AI技术赋能教育安全风险评估与应对策略

——评《AI大模型安全观:通用人工智能的应用场景、安全挑战与未来影响》

在数字经济崛起与人工智能(Artificial Intelligence,AI)技术深度渗透的时代背景下,教育作为人才培养、知识传递的核心领域,正全面迎来AI技术的赋能重构。从智能教学系统依据学情动态推送个性化课程内容,到AI助教实时解答学生课堂疑问、辅助教师批改日常作业;从基于大数据分析的教育管理决策优化,到计算机视觉支撑的课堂行为分析与校园安全监控,AI技术正以高效、精准、便捷的优势,贯穿"教学实施-学习支持-管理服务-质量评价"教育全流程。

笔者在开展福建省教育科学规划2024年教育考试招生重点专项课题(FJJKKS24-39)研究过程中,认真阅读了《AI大模型安全观:通用人工智能的应用场景、安全挑战与未来影响》一书,书中以数字技术引领、AI主导的全新发展范式为背景展开论述。全书共8章,适合教育管理人员、AI企业从业人员、网络安全监管人员及相关专业研究者参考,旨在帮助读者客观、全面认识大模型技术,实现与AI的和谐共处。其中,第1章和第2章简要介绍了ChatGPT与自然语言大模型的基本概念、GPT-4的核心技术特点及AI未来发展趋势,阐述了数字化时代的基础安全问题;第3—6章深入分析了AI技术在内容安全、网络安全、隐私安全、版权合规及伦理道德等方面的新挑战与新风险,并从AI技术监管角度给出了策略建议;第7章从反向视角探讨如何利用ChatGPT等AI工具提升安全防护能力,强调技术价值取决于使用者,凸显"技术中立性";第8章从行业维度分析AI技术对经济社会的影响,解读其局限性、国内外大模型发展差异,并提及我国相关政策、AI产业规模及发展地位。

书名:AI大模型安全观:通用人工智能的应用场景、安全挑战与未来影响
作者:秘蓉新,翟尤
出版社:人民邮电出版社
ISBN:9787115622761
出版时间:2023年7月
定价:69.8元

作者指出,AI技术的快速应用必然伴随着新的安全风险与挑战。在教育领域,AI技术的赋能并非无懈可击:学生利用ChatGPT、豆包等AI工具代写作业、伪造试验报告的现象日益普遍,AI教学系统推送的知识内容存在事实偏差,学生学习数据在AI系统存储与传输中泄漏风险凸显,甚至AI训练数据的版权问题可能波及教学资源的合规性。在此背景下,系统评估AI技术赋能教育的安全风险,构建科学、可行的应对策略,不仅是保障教育公平性、守护教育本质的必然要求,更是推动AI技术与教育深度融合、实现教育数字化转型的关键前提。

笔者认为,AI技术在教育领域的应用已覆盖"教学-学习-管理-评价"全链条,但其引发的安全风险呈现出"多维度、隐蔽性、传导性"特征,须从公平性、数据安全、技术可靠性、伦理合规4个核心维度展开系统评估。

1)公平性风险。教育的核心价值在于公平性,而AI技术的不当应用正从资源获取、过程评价、结果认定3个层面侵蚀这一底线。从资源壁垒来看,优质AI教育工具往往伴随较高使用成本,仅能被经济条件优越的学校或家庭获取,导致教育资源分配差距进一步扩大,形成"AI教育鸿沟"。从过程评价来看,AI评价系统的偏差可能导致不公平判定,导致评价结果失真。从结果认定来看,AI工具的滥用破坏了能力评估的真实性,除了考试作弊,学生利用AI代写课程论文、伪造实践报告的现象更为普遍,这类AI生成内容往往结构完整、表述规范,但缺乏真实的知识积累与能力训练,使教师难以准确判断学生的真实学习水平,最终影响升学、评优等结果的公正性。如该书中提及的"生成内容准确性问题",AI生成的学习成果不仅可能存在逻辑漏洞,更重要的是其破坏了学生自主学习的教育基本原则,使教育评价失去对真实能力的衡量价值。

2)数据安全风险。教育领域涉及大量高敏感数据,包括学生的个人基础信息、学习行为数据、心理测评数据,以及教师的教学方案、教育机构的管理决策数据等。这些数据在AI系统的"收集-存储-处理-传输"全流程中,面临泄漏、滥用、篡改等多重风险。在数据收集环节,部分AI教育产品存在"过度收集"问题,本质上违背

了数据最小必要原则,为隐私泄露埋下隐患,这与该书中“隐私安全是 AI 应用的核心底线”的观点高度契合。在数据存储环节,AI 系统的“技术节点多、接口开放”特性,增加了安全漏洞的暴露概率。在数据处理与传输环节,滥用与篡改风险尤为突出,考试数据、升学推荐数据在 AI 系统间传输时,若未采用加密技术,易被第三方拦截篡改,影响教育决策的公正性。

