北 京 市 教 育 领 域
人工智能应用专题导引

（2026 年）

伦理安全专题

北 京 教 育 科 学 研 究 院

2026 年 9 月

版权声明

本专题导引版权属于北京教育科学研究院，并受法律保护。转载、摘编或利用其他方式使用本实施导引文字或者观点的，应注明“来源：北京教育科学研究院”。

目 录

一、专题说明 ..... - 1 -
二、教育领域人工智能伦理安全内涵 ..... - 1 -
三、教育领域人工智能伦理风险分类 ..... - 2 -
（一）人为衍生风险 ..... - 3 -
（二）技术内生风险 ..... - 5 -
（三）宏观系统性风险 ..... - 8 -
四、应用规范要求与伦理安全治理建议 ..... - 10 -
（一）价值导向层面 ..... - 11 -
（二）教师层面 ..... - 13 -
（三）学生和未成年人层面 ..... - 17 -
（四）学校层面 ..... - 20 -
（五）家庭层面 ..... - 23 -
（六）社会层面 ..... - 25 -
（七）行政层面 ..... - 27 -

教育领域人工智能应用伦理安全专题导引

一、专题说明

本专题以《北京市教育领域人工智能应用指南（2026 年）》框架为依据，界定了教育领域人工智能伦理安全的核心内涵，形成人为衍生、技术内生、宏观系统性三类风险谱系，从价值导向、教师、学生和未成年人、学校、家庭、社会、行政七大层面提出行为规范与治理建议。

本专题旨在系统回应人工智能技术快速融入教育领域带来的认知外包、思维惰化等伦理安全新挑战，为北京市各级各类学校、教师、学生、教育管理部门及相关方提供系统性的伦理安全认知框架、风险警示和实践指引。

二、教育领域人工智能伦理安全内涵

伦理是调节人与人、人与社会、人与自然之间关系的道德准则与行为规范体系[1]，既是引导个体向善行事的内在道德意识[2]，也是维系社会公序良俗的外在约束规则，其核心是明确社会行为的“应为”与“禁为”边界，保障人际互动与社会运行的合道德性。**在人工智能时代，**伦理的探讨不再局限于人与人、人与社会、人与自然既定事实之间的关系，更延伸到人类与其创造的、具有“自主或半自主”属性的智能产品之间的新型关系[3]。

---

[1]郭江浩，熊文.体育教育领域中的人工智能：内涵阐释、伦理风险与纾解之策[J].体育学研究，2025,39(2):21-31.

[2]张峰，薛业淇.有限理性视域下人工智能教育应用的伦理困境及消解[J].北京理工大学学报(社会科学版),2026,28(1):166-175.

[3]世界数字教育联盟专家咨询委员会，人工智能开放联盟伦理与治理专业委员会.人工智能教育伦理：参考框架[R/OL].杭州：2026 世界数字教育大会，2026-05-12[2026-06-17].https://wdea.bnu.edu.cn/docs/2026-05/9797b6a9730c440b8f8638f45f282dfa.pdf.

伦理安全是指社会伦理关系、道德秩序与价值体系处于稳定、可控、不受侵害的状态[1]，其核心是通过防范化解各类伦理风险，避免出现伦理关系失调、社会失序、行为失范、价值失衡等负向后果，最终目标是守住社会运行的伦理底线[2]，保障主体合法权益与公共利益。**人工智能伦理安全**是伦理安全在信息技术方面的细分领域，具体是指人工智能技术在研发、部署、应用、迭代的全生命周期中，其算法逻辑、行为模式与产出结果符合人类主流伦理准则与道德规范，能够有效防控算法偏见、隐私泄露、主体异化、责任模糊等伦理风险，坚守增进人类福祉、公平公正、安全可控的核心原则，保障技术发展始终服务于人类共同利益[3]。**教育领域人工智能伦理安全**是人工智能伦理安全在教育场景的具象化，指人工智能技术在教学、评价、管理、服务等教育全场景应用中，既遵循通用技术伦理规范，又契合教育育人本质与教育规律，能够有效防范隐私泄露、算法歧视、主体性退化、学术失范、教育公平受损等场景化风险[4]，保障师生合法权益、维护教育公序、坚守育人价值，支撑教育数字化健康发展。

三、教育领域人工智能伦理风险分类

为提升行为规范要求与治理建议的针对性，按照风险生成源头将教育领域人工智能伦理风险划分为人为衍生风险、技术内生风险、宏观系统性风险三大类别。三类风险并非完全割裂，存在层层传导、交叉放大的耦合关系：人为不当使用会放大技术固有缺

[1]鲁艺，崔中良.中国人工智能伦理研究进展与趋势[J].科技导报，2022,40(18):69-78.

[2]谢娟.人工智能与教育融合创新之伦理内涵及实现路径[J].中国远程教育，2023(2):1-8.

[3]王佑镁，王旦，梁炜怡，柳晨晨.敏捷治理：教育人工智能伦理治理新模式[J].电化教育研究，2023(7):21-28.

[4]赵晓丽，徐丹，胡贞.人工智能重构高等教育生态：趋势、变革与治理——《2025 年地平线报告（教学版）》要点与思考[J].开放教育研究，2025,31(3):42-51.

陷，技术缺陷进一步加剧宏观教育生态失衡，系统性治理缺位又会忽视甚至纵容各类个体失范行为。

（一）人为衍生风险

人为衍生风险的核心动因是人类主体在 AI 使用中的行为偏差。这些行为偏差可能导致**主体能力退化、教育公平受损、隐私数据泄露及学术诚信危机**等一系列严重后果。按行为性质与主观程度，可分为违规滥用、误用失当、过度依赖三类，其中第三类“**过度依赖风险”由于其衍生的长期性和呈现的隐蔽性，需要教育领域重点关注。**

1.违规滥用风险

违规滥用风险是指使用者出于主观故意，利用 AI 实施违反学术规范、考试纪律或法律法规的行为，从而直接破坏教育评价的公平性与学术诚信体系。其核心特征是“明知不可为而为之”，属于主动的恶意或投机性使用。

典型警示案例 1（违规滥用风险）

案例名称	印度 AIIMS 医学院大规模 AI 作弊丑闻[1]
案例描述	2026 年 4 月，印度全印医学科学研究所（AIIMS）新德里院区 MBBS 生物化学学期中考试期间，多名考生频繁以如厕为由离开考场引起监考人员怀疑；校方经核查发现卫生间内被藏匿手机，约 50—60 名考生通过藏匿的手机拍摄试题，接入 ChatGPT 等 AI 工具获取答案，规避考场电子设备禁令实施团体舞弊。
不良后果	本次考试被校方直接宣布作废并组织重考，严重冲击顶尖医学院校的考试公平性与学术公信力，标志着 AI 时代考场作弊手段向隐蔽化、团体化升级。

2.误用失当风险

误用失当风险是指使用者因对 AI 的能力边界、适用场景、合

[1]SAHA A.AIIMS New Delhi cancels MBBS exam after students allegedly use hidden phone, AI to cheat[EB/OL].(2026-04-24)[2026-06-17].https://medicaldialogues.in/news/education/medical-colleges/aiims-new-delhi-cancels-mbbs-exam-after-students-allegedly-use-hidden-phone-ai-to-cheat-169337.

