中小学人工智能教育工作方案(2025—2027年)[EB/OL]. https://www.beijing.gov.cn/zhengce/zhengcefagui/202503/t20250310_4029667.html, 2025-03-12.
[26]王思北,阳娜,周琳,等.大数据“杀熟”不能再“杀”了,算法推荐不能乱“推”了[N].新华每日电讯,2022-01-07(007).
[27]吴文涛,于浩然,刘翠,等.究竟什么是“好”的智能教育产品——家长视角的文本扎根分析[J].中国电化教育,2024(12).
[28]世界慕课与在线教育联盟秘书处.高等教育数字化变革与挑战——《无限的可能:世界高等教育数字化发展报告》节选五[J].中国教育信息化,2023(1).
[29]邱晨辉.专家谈“智能头环事件”:学生被机器监测不符合教改方向[EB/OL]. https://edu.youth.cn/jyzx/jyxw/201911/t20191111_12116180.htm, 2025-01-16.
[30]中华人民共和国国家互联网信息办公室.生成式人工智能服务管理暂行办法[EB/OL]. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm, 2025-02-11.
[31]赵瑞斌,杨现民,张燕玲,等.“5G+AI”技术场域中的教学形态创新及关键问题分析[J].远程教育杂志,2021(2).
[32]黄荣怀,张国良,刘梦彧.面向智慧教育的技术伦理取向与风险规约[J].现代教育技术,2024(2).
[33][美]道格拉斯·C·诺思.制度、制度变迁与经济绩效[M].杭行,译.上海:格致出版社,2008.
[34]UK Government. *Ai Regulation:A Pro-innovation Approach*[EB/OL]. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, 2023-08-09.
[35]TEQSA. *Assessment Reform for the Age of Artificial Intelligence*[EB/OL]. https://www.teqsa.gov.au/about-us/news-and-events/latest-news/assessment-reform-age-artificial-intelligence, 2023-11-30.
[36][美]罗伯特·索尔所,奥托·麦克林,金伯利·麦克林.认知心理学[M].邵志芳,译.上海:上海人民出版社,2008.
[37]陈翠荣,李海龙.数字化赋能创新创业教育生态系统建设:价值、逻辑与路径[J].高等工程教育研究,2024(3).
[38]张维迎.博弈论与信息经济学[M].上海:格致出版社,2012.
[39]European Parliament News. *AI Act:A Step Closer to the First Rules on Artificial Intelligence*[EB/OL]. https://ww. europarl. europa. eu/news/en/pressroom/20230505IPR84904/ai-act-a-step-closer-to-the-firstru

les-on-artificial-intelligence, 2023-05-21.
[40]杨耀,施筱勇.新兴技术伦理治理的国际实践与启示[J].中国科技论坛,2025(1).

Research on the Governance Dilemma of Ethical Risk of Educational Artificial Intelligence and Resolution Strategies from the Perspective of Game Theory

CHEN Cui-rong, CUI Hong-yan

Abstract: The governance of ethical risk of AI in education faces dilemma in ethical relationship, teacher-student rights and interests, educational fairness, emotional value, and the boundary of responsibility and power. From the perspective of game theory, the ethical risk governance process involves multiple stakeholders including the government, research and development enterprises, schools, teachers and students, etc., who make strategic choices based on their respective interests. Therefore, the governance dilemma is essentially the result of the game among the stakeholders. To resolve this dilemma, it is imperative to accelerate the development of front-loaded educational AI ethical regulations to prevent potential ethical risks; enhance the rational cognition of the game subjects and promote the recognition of the ethical value of educational AI; establish smooth information communication channels to mobilize stakeholders' enthusiasm for governance participation; strengthen the supervision of the ethical process of educational AI and increase the cost of ethical misconduct.

Key words: game theory; artificial intelligence in education; ethical risk; strategic choice

（责任编辑　周振新）