3)技术可靠性风险。AI 技术的“黑箱”特性与训练数据局限性,导致其在教育应用中存在显著的技术可靠性风险,直接影响教学秩序、误导学习方向,甚至引发师生对 AI 教育工具的信任危机。从系统故障来看,AI 教育系统的复杂性使其故障影响范围更广,若出现 AI 课堂互动系统崩溃可能导致许多学校课程中断等问题。从功能误判来看,AI 系统的场景适应性不足问题在教育领域尤为明显,AI 作文批改系统可能将“富有个性的表达”判定为“语法错误”,AI 答疑系统可能因训练数据陈旧,给出与现行教材不符的知识点解释。从能力边界突破来看,“对抗性攻击”成为新隐患,第三方机构可能通过技术手段篡改 AI 升学推荐算法,为特定学生获取优质教育资源。这正是《AI 大模型安全观》中“大模型技术存在被恶意利用的风险”在教育领域的具体体现,表明 AI 系统的能力边界并非不可突破,需加强防护措施。

4)伦理与合规风险。AI 技术在教育中的应用,还面临伦理边界模糊与合规性不足的双重风险。从伦理风险来看,核心问题在于技术异化教育:①教育目标功利化,部分 AI 教育工具过度关注“分数提升”,通过高强度刷题、应试技巧训练替代素质教育,忽视学生的创造力、情感能力培养;②学生自主能力弱化,过度依赖 AI 答疑、AI 作业辅助的学生,逐渐丧失独立思考与问题解决能力;③教师角色边缘化,部分学校引入 AI 教学系统后,将教师定位为“AI 辅助者”,削弱了教师在育人过程中的情感关怀与价值引导作用。从合规风险来看,主要体现在算法透明性与政策适配性不足。多数 AI 教育系统未向师生、教育监管部门公开核心逻辑,当学生对 AI 评价结果有异议、教师对教学推荐内容存疑时,难以获得合理解释;部分 AI 教育产品未符合《个人信息保护法》《未成年人网络保护条例》等法规要求,在数据使用、未成年人保护等方面存在合规漏洞,可能面临行政处罚。

针对上述安全风险,结合书中推测的“技术防护+制度监管+伦理引导”的综合防控思路,需从技术、制度、伦理、人才 4 个维度构建全流程、多层次、立体化的应对策略体系,实现风险可防、可控、可追溯,推动 AI 技术与教育的安全融合、良性互动。

1)技术防护。围绕“作弊检测、数据保护、系统稳定、算法优化”4 个核心目标,开发针对性技术方案,同时,发挥人工复核的“兜底作用”,形成“AI 预警+人工判断”的双重防控机制。在作弊检测方面,开发多模态 AI 作弊检测系统,覆盖考前-考中-考后全流程,建立作弊检测结果人工复核机制,避免误判。在数据保护方面,构建全生命周期数据安全防护体系,保障教育数据安全。在系统稳定方面,建立 AI 系统可靠性保障机制,降低故障风险。在算法优化方面,提升 AI 阅卷、组卷系统的准确性与透明度。

2)制度监管。制度监管是应对安全风险的关键,需结合教育行业特点,完善相关法律法规与行业标准,建立全链条、跨部门的监管机制,明确 AI 技术在教育中的应用边界与责任主体。在法规完善方面,需细化 AI 教育考试应用的法律条款,出台《人工智能教育应用管理细则》,明确 AI 技术在教育考试中的应用范围、数据使用规范、算法合规要求。在标准制定方面,制定涵盖技术要求、安全要求、评估要求的系列标准,构建 AI 教育技术标准体系,规范技术应用。在监管机制方面,建立跨部门协同监管平台,实现全链条监管。

3)伦理引导。伦理引导是应对伦理风险的核心,需围绕公平、尊重、育人 3 个核心伦理原则,建立 AI 教育伦理审查机制,引导 AI 技术应用回归教育本质,避免技术“异化”教育考试。明确 AI 教育应用的伦理边界,制定《AI 教育伦理准则》,需坚持公平、尊重和育人 3 大原则。建立 AI 教育伦理审查委员会,对 AI 系统进行伦理评估。加强 AI 应用伦理教育,提升考生、教师、AI 从业者的伦理意识。

4)人才培养。人才是应对安全风险的保障,需培养既懂教育领域业务,又懂 AI 技术、数据安全、法律法规的复合型人才,为 AI 教育领域安全风险防控提供智力支持。构建“高校+企业+教育机构”协同培养模式。高校可在教育技术学、计算机科学与技术、法学等专业中增设 AI 教育安全相关课程,培养具备跨学科知识的本科生、研究生;教育机构开展在职人员培训,定期组织教育机构工作人员、AI 企业技术人员参加培训,内容包括 AI 技术最新进展、安全风险防控技术、相关法律法规,提升在职人员的专业能力。

(马小芳/福州外语外贸学院教育学院/副教授)