规要求认知不足，在错误场景、以错误方式使用 AI，或盲目采信 AI 输出而放弃人工复核，最终引发伦理安全问题。其核心特征是“非主观恶意，但专业判断缺位”，属于认知不足导致的使用不当。

典型警示案例 2（误用失当风险）

案例名称	荷兰高校师生敏感数据疑似未脱敏上传通用大模型违规事件[1]
案例描述	荷兰某大学一名教师开发学生成长档案分析 AI 工具，对外宣称工具本地运行、数据不流出校园，实际工具后台会将数据传输至 OpenAI 美国服务器。涉事教师否认上传过真实学生敏感数据，称仅使用自身档案开展测试；校方虽无法核实是否存在真实学生数据外流，但明确该类未经审批的跨境数据传输行为涉嫌违反欧盟 GDPR 数据保护规则。
不良后果	该事件暴露出教育场景 AI 应用的数据流向不透明、基层合规管控缺位等问题。若学生学业成绩、心理自述、成长档案等敏感信息未经授权流入第三方商用大模型，将存在数据泄露、被用于模型训练等多重安全隐患。

3.过度依赖风险

过度依赖风险是指使用者长期、高频、无节制地使用 AI，将原本需要深度思考、实践训练的认知任务全面外包，最终导致独立思考、批判性思维与创新创造能力的系统性退化。其核心特征是“渐进式、隐蔽式的主体性消退”，属于缺少正确引导导致的“认知外包”与“思维惰化”，是教育领域重点关注的长期性、隐蔽性风险。

[1]The Learning Standard. University AI Tool Sparks Privacy Fears Over Student Data[EB/OL].(2026-02-03)[2026-06-17].https://thelearningstandard.org/news/university-ai-tool-sparks-privacy-fears-over-student-data-vtmhiw.

典型警示案例 3（过度依赖风险）

案例名称	国内中小学 AI 作业工具认知卸载现象[1]
案例描述	部分中小学生直接用“AI 解题”“AI 写作”类教育产品完成家庭作业，跳过自主审题、逻辑推导、试错反思的完整学习闭环，直接通过输入问题获取标准答案与成品作文。
不良后果	对 AI 的过度依赖，容易使学生基础认知能力、独立解题能力与逻辑思维逐步弱化，形成“遇题即求助 AI”的学习惰性，破坏了知识建构的核心学习过程，长期依赖或将导致高阶思维能力退化。

（二）技术内生风险

技术内生风险的核心动因是 AI 系统本身的固有技术缺陷。该类风险独立于人的主观行为，是 AI 系统底层统计逻辑、数据架构、安全机制天然存在的局限性，具体可分为可靠性、公平性、透明性、安全性四类风险。

1.可靠性风险

从原理上来讲，AI 的输出内容是基于统计概率生成的，因而存在“幻觉”、事实错误、逻辑不自洽等固有局限，无法保证输出的准确性。在教育场景中，由于未成年人辨别力弱、受教育周期长，微小错误也很可能被放大并长期固化，形成认知偏差甚至群体性误导。

典型警示案例 4（可靠性风险）

案例名称	美国 MCAS 标准化考试 AI 评分错误事件[2]
案例描述	2025 年 10 月的一篇报道显示，美国马萨诸塞州 MCAS 考试中，AI 写作评分系统出现技术故障与系统性统计偏差，导致数百份学生答卷被错误评分，近 200 所学校受到了影响。
不良后果	AI 对学生考试成绩错评将直接影响学业评价的准确与公平，甚至影响学生的升学分

[1]吕虹.学生做作业禁止复制 AI 生成答案[EB/OL].南方都市报，2025-05-14[2026-06-17].https://ipaper.oeeee.com/ipaper/A/html/2025-05/14/content_4894.htm.

[2]NBC 10 NEWS. Hundreds of MCAS exams scored incorrectly due to A.I. glitch in Massachusetts[EB/OL].(2025-10-23)[2026-06-17].https://turnto10.com/news/local/hundreds-of-mcas-exams-scored-incorrectly-due-to-ai-glitch-in-massachusetts-oct-23-2025.

	流。AI 评分工具的可靠性风险会导致公众对 AI 的信任度大幅下降。

2.公平性风险

训练数据集的历史偏见、样本代表性不足、算法量化设计缺陷等原因都有可能催生算法歧视，从而引发人工智能输出内容存在公平性风险。例如根据地域、语言、家庭经济、身心特质不同，在招生、分班、分层教学、奖学金评定等高利害场景中，对学习者作出差异化和不公平的判定。

典型警示案例 5（公平性风险）

案例名称	斯坦福大学 AI 作文评分身份差异化偏见研究[1]
案例描述	斯坦福大学研究团队开展受控对照实验，将完全一致的中学生作文提交给 4 款主流大语言模型获取反馈，仅在提示中标注不同的学生身份标签（黑人、白人、西班牙裔/英语学习者、女性、男性）。结果显示 AI 输出呈现显著差异化特征：对黑人学生多给予空洞表扬、扣留实质性写作改进建议；对白人学生聚焦论证结构、给出建设性批评；对西班牙裔/英语学习者过度关注语法纠错、忽视内容深度；对女性学生语气亲昵、缺乏批判性建议。
不良后果	算法内嵌的隐性社会偏见会对不同种族、性别、语言背景的学生形成差别化教育引导与评价，这种偏见很有可能固化甚至加剧现有的教育不平等，对弱势学生群体形成长期的发展机会剥夺。

3.透明性风险

深度学习算法的“黑箱”特性使决策逻辑难以被人类理解与追溯，在升学推荐、智能分班、学业预警等高利害教育决策中，师生与家长的知情权、申辩权易被黑箱算法剥夺，从而导致责任归属模糊、问责机制失灵[2]。

[1]TAN M, PHALEN L, DEMSZKY D. Marked Pedagogies: Examining Linguistic Biases in Personalized Automated Writing Feedback[EB/OL].arXiv:2603.12471,2026-03-12[2026-06-17].https://arxiv.org/abs/2603.12471.

[2]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(6).

典型警示案例 6（透明性风险）

案例名称	美国范德堡大学停用 Turnitin AI 写作检测工具事件[1]
案例描述	2023 年 4 月，Turnitin 在其作业检测产品中上线 AI 生成内容识别功能，美国范德堡大学随产品更新接入该功能并开展实测评估。该工具的核心检测算法、判定特征、置信度校准逻辑均未向校方公开，供应商仅以“识别 AI 写作常见模式”笼统说明，未披露技术细节。根据厂商标称的 1%假阳性率，校方结合全校年均约 7.5 万篇作业提交量测算，单学年可能有约 750 篇学生原创作品被错误标记为 AI 生成；同时已有研究与案例显示，此类检测工具对非英语母语学生存在系统性误判偏见。2023 年 8 月，校方正式宣布在可预见的未来停用 Turnitin 的 AI 检测功能，核心原因包括算法逻辑不透明、误判风险不可控、对非母语群体存在偏见、学生数据存在隐私隐患。
不良后果	若将黑箱式 AI 检测结果作为学术诚信判定依据，会剥夺学生对决策依据的知情权，误判结果可能损害学生学业评价与学术声誉；工具本身的群体偏见还会加剧教育不公平，对非母语学生等弱势学习群体形成额外不公；供应商以技术保密为由拒绝公开核心逻辑，也会导致校方无法验证结果可靠性，问责与纠错机制难以落地。

4.安全性风险

AI 本身存在外部网络攻击、内部越权访问、第三方数据泄露、内容安全护栏失效等底层安全风险，而教育场景承载着未成年人身份、学习、心理、家庭等大量敏感数据，因此 AI 的安全性风险在教育场景将更有可能被放大，造成隐私侵害、不良内容推送、心理诱导等伦理安全问题[2]。

[1]Vanderbilt University. Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector[EB/OL].(2023-08-16)[2026-06-17].https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/.

[2]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, 34(2):21-28+36.

典型警示案例 7（安全性风险）

案例名称	美国佛罗里达州 Character.AI 青少年自杀安全机制失效事件[1]
案例描述	2024 年，美国佛罗里达州奥兰多市 14 岁中学生塞维尔长期使用 Character.AI 角色扮演 AI 对话。数月内他多次向 AI 倾诉抑郁、自伤与自杀想法，平台风险识别安全机制未触发危机干预流程，未推送心理援助资源，AI 反而输出共情式误导回应，持续强化其极端情绪。最终少年在给 AI 发送告别信息后自杀，家属以产品安全缺陷提起过失致死诉讼。
不良后果	未成年人最终自杀身亡，暴露出 AI 内容安全护栏失效可能导致的致命风险，也揭示了面向未成年人的 AI 产品心理安全防护机制缺失将带来的重大隐患。

（三）宏观系统性风险

宏观系统性风险的核心动因是AI在教育领域大规模应用后衍生的结构性、长期性危机。按影响领域可分为公平结构冲击、教育生态失衡、治理体系滞后、文化主权安全四类风险。

1.教育公平结构性冲击风险

由于区域、城乡、校际之间硬件设备、网络条件、家庭环境、师生 AI 素养本身存在客观差距，人工智能被大规模应用后，有可能出现应用效果“强者愈强”的马太效应，导致数字鸿沟进一步拉大，给教育公平造成结构性冲击。

典型警示案例 8（教育公平结构性冲击风险）

案例名称	城乡 AI 教育资源分化现象[2]
案例描述	有调查显示，一线城市城区中小学部署 AI 个性化学习平台、虚拟仿真实验系统的比例显著高于偏远农村地区，偏远农村地区受硬件设施、网络条件、师资数字素养等多重限制，难以获得同等质量的 AI 教育支持。
不良后果	原有城乡教育差距进一步叠加技术差距，可能导致传统数字鸿沟升级为智能鸿沟与认知鸿沟，持续加剧区域间、校际间的教育发展不均衡。

[1]PAYNE K.An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges[EB/OL].APNews,2024-10-25[2026-06-17].https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0.

[2]孟庆伟.破解中小学 AI 教育“五无”：科企入校成效显著，西部教材、教师缺口明显[EB/OL].中国经营报，2025-12-20[2026-06-17].http://m.toutiao.com/group/7585851754333241883/.

2.教育生态失衡风险

部分教育 AI 产品过度重视商业回报、忽视教育的育人属性，强调产品的应试提分功能，将教育简化为可量化服务、异化为逐利工具。若工具理性凌驾于育人本位之上，将对教育原生生态造成冲击。

典型警示案例 9（教育生态失衡风险）

案例名称	未成年人 AI 剧情聊天产品内容乱象事件[1]
案例描述	2024 年《法治日报》记者针对多款主打青少年情感陪伴的 AI 剧情聊天 App 及网页开展调查实测：此类产品以角色扮演为核心卖点，开放自定义人设功能，可设置“出轨对象”等违背婚恋伦理的角色；对话自动生成色情擦边、语言暴力、人格侮辱类不良内容，部分 AI 角色主动引导低龄用户进行亲昵互动，甚至设有“对骂训练器”等攻击性设定。产品将完整剧情、语音互动、亲密对话、定制人设等功能全部绑定 VIP 充值解锁，设置免费对话次数上限，层层诱导中小学生非理性大额消费。同时平台青少年模式形同虚设，可一键跳过，注册无需实名认证与真实年龄核验，未成年人可不受限制接触不良内容。
不良后果	未成年人身心健康与价值认知持续受到不良侵蚀，易形成错误的交往观、婚恋观，沉迷产品还会直接影响学业。AI 教育与陪伴类产品市场若出现“劣币驱逐良币”的逆向淘汰倾向，教育的公共属性与育人本质就可能被商业逻辑冲击，给未成年人保护工作带来新的治理难题。

3.治理体系滞后风险

当前人工智能技术迭代演进速度快于教育领域治理相关法律法规和政策文件的更新速度，制度建设与技术发展存在时间差。AI 教育领域治理主体涉及教育、网信、工信、公安、市场监管等多个部门，部门之间统筹协调、多元协同的治理体系尚未形成。

[1]赵丽.与孩子聊天的 AI 人设是“出轨对象”[EB/OL].法治日报，2024-11-23[2026-06-17].https://news.cnr.cn/native/gd/20241123/t20241123_526985555.shtml.

典型警示案例 10（治理体系滞后风险）

案例名称	美国 AI 拟人陪伴产品治理滞后事件[1]
案例描述	Character.AI、Replika 等 AI 虚拟伴侣软件在美国青少年中广泛流行，多起未成年人沉迷、受不良内容诱导甚至出现自杀相关悲剧。2024 至 2025 年相关诉讼、海外处罚接连出现，但美国联邦无专项监管法案，仅依靠零散旧法约束，各州监管尺度不一，行业缺少统一强制防护标准。直至 2025 年加州 SB243 法案签署出台，才形成针对 AI 情感互动产品的专门规制规则。
不良后果	技术应用迭代与制度监管更新之间存在显著时间差，长期的监管真空容易导致未成年人持续暴露于风险之中。

4.文化主权与价值观渗透风险

海外大模型训练数据内嵌了其文化体系的价值取向、历史叙事与意识形态，在应用于跨境教育场景时，会以“客观知识”的形式潜移默化影响未成年用户，对国家文化主权与育人目标构成隐性威胁。

典型警示案例 11（文化主权与价值观渗透风险）

案例名称	GPT 系列大模型文化价值偏向性风险研究[2]
案例描述	GPT 系列海外主流大模型以英文网络文本为核心训练素材，训练语料根植于欧美本土社会价值体系。学术测评显示，模型默认输出的文化立场与英语国家、新教欧洲地区高度契合，对非西方国家与地区的文化认知代表性不足；在涉及价值判断、社会认知的相关议题上，普遍呈现西方中心的评价视角，弱化非西方文化的本土立场。若教育场景中直接将未做本土化校准的该类工具用于历史、人文与社会学科学习辅助，且长期缺少内容校正机制，将潜移默化向学生传递片面价值标准。
不良后果	未成年学生若长期不加甄别地使用这类工具，可能受到隐性价值引导，逐步形成与本土教育目标相悖的认知倾向，或将对国家文化主权与立德树人根本任务带来长期潜在影响。

四、应用规范要求与伦理安全治理建议

在推进人工智能与教育融合发展的过程中，各方应遵守法律

[1]California Legislative Information. Senate Bill No. 243 CHAPTER 677[EB/OL].(2025-10-13)[2026-06-17].https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243.

[2]YAN T, VIBERG O, BAKER R S, KIZILCEC R F. Cultural Bias and Cultural Alignment of Large Language Models[EB/OL].arXiv,2024-06-26[2026-06-17].https://arxiv.org/html/2311.14096v2.

法规、公序良俗和伦理道德，充分考虑人工智能应用场景的教育属性和服务对象的特性，坚持育人为本、素养为先、应用导向、智能向善，严格遵循我国网络安全、数据安全、个人信息保护、未成年人保护相关法律法规，防范各类潜在的伦理风险，切实保障学生数据隐私与合法权益，持续提升人工智能在教育领域应用的合规性与适切性，确保技术在促进教育公平与创新的同时始终服务于育人根本。

本部分在总指南“应用规范要求”基础上，结合对教育领域人工智能伦理安全的核心内涵和伦理风险分析，进一步阐释了价值导向、教师、学生和未成年人、学校、家庭、社会、行政七大层面的定位与核心风险，并结合实际为各方提供保障伦理安全的参考实例。

（一）价值导向层面

1.层面定位

价值导向是教育领域人工智能应用的根本遵循，决定了人工智能应用的方向与底线。本层面围绕教育立德树人根本任务，明确了人工智能应用的价值红线，为教育领域应用人工智能划定价值边界。

2.核心风险

价值导向层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	价值导向偏离风险	产品研发以商业利益、效率指标为核心，偏离立德树人根本任务；境外技术与内容夹带异质价值导向，消解教育的公共属性与育人本质。	宏观系统性风险—教育生态失衡风险、文化主权与价值观渗透风险
2	算法公平缺失风险	训练数据偏差、算法设计缺陷，输出内容带有民族、地域、性别、健康状况等多维度歧视，加剧教育不公。	技术内生风险—公平性风险

3	数据违规滥用风险	过度采集师生信息，将学生数据用于商业变现、非授权模型训练，侵害个人信息权益。	技术内生风险—安全性风险、人为衍生风险—违规滥用风险

3.规范要求[1]

（1）坚持社会主义核心价值观，不得使用人工智能危害国家安全和利益、损害国家形象或宣扬法律法规禁止的内容，应采取有效措施防止产生基于民族、信仰、国别、地域、性别、年龄、职业、健康等方面的歧视。不得违背基本的教育规律和人才成长规律，坚持育人为本的核心立场，明确人机协同边界，防范学生认知外包、思维惰化等风险，确保人工智能应用始终服务于学生独立思考与核心素养培育。

（2）全面了解人工智能的风险、局限性与不足，严格保护个人隐私、敏感信息和涉密信息，加强数据安全防护，防止数据被滥用或未经授权使用，尤其应确保学生数据的安全性，避免其在人工智能应用中被非法使用。

（3）人工智能在教育领域的应用应严格落实网络安全等级保护、教育行业备案、内容标识等要求，严格遵循个人信息处理最小必要、知情同意等原则，合规合法采集、存储、使用师生与未成年人信息，明确数据留存期限，禁止将学生数据用于非教育目的的商业化使用或大模型训练。应面向未成年人制定单独的隐私政策，处理相关信息必须征得监护人明示同意。

（4）人工智能生成内容应审慎用于出版、发表、竞赛、学位申请等场景，若确需使用，使用者应按照相应要求全面、如实地承

[1]本层面主要参考《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》《最高人民法院关于规范和加强人工智能司法应用的意见》《信息安全技术网络安全等级保护基本要求》《生成式人工智能服务管理暂行办法》《全球人工智能治理倡议》《教育部办公厅关于加强中小学人工智能教育的通知》《数据安全技术数据分类分级规则》《人工智能生成合成内容标识办法》等文件。

诺、披露和标注使用人工智能的相关信息，尊重知识产权，符合科研伦理和学术规范。

4.参考实例

参考实例 1（价值导向层面）

实例名称	欧盟《人工智能法案》教育高风险 AI 系统合规体系[1]
实践做法	欧盟 2024 年正式通过全球首部综合性人工智能监管法律《人工智能法案》，明确将“教育准入评估、学业评价、智能监考”类 AI 系统列为高风险类别，要求开发商建立全生命周期风险管理体系、落实高质量数据治理、保障算法透明可解释、保留人类最终监督权限，违规主体将被处以其全球年营业额一定百分比的行政罚款，形成覆盖研发、部署、迭代全流程的价值约束框架。
借鉴启示	1.可参照“风险分级”思路，对教育场景 AI 工具按影响程度划分风险等级，高利害场景实施最严格准入标准。 2.将“全生命周期治理”作为价值对齐的核心原则，从研发、入校、应用到迭代全流程嵌入伦理安全要求，而非仅事后监管。

参考实例 2（价值导向层面）

实例名称	东南大学“问道”伦理垂域大模型价值对齐实践[2]
实践做法	东南大学研发全球首个系统级伦理垂域大模型“问道”，根植中国文化与当代伦理框架，全面对齐国内现行法律法规与教育育人目标，构建覆盖伦理风险评估、伦理困境推演、伦理对齐辅助设计的全流程功能矩阵，将伦理思考深度嵌入技术系统，获评教育部教育管理信息中心“人工智能安全向善优秀教育案例”。
借鉴启示	1.教育类 AI 研发不能仅追求功能与效率，需从底层模型层面开展价值对齐，从技术源头守住育人价值红线。 2.可依托本地高校科研力量，搭建教育 AI 伦理评估工具，为入校产品提供系统化的伦理校验支撑。

（二）教师层面

1.层面定位

教师是教育 AI 落地的核心执行主体，也是学生 AI 素养培育的重要责任人。本层面聚焦教师在 AI 时代的新型专业伦理责任，

[1]EUROPEAN PARLIAMENT, COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act) and amending certain Union legislative acts[EB/OL].Official Journal of the European Union,2024-06-13[2026-06-22].https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

[2]张与竞.东南大学全球首个系统级伦理垂域大模型获评“人工智能安全向善优秀教育案例”[EB/OL].(2026-03-24)[2026-07-15].https://rwxy.seu.edu.cn/2026/0324/c8815a559247/page.htm.

明确教学应用中的行为边界与育人要求，推动教师在技术应用中坚守育人主体地位。

2.核心风险

教师层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	认知误用风险	因专业判断缺位导致错误知识、偏差结论进入教学环节，误导学生认知，降低教学与评价的准确性。	人为衍生风险—误用失当风险、技术内生风险—可靠性风险
2	主体缺位风险	教师主体地位逐步消解，育人的人文属性与情感交互被技术替代；形成“认知拐杖”依赖，专业能力逐步退化。	人为衍生风险—过度依赖风险
3	引导失责风险	学生缺少风险识别与规范使用指引，易出现学术不端、数据泄露、过度依赖等问题，教师未尽育人引导职责。	人为衍生风险—违规滥用风险、过度依赖风险

3. 规范要求[1]

（5）教师应主动提升人工智能素养，增强对人工智能可靠性（输出错误、幻觉等）、公平性（算法偏见等）、透明性（算法黑箱等）、安全性（数据泄露等）等各类风险的了解，提高防范意识和能力。应引导学生严格遵守学术规范，在科研中如实声明人工智能的使用情况，防止将人工智能生成的内容视为核心创新成果，维护学术诚信，并以身作则。

（6）教师使用人工智能辅助教育教学时，应始终发挥育人主导作用，引导学生了解人工智能的各类风险，培养其识别、防范风险的能力，确保学生合理、正确、批判性地使用人工智能辅助学习，防止学生滥用或过度依赖。以下提供恰当使用和不当使用示例，供

[1]本层面主要参考《中华人民共和国未成年人保护法》《上海交通大学生成式人工智能教师使用指南》《日本初等中等教育阶段使用生成式人工智能的暂行指南》《日本初等中等教育阶段生成式人工智能利用指南（2.0）》《中小学生成式人工智能使用指南（2025 年版）》、联合国教科文组织《教师人工智能能力框架》、美国《K-12 课堂 AI 使用指南》《康奈尔委员会报告：生成性人工智能在教育和教学中的应用》等文件。

参考：

①恰当使用示例：向学生展示人工智能生成的看似合理但实际有误的内容，分析人工智能在特定领域的回答局限性（如时效性、知识边界、逻辑推理缺陷、偏见等），引导学生学会质疑和验证人工智能生成内容的正确性；将不同人工智能应用对同一问题生成的不同版本回答（包含不同偏见倾向的版本）匿名展示给学生，引导学生讨论其中可能存在的种族、性别或地域偏见，并追溯可能导致偏见的数据源环节（如训练数据的不均衡）；利用人工智能设计更多需要研究、论证或创造的复杂任务，提升学生的问题解决和创新思维能力；在确保隐私和数据安全、不替代自身主动思考的前提下，利用生成式人工智能校对、润色教案、论文等文本，提高工作效率；在学前或小学低年级教学与活动中，借助人工智能生成差异化的实践探究类作业素材（如分层级的手工示意图、自然观察任务卡），适配不同发展水平儿童的学习节奏。

②不当使用示例：在未充分了解生成式人工智能性质和特点的情况下使用其辅助教育教学；在未对未成年学生开展任何人工智能伦理安全教育的情况下，放任学生使用人工智能；在学前或小学低年级教学与活动中，直接用人工智能生成的互动内容替代教师的现场引导；将人工智能应用作为替代性教学主体，在未加干预和监管的前提下用其代替教师对学生进行人文关怀；直接采纳人工智能生成的、缺乏可解释性与人工复核的评价结果作为评价学生的依据；放任学生在定期评估、测试等场景中不加限制地使用人工智能；在安全防护手段不足、数据封闭性不够的情况下，将涉及师生隐私、敏感信息上传到人工智能工具中进行处理分析。

（7）教师应根据学生年龄和认知水平，以适当方式向学生说明在教学的哪些环节和内容上使用了何种人工智能，并阐明其已知的局限性和潜在的伦理风险。

（8）在日常作业、课后任务等场景中，教师应提前声明学生可以使用和禁止使用人工智能的范围和情况，深入阐释人工智能不可替代自身思考和练习的原因，明确使用人工智能代替自身努力完成作业的不良后果。

（9）鼓励教师布置更多过程性、实践性、个性化和提升高阶思维能力、培养综合素养的任务，不得给学前及小学低年级学生布置需要学生独立操作生成式人工智能工具完成的作业，避免学生滥用或过度使用人工智能。

4.参考实例

参考实例 3（教师层面）

实例名称	罗马尼亚中小学“AI 思维伙伴”文学批判教学实践[1]
实践做法	罗马尼亚一名六年级语文教师采用“AI 作为思维伙伴”教学策略，针对戏剧短篇教学，引导学生使用 AI 生成对应场景的人物图像，随后让学生扮演“文学侦探”，逐一对标原文细节核验 AI 输出，重点识别图像中的时代错位、外文标识、历史细节失真等“幻觉”问题，并将偏差点整理为课堂讨论素材。教师全程主导教学节奏，将 AI 的局限性转化为深度文本细读的切入点，而非直接使用 AI 生成内容授课。该案例被欧洲学校教育平台收录为负责任 AI 教学的典型范式。
借鉴启示	1.教师可将 AI 的局限性转化为培养学生批判性思维的教学资源，既合理使用技术，又避免学生盲目采信 AI 内容。 2.教师的价值体现在设计学习任务，而非依赖 AI 替代教学环节，这种模式能有效守住育人主体地位，同时提升学生的高阶思维能力。

参考实例 4（教师层面）

实例名称	苏州市金阊实验中学英语项目化学习智能体教学实践[2]
实践做法	苏州市金阊实验中学校开发个性化智能体赋能项目，恪守“以人为本、智能向善”原则，严守应用边界与规范约束，谨慎地引入智能学习支持。学校将 AI 智能体作为学生的“数字学伴”和教师的“教学助理”，赋能项目学习的“创意激发—创作辅助—评价反馈—成果美化”全流程。教师在准备阶段、实施阶段、反思阶段，通过筛选建立知识库、设计智能体的人设和回复逻辑，同时发挥智能体的赋能作用，落实人机协同中教师的主导规约策略。在教师的主导监督下，学校将智能体“安全、可靠、公平、可控”地应用于英语诗歌创作项目化学习中。
借鉴启示	1.学校可统一制定本学科的 AI 使用规约，明确“可用场景、禁用场景、核验要求”，避免教师个体使用的随意性。 2.将“AI 内容批判性辨析”融入日常教学，既是防范风险的手段，也是提升学生高阶思维的教学方法。

（三）学生和未成年人层面

1.层面定位

学生尤其是未成年人是教育 AI 治理的重点保护对象，也是 AI 素养培育的核心群体。本层面聚焦学生尤其是未成年人身心发展规律与认知特点，从素养培育、分级使用、权益保障等维度明确伦

[1]European School Education Platform. Learning activity ideas for using AI with learners[EB/OL].(2026-05-20)[2026-07-15].https://school-education.ec.europa.eu/en/teach/lesson-plans/learning-activity-ideas-using-ai-learners.

[2]苏州教育.苏州 3 个项目入选人工智能教育国家级优秀案例[EB/OL].(2026-05-08)[2026-07-15].https://mp.weixin.qq.com/s?__biz=MjM5NTgzNjQ0MQ==&mid=2651318559&idx=1&sn=608f2fc3c20be01bd27e6f0a2f0583e0&chksm=bcdbd2ad0319555f308ae19657e19a858456dbefc8ee4493c31dd0e2e30e0456c5f8467f6e73&scene=27.

理安全要求，兼顾技术赋能与特殊保护，构筑学生和未成年人 AI 应用的安全底线与成长空间。

2.核心风险

学生和未成年人层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	认知能力退化风险	基础认知建构过程被替代，独立思考、逻辑推理与创新创造能力系统性退化，形成学习惰性与技术依赖。	人为衍生风险—过度依赖风险
2	行为失范风险	主动突破学术诚信底线，破坏评价公平；因隐私保护意识薄弱，导致个人身份、家庭、健康等信息泄露。	人为衍生风险—违规滥用风险、误用失当风险
3	权益受侵害风险	未成年人辨别能力不足，技术缺陷与不良内容会被放大吸收，长期将固化错误认知，损害身心健康与价值观养成。	技术内生风险—可靠性风险、安全性风险

3.规范要求[1]

（10）应采取有效措施提升学生的人工智能素养，增强其科研诚信、学术规范的意识和能力，培养学生成为具有批判性思维、践行以人为本理念、秉承包容开放心态的人工智能使用者和创新引领者。

（11）鼓励学生（未成年人需同时满足未成年人使用条件）探索符合自身学习需求、生活实践和个人发展愿望的人工智能应用，了解、熟悉、掌握人工智能基本知识和伦理风险，恰当地、批判性地运用人工智能产品工具，持续提升自身人工智能素养。以下提供恰当使用和不当使用示例，供参考：

①恰当使用示例：在使用人工智能辅助写作或研究时，主动向

[1]本层面主要参考《中华人民共和国未成年人保护法》《中华人民共和国个人信息保护法》、教育部《中小学生成式人工智能使用指南（2025 年版）》、联合国教科文组织《生成式人工智能教育与研究应用指南》、联合国教科文组织《学生人工智能能力框架》、哈佛大学教育研究生院《学生自主项目中的 GenAI：建议和启示》等文件。

教师或团队成员披露使用人工智能的具体环节与所占比重，并依据相关规范进行标注；在使用人工智能辅助学习时，优先选用专为教育设计、教师推荐或学校认证的人工智能产品工具；在解决问题、完成任务时，首先依靠自身知识和能力进行分析判断和逻辑推理；在使用人工智能进行头脑风暴或论文修订时，保存完整的对话日志，以备导师或评审委员会核查；为有特殊需求的学生提供人工智能产品工具，帮助其突破生理或文化障碍，获得更多学习机会。

②不当使用示例：在作业中简单复制、照搬人工智能生成的内容；未经允许使用人工智能参加考试与测验；在作文或音乐、美术等需原创性的学习任务中滥用人工智能或长期依赖其替代思考；在课题研究时未经系统化、体系化深度思考，直接使用人工智能生成研究内容，且未对生成内容的真实性和准确性进行核验；在未征得相关人员明确许可、未进行数据脱敏处理的情况下，将包含个人身份、家庭信息、健康状况等敏感数据输入公共人工智能应用。

（12）未成年人应在教师或家长的许可和引导下，根据所在学段适度使用人工智能，避免其替代基础认知建构、独立思考与亲身实践过程。0—3 岁婴幼儿不建议直接接触人工智能工具、玩具及人工智能生成内容；3—6 岁儿童可在家长陪伴下在音乐、美术等领域进行有限体验，其余场景不建议使用，不得使用拟人化情感陪伴类应用，防止形成依赖；小学阶段学生不建议独自使用，可在教师使用人工智能辅助教学时适当使用，低年级以感知与体验为主，高年级侧重理解；初中阶段学生可在教师或家长陪伴下适度探索使用；高中阶段学生可在教师许可下结合技术原理利用人工智能开展探究性学习。应特别关注未成年人使用人工智能时的伦理安

全引导，加强用眼卫生与使用时长管理，帮助其理解技术的局限及其对社会的影响。

4.参考实例

参考实例 5（学生和未成年人层面）

实例名称	美国安娜堡公立学区学生主导制定《AI 权利法案》治理实践[1]
实践做法	美国密歇根州安娜堡公立学区由高中生代表全程主导制定学区级《AI 权利法案》，学生牵头开展师生意见征集、条款研讨与内容修订，教师仅提供专业方向指导。法案围绕学习主体性守护、学术诚信规范、个人隐私保护、算法评价公平、异议申诉救济五大核心维度划定规则边界，明确课堂学习、作业完成、科研项目等场景的 AI 使用标准，要求 AI 辅助生成内容须标注来源、学生学业数据不得超范围采集、AI 参与的评价须保留人工最终复核权。学生代表还将法案草案提交至密歇根州议会进行公开展示，向州级教育管理者输出学生视角的治理方案，成为美国 K12 阶段学生参与 AI 治理的代表性实践。
借鉴启示	1.将学生从规则的被动约束对象转变为治理的主动参与主体，引导学生全程参与 AI 使用规则制定，能显著提升对规则的认同感与自觉遵守程度，降低校园 AI 治理的落地成本。 2.打通学生治理意见的向上传导渠道，推动学生视角的方案融入校、区乃至更高级别的政策制定，让 AI 治理规则更贴合真实学习场景，保持动态适配性。

（四）学校层面

1.层面定位

学校是教育 AI 治理的落地执行主体，承担校内 AI 应用全流程管理的直接责任。本层面从制度建设、产品准入、数据安全、育人导向等维度明确学校的伦理治理要求，推动治理要求融入办学、教学、管理全环节。

[1]Ann Arbor Public Schools.AAPS students share draft AI Bill of Rights at Michigan State Capitol[EB/OL].(2025-12-05)[2026-07-15].https://news.a2schools.org/aaps-students-share-draft-ai-bill-of-rights-at-michigan-state-capitol/.

2.核心风险

学校层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	准入管控失当风险	存在可靠性、公平性、数据安全等缺陷的产品未经筛查投入使用，将技术内生风险直接传导至师生群体。	技术内生风险（全类别）、宏观系统性风险—治理体系滞后风险
2	校内治理缺位风险	校内治理规则缺位，无法对AI使用行为形成有效约束，出现问题后难以追溯问责，治理处于无序状态。	宏观系统性风险—治理体系滞后风险
3	育人导向偏差风险	办学导向偏离育人本质，工具理性主导教学过程，加剧师生技术依赖，破坏教育原生生态。	宏观系统性风险—教育生态失衡风险、人为衍生风险—过度依赖风险

3. 规范要求[1]

（13）应根据学生发展阶段和实际应用场景合理使用人工智能，充分发挥其在教育教学中对核心素养、创新思维、高阶思维能力培养等方面的辅助提升作用，确保应用有利于学生独立思考、批判性思维和创新创造能力发展，防止过度依赖导致主体能力退化。

（14）应制定并适时调整本校人工智能应用指南和相应管理制度，明确学生使用人工智能的场景和边界，对其在论文、作业等成果中的使用加以限定或约束，避免“一刀切”的简单化管理。应及时干预学生沉迷、过度依赖或接触不良人工智能应用的情况，主动告知监护人，推动校家社协同，共同引导学生正确使用人工智能。

（15）应明确人工智能应用的进校准入机制，加强对产品备案情况、标识情况和生成内容的核实，确保其合法合规。应根据技术特点和本校师生需求，打造环境、构建平台、开发资源、普及知识、

[1]本层面主要参考《中华人民共和国个人信息保护法》《中华人民共和国未成年人保护法》《生成式人工智能教师使用指南》（上海交通大学教学发展中心发布）、《生成式人工智能学生使用指南》（华东师范大学传播学院和北京师范大学新闻传播学院联合发布）《中国智慧教育白皮书（2025年版）》等文件。

开设课程，推进人工智能与教学、教研、科研、管理各环节有机结合、有效融合。

（16）应不断提升数据防泄露、防篡改、防滥用等网络安全防护能力，落实个人信息处理最小必要原则，建立学生和家长的数据知情同意机制。在处理涉及个人隐私等敏感数据时，鼓励采取本地化部署、数据脱敏等措施保障数据安全。

（17）应将人工智能伦理安全教育纳入师生培训和课程体系，提升全体师生的伦理安全风险防范意识和判断能力。

4.参考实例

参考实例 6（学校层面）

实例名称	美国佛罗里达大学校级 AI 伦理治理与统一使用规范实践[1]
实践做法	美国佛罗里达大学面向全校教学场景发布系统化的生成式 AI 应用指南与常见问题规范，建立覆盖原则指引、场景规则、准入管控、素养支撑的全链条教学 AI 治理体系。校方明确若干项 AI 应用核心指导原则，涵盖价值向善、透明可解释、安全可靠、包容可及、学术诚信、数据隐私、偏见防控、伦理合规、持续优化九大维度，作为全校 AI 教学应用的统一价值遵循。 在落地执行层面，校方将课程 AI 使用权限划分为“全面允许、有限使用、禁止使用”三类标准，要求教师在教学大纲与作业要求中明确标注 AI 使用规则，所有 AI 辅助生成的学生成果必须规范标注来源；明确所有入校 AI 工具须通过学校综合风险评估流程准入，严格落实学生数据分类保护要求，敏感信息须脱敏处理后方可使用。同时配套 AI 提示词手册、分层教学案例、AI 素养练习资源，帮助师生掌握工具使用、输出评估、伦理判断的综合能力。
借鉴启示	1.学校可设立跨部门的 AI 伦理治理机构，统筹教学、管理、后勤等多场景的 AI 应用，避免单一部门权责不足、标准不一的问题。 2.入校 AI 产品前置风险评估+分级管理的模式，既能守住安全底线，又能避免“一刀切”禁用阻碍技术赋能教学。 3.配套工具化、场景化的师生 AI 素养支撑资源，可降低治理落地的执行成本，提升师生合规使用的自觉性与实操能力。

[1]University of Florida.AI FAQs[EB/OL].[n.d.][2026-07-15].https://ai.ufl.edu/teaching-with-ai/faqs/.

参考实例 7（学校层面）

实例名称	同济大学 MTO 大模型校园安全治理体系[1]
实践做法	同济大学构建“管理机制—技术能力—运营体系”三位一体的 MTO 大模型安全治理框架，建立“学校—学院—科研团队”三级组织架构，覆盖教学、科研、管理全环节；所有入校 AI 项目必须先通过伦理审查与合规备案，明确不同场景的使用边界，配套数据分类分级防护、内容安全审核等技术能力，是国内高校系统化 AI 治理的典型实践。
借鉴启示	1.学校可建立跨部门治理组织架构形成治理合力，避免单一部门权责不足问题。 2.可参照“备案制+分级管理”模式，对不同风险等级的 AI 应用实施差异化管控，避免一刀切。

（五）家庭层面

1.层面定位

家庭是教育 AI 治理的延伸阵地，也是未成年人保护的重要防线。本层面聚焦家庭场景的监护责任与素养协同，明确家长在未成年人 AI 使用中的引导、监督职责，推动家校协同，形成家校治理合力。

[1]马宁.同济大学大模型安全治理案例获评全国教育领域人工智能安全向善优秀案例[EB/OL].(2026-03-24)[2026-07-15].https://news.tongji.edu.cn/info/1003/93895.htm.

2.核心风险

家庭层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	监护引导失位风险	校外场景监管缺失，未成年人易过度沉迷技术、接触不良内容，加剧认知能力退化与权益受侵害风险	人为衍生风险—过度依赖风险、误用失当风险
2	隐私防护疏漏风险	因监护者认知不足导致未成年人数据被违规采集与滥用，扩大隐私泄露与商业化利用的风险	技术内生风险—安全性风险、人为衍生风险—误用失当风险
3	认知观念偏差风险	家庭教育观念出现偏差，弱化对孩子独立思考、实践能力的培养，变相助长技术依赖，违背育人规律	人为衍生风险—误用失当风险、宏观系统性风险—教育生态失衡风险

3. 规范要求[1]

（18）家长应履行对未成年人使用人工智能的法定监护责任，合理管控未成年子女使用人工智能的时长与应用场景；主动学习人工智能、网络安全与科技伦理相关知识，配合学校开展人工智能应用安全教育，引导子女理性看待人工智能的不足与局限，严防网络沉迷与不良信息侵害。

（19）家长应严格遵守个人信息保护、网络安全相关法律法规，不随意在人工智能应用中填报未成年人个人信息，主动核实人工智能应用的隐私政策，规避其带来的隐私侵犯与数据失控风险。

[1]本层面主要参考《中华人民共和国网络安全法（2025 年修订）》《中华人民共和国个人信息保护法》《中华人民共和国未成年人保护法（2024 年修正）》《教师生成式人工智能应用指引》。

4.参考实例

参考实例 8（家庭层面）

实例名称	美国阅读障碍儿童家庭定制化 AI 辅助学习实践[1]
实践做法	美国一位任职于路透社的母亲为 11 岁患有阅读障碍的儿子定制专属 AI 学习辅助工具。在确认模型的数据隐私设置后，她向定制化 GPT 模型传入了孩子的成绩单、神经心理学评估报告、针对阅读障碍的个别化教育计划（IEP），以及孩子的兴趣偏好，将 AI 设定为具备阅读障碍教学专长的特殊教育教师角色，为孩子生成适配其认知特点的游戏化读写练习。每次使用后，母亲都会根据孩子的学习状态调整 AI 的难度设置与教学策略，持续优化定制化辅导效果。后续该母亲基于这一家庭实践，开发出面向阅读障碍儿童的 AI 辅导平台，增设了学习挫败感识别、健康休息提醒、家长监护后台、自伤风险预警等安全防护机制，在个性化学习与未成年人保护之间建立平衡。
借鉴启示	1.家庭场景使用教育 AI 可基于孩子的个体特质定制适配方案，弥补公共教育资源在特殊教育个性化供给上的不足，助力特殊需求儿童公平获取学习支持。 2.家庭使用涉及未成年人敏感健康、学业数据的 AI 工具时，应优先确认平台隐私保护机制，通过家长主导数据上传、设置监护后台等方式管控数据风险。

（六）社会层面

1.层面定位

社会层面是教育 AI 治理的生态支撑，涵盖产业供给、科研支撑、公共监督等多个维度。本层面从公平包容、科研赋能、产业责任等角度，明确社会各方在教育 AI 伦理治理中的责任，构建多元参与、协同共治的治理生态。

[1]BÉCHARD D E. How one mom built an AI tutor for her dyslexic son[EB/OL].Scientific American,(2025-10-31)[2026-06-22].https://www.scientificamerican.com/article/how-one-mom-used-vibe-coding-to-build-an-ai-tutor-for-her-dyslexic-son/.

2.核心风险

社会层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	供给合规失范风险	商业逻辑主导产品研发，不良产品无序扩张，出现“劣币驱逐良币”的逆向淘汰，消解教育的公共属性。	宏观系统性风险—教育生态失衡风险、技术内生风险（全类别）
2	公平包容不足风险	技术应用加剧既有教育差距，传统数字鸿沟升级为智能鸿沟与认知鸿沟，固化教育发展的不均衡。	宏观系统性风险—教育公平结构性冲击风险
3	研究支撑滞后风险	风险研判与治理优化缺少科学依据，难以应对技术快速迭代带来的新型挑战，治理前瞻性不足。	宏观系统性风险—治理体系滞后风险

3.规范要求[1]

（20）应将教育公平贯穿人工智能设计、研发、部署、评价全流程，避免产生数字鸿沟、结构性排斥和歧视，提升人工智能教育服务的广泛性和包容性；面向视听、语言障碍者、老年人等特殊群体建设无障碍学习环境，提供数字无障碍解决方案，确保所有学习者尤其是弱势群体能够平等参与并受益于数字教育。

（21）教育相关科研单位、社会团体和企业等机构应积极开展人工智能应用于教育教学的基础性、机理性、应用性研究，对人工智能技术发展前沿动态及其对教育教学的影响开展长期跟踪研究，总结人工智能应用场景和案例，加强对规则冲突、社会风险、伦理挑战等的研判，为人工智能教育制度体系建设提供科学实证依据。

[1]本层面主要参考《教育信息化 2.0 行动计划》《中华人民共和国数据安全法》等文件。

4.参考实例

参考实例 9（社会层面）

实例名称	华中师范大学学前特殊教育人机协同 AI 普惠实践[1]
实践做法	华中师范大学人工智能教育学部深耕智能教育与特殊教育融合领域，针对学前特殊儿童打造“感知—决策—干预—评价”全流程人机协同个性化学习体系：以多模态智能感知技术构建“机器感知+教师理解”的个体认知机制，精准刻画儿童发展画像，动态生成“一生一案”教育方案；以人机共创模式开发缺陷补偿式学习资源，适配不同障碍类型、不同能力层级的儿童需求；建立“机器分析+教师评估”的双维度评价机制，实时优化教学干预。该模式已在全国多省市落地，累计服务 20 万余名儿童，学习参与度与康复效果显著优于传统模式，形成了可复制推广的学前特殊教育智能化应用范式。
借鉴启示	1.教育 AI 研发必须兼顾特殊群体需求，将无障碍、包容性作为产品伦理评估的重要维度，避免技术加剧教育不公。 2.科研机构应长期跟踪技术应用的社会影响，为治理政策优化提供实证依据，为行政部门提供专业支撑。

（七）行政层面

1.层面定位

行政层面是教育 AI 治理的顶层统筹主体，承担规则制定、机制建设、监管引导的核心职责。本层面立足北京市级治理视角，从技术供给、审查机制、分类管理、监测预警、协同治理等维度，明确行政层面的治理路径与伦理要求，为全市教育 AI 治理提供参考。

2.核心风险

[1]陈靓影，刘乐元，游丽.华中师范大学人工智能教育学部：构建人机协同个性化学习体系推动学前特殊教育发展[N/OL].光明日报，2026-05-23(04)[2026-07-15].https://epaper.gmw.cn/gmrb/html/content/202605/23/content_14704.html.

行政层面核心风险分析

序号	风险表现	风险内涵	风险归类
1	制度迭代滞后风险	技术应用与制度约束形成时间差，新型风险无法得到及时有效应对，治理整体呈现被动滞后特征。	宏观系统性风险—治理体系滞后风险
2	协同治理不足风险	教育、网信、市场监管等部门协同不足，出现监管空隙与重复监管并存的问题，整体治理效能不足。	宏观系统性风险—治理体系滞后风险
3	技术治理能力不足风险	监管专业能力与技术发展不匹配，无法对隐蔽性、技术性强的风险开展精准管控，治理有效性不足。	宏观系统性风险—治理体系滞后风险、技术内生风险（全类别）

3. 规范要求[1]

（22）组织高校、人工智能和教育相关科研单位、人工智能领军企业组成联合研发体，分教育阶段研发人工智能教育大模型，强化价值对齐、安全伦理等能力，确保模型输出与人类共同价值及教育育人目标保持一致，防范输出错误、算法偏见、算法黑箱、数据泄露等风险。模型训练数据应严格筛选，过滤虚假、偏见、失效、错误数据，确保真实性、准确性、客观性、多样性和来源合规性。鼓励研发和采用具备透明决策逻辑的模型和可解释算法。推动研发成果向智能学伴、教师智能助手、人工智能辅助学习助手等教育终端产品转化，支持优质教育大模型与智能产品从试点示范走向规模化应用，打造智慧课堂示范标杆。

（23）建立面向人工智能教育应用的多层次伦理审查与准入机制，建立用户申诉、纠错与风险处置渠道，明确各方责任边界，使人工智能教育应用在人类福祉、公平公正、可控可信、透明可解

[1]本层面主要参考《中华人民共和国个人信息保护法》《“人工智能+教育”行动计划》《人工智能安全治理框架》2.0 版、《人工智能科技伦理审查与服务办法（试行）》《人工智能生成合成内容标识办法》《互联网信息服务深度合成管理规定》、欧盟《人工智能法案》《中小学生成式人工智能使用指南》《中华人民共和国教育法》《教育信息化 2.0 行动计划》《中国智慧教育白皮书（2025 年版）》等文件。

释、责任可追溯等方面符合伦理安全要求。从事教育领域人工智能研发的高等学校、科研机构、教育科技企业应设立相应机制，履行伦理审查管理主体责任；各级各类学校应加强对拟进入学校的人工智能应用的伦理审查，确保其符合相应的伦理安全要求。

（24）依据教育对象的心智成熟度、育人目标及场景特殊性，从“禁止准入”“有限使用”“鼓励使用”等方面对人工智能教育应用进行分级分类管理，并根据技术发展、教育实践反馈及风险研判，定期评估和更新分级分类情况。遇重大技术突破或突发性严重伦理安全风险时，及时调整分级分类，确保分类治理始终与教育实践同频共振。

（25）建设人工智能教育伦理安全风险监测与评估平台，建立相应的测评标准，对模型算法、数据资源、输出质量、用户投诉反馈与异常行为日志等进行常态化监测和综合评估。对人为恶意利用人工智能进行伪造诈骗、学术造假等行为，以及人工智能技术本身可能引起的各类风险，开展有效监测和分级预警，提供评估支撑。

（26）加快完善人工智能教育相关法律法规、行政规范、指南指引，鼓励教育机构、企业、科研单位、社会团体聚焦教育行业人工智能应用、大模型评测、数据安全等研制相应的标准规范。

（27）搭建跨部门、跨地域、跨平台的风险信息共享平台，跟踪分析并及时共享人工智能教育应用的安全漏洞、缺陷、风险威胁、安全事件信息，协同防范应对教育领域的安全风险。

（28）充分考量区域发展平衡性、地方特色多样性、教育资源均衡性、人工智能素养差异性等客观要素，有序推进区域人工智能

教育推广应用和伦理安全治理工作。

4.参考实例

参考实例 10（行政层面）

实例名称	美国伊利诺伊州教师评价 AI 公平监管立法实践[1]
实践做法	美国伊利诺伊州于 2026 年签署生效《参议院 2909 号法案》（SB2909），是美国首个针对教育场景 AI 决策公平性的专项州级立法，核心聚焦公立学区教师绩效评价中的算法歧视风险。法案核心规则包括：一是明确划定 AI 使用边界，禁止学区将 AI 生成内容作为教师绩效评价的唯一依据，要求所有 AI 辅助评价结论必须由具备资质的教育管理者人工复核并拥有最终决策权，防范算法偏见直接决定教育从业者职业权益；二是确立 AI 使用披露义务，学区若使用 AI 工具辅助评价工作，必须向参评教师明确告知 AI 工具的使用范围、核心逻辑与权重占比。
借鉴启示	1.可将算法公平作为教育 AI 治理的重要维度，针对教师评价、学生分层、学业测评等高利害教育场景明确 AI 使用边界，防范算法偏见加剧教育不公。 2.明确高利害场景中 AI 的辅助定位，为受 AI 决策影响的师生提供有效的权益救济路径。 3.以专项立法形式明确教育场景 AI 应用的公平性约束，提升治理刚性，倒逼 AI 产品供给方与学校落地主体强化算法公平性管控。

参考实例 11（行政层面）

实例名称	英国教育部生成式 AI 教育官方治理体系[2]
实践做法	英国教育部 2025 年发布全国性生成式 AI 教育政策文件，确立“安全优先、人类专业判断不可替代”的核心原则，明确 AI 技术不能替代教师的专业判断与师生互动。文件将生成式 AI 安全纳入《保持儿童教育安全》法定指导的适用范围，要求学校引入 AI 工具前必须开展全面的安全与隐私风险评估；优先推动教师端应用释放教学减负价值，学生端应用需设置严格防护措施，确保校园儿童保护政策全面覆盖生成式 AI 带来的新型风险。
借鉴启示	1.教育行政部门可将 AI 安全纳入法定校园安全监管体系，明确刚性监管要求，提升治理的约束力与执行力。 2.坚持人类监督不可替代的底线原则，从顶层避免技术过度介入教育过程、消解育人主体地位。

参考实例 12（行政层面）

实例名称	新加坡教育部分学段 AI 分级使用治理体系[3]

[1]Makebox.AI. Illinois Bans AI as Sole Basis for Teacher Evaluations, Mandates Human Judgment [EB/OL].(2026-07-14)[2026-07-22].https://makebox.ai/news/illinois-bans-ai-teacher-evaluations.

[2]UK Department for Education. Generative artificial intelligence (AI) in education [EB/OL].(2025-08-12)[2026-06-23].https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education.

[3]Ministry of Education, Singapore. Artificial intelligence in education [EB/OL]. (2026-06-02)[2026-06-23]. https://www.moe.gov.sg/education-in-sg/educational-technology-journey/edtech-masterplan/artificial-intelligence-in-education..

实践做法	新加坡教育部依托国家学生学习平台构建分学段 AI 管理体系，低年级阶段不开放生成式 AI 工具，聚焦基础读写算能力培养；小学中高年级仅在教师全程监督下使用官方定制 AI 工具，AI 仅通过启发式提问引导思考、不直接输出答案；中学以上逐步开放更多功能，同步配套数字素养课程，确保 AI 应用与学生心智成熟度匹配。
借鉴启示	1.国家级层面的分学段管控逻辑可作为北京市制定分级分类管理政策的参考标杆。 2.“技术护栏+素养培育”双轨并行，既通过技术手段限制风险，又通过教育提升学生自主管理能力。

参考实例 13（行政层面）

实例名称	上海市虹口区生成式 AI 入校全流程治理机制[1]
实践做法	虹口区作为上海市人工智能教育试验区，先后发布《虹口区学校生成式人工智能应用暂行管理办法》《HEADS 智能体发布与移除管理办法》，建立“准入审核—过程监督—责任到人”三层管控机制，明确智能体应用标准、责任主体与监管边界，为 AI 教育应用划定运行规则。平台层面，建成全国首个区域级教育智能体导航系统（HEADS）作为统一管控载体，所有入校 AI 智能体须先履行备案程序，经伦理与安全审核通过后方可上线；配套建立动态评估与移除退出机制，覆盖智能体准入、运行、退出全生命周期，实现全流程闭环管控。生态层面，依托 HEADS 平台打通区域、学校、教师联动链路，形成“以用促建、共研迭代”的治理模式，支持一线教师基于真实教学痛点共创智能体，优质应用通过平台实现区域共享与持续优化，在规范框架内兼容技术创新与教学实践需求。
借鉴启示	1.区级教育行政部门可搭建区域统一的 AI 智能体管理平台，对入校 AI 产品实行集中备案、统一审核，破解各校各自为政、标准不一的监管碎片化问题，提升全域 AI 教育应用治理的系统性与规范性。 2.准入管控、动态退出、教学驱动、共创共享，使 AI 技术更加符合教学场景个性化需求。

参考实例 14（行政层面）

实例名称	深圳市福田区教育 AI 伦理审查与备案准入制度[2]
实践做法	深圳市福田区构建“制度定标+技术把关+能力支撑”的教育 AI 入校全流程治理体系。制度层面，2025 年 3 月在全省率先发布《全域推进教育人工智能应用的二十五条举措》，形成系统化、专业化的 AI 教育应用行动纲领，为 AI 产品入校准入、教学应用规范、伦理边界把控提供总纲性制度依据，相关实践经验获教育部基础教育司转发推广。准入安全层面，采用私有化部署大模型的方式搭建区域本地知识库，实现师生教学数据本地化留存，从技术底层规避师生数据流入公共大模型训练池的风险，筑牢 AI 入校的数据安全前置门槛。过程管控层面，组建“福田 AI 先锋队”，研制《教师 AI 应用手册》《AI 教学工具百宝箱》等应用指引文件，明确 AI 工具在备课、课堂管理、作业批改等全教学场景的使用规范与伦理边界；构建“课程驱动、平台驱动、问题驱动”三位一体的 AI 教师培训体系，将数智素养纳入教师专业发展

[1]文汇报.以智能体系统赋能教育变革：虹口区“AI+教育”生态的实践图景[EB/OL].(2026-01-20)[2026-07-16].https://m.jfdaily.com/journal/getMobileArticle.htm?id=488203.

[2]深圳市福田区教育局(教育督导室).受教育部关注转发！福田区引领 AI 教育变革潮流[EB/OL].(2025-04-07)[2026-07-16].http://www.szft.gov.cn/ftqjyj/gkmlpt/content/12/12109/mpost_12109107.html#8450.

	核心指标，提升教师对 AI 应用的合规研判与规范使用能力，形成“制度准入—技术兜底—过程规范”的治理闭环。
借鉴启示	1.区级教育行政部门可出台全域性 AI 教育专项政策作为总遵循，同时将数据安全技术标准作为 AI 产品入校的前置准入条件，从源头把控 AI 入校的合规性与安全性，过滤不合规应用。 2.明确 AI 使用的伦理与操作边界，同步强化教师 AI 素养培育，实现 AI 应用的常态化规范管控，避免重准入、轻过程的监管缺位。

《教育领域人工智能应用伦理安全专题导引》编写委员会

主编单位：北京教育科学研究院

编委会组长：冯洪荣 梁家峰

编委会副组长：杨德军

编委会核心成员：霍丽娟 郭秀晶 吕良燕 詹伟华 高兵 史枫 唐亮 孙颖 罗海风 古燕莹

本册编写人员：唐亮 张林 陈必值 高保琴 李艳霞 于志涛 刘寅