教育领域生成式人工智能应用的伦理风险与应对举措

杨伟涛　　李晨伟

摘　要：以 ChatGPT 为代表的生成式人工智能正引发科技、人文的深刻变革。这一技术在教育领域的应用显现出巨大的潜力与价值，包括推进教学精准化变革、驱动教学场景的数智化转型等。生成式人工智能无疑是教育未来发展的重要引擎。然而，当前生成式人工智能处于方兴未艾的发展阶段，在赋能教育的实践中仍有局限性，衍生出数据安全失范、主体性消解及意识形态渗透等多重伦理风险。为防范风险并形成最大化技术效能，应筑牢数据安全防线、复归人的主体地位、巩固主流意识形态引领，通过“技术监管—人文关怀—价值引领”三维路径，推动生成式人工智能教育应用规范化，赋能教育的高质量发展。

关键词：生成式人工智能；教育赋能；伦理风险

中图分类号：B82　　**文献标识码：**A　　**文章编号：**1003-0751(2026)02-0111-08

习近平总书记在向国际人工智能与教育大会致贺信中明确指出，“中国高度重视人工智能对教育的深刻影响，积极推动人工智能和教育的深度融合，促进教育变革创新”[1]。2025 年 1 月 19 日，中共中央、国务院印发的《教育强国建设规划纲要(2024—2035 年)》进一步作出战略部署，强调要“建设学习型社会，以教育数字化开辟发展新赛道、塑造发展新优势”[2]。在全球数字科技浪潮的推动下，教育正经历颠覆性变革。以 ChatGPT、Sora、DeepSeek 等为代表的生成式人工智能(Generative Artificial Intelligence，GAI)，凭借强大的数据化处理、智能化交互、创造性生成等能力，展现出提高知识生产效率、拓展教育空间以及支持沉浸式、个性化学习的潜力，甚至在逻辑推理、创意写作、对话交互等方面表现突出。学者普遍认为它将引领教育进入人机协同的新时代。“精准教学”“智能批改”“AI 助教”“智能学伴”“智慧校园”等观点再次华丽登场，为传统教育模式注入前所未有的活力与可能。毋庸讳言，生成式人工智能的兴起推动“以人为本”教育理念走向深入实践，其应用呈现出鲜明的“育人”导向。在人机协同的新生态中，这种深度介入教与学的模式，实现了规模化教育场景中的“因材施教”，有效填补了人类在某些教育环节中的“缺陷存在”。

从当前学界相关研究动态看，数字技术在推动教育创新发展的同时，其应用中的一系列风险也日益呈现。目前，部分学者聚焦数据与算法伦理维度，揭示了 GAI 在教育落地过程中潜藏的隐私泄露、算法黑箱、过度监控等问题；另有研究从教育主客体关系的嬗变出发，批判性审视了人机共处中面临的教师权威弱化、教学情感淡化等挑战；抑或基于宏观维度，深入剖析算法可能引发的公平失衡与责任归属等困境。由此，理性审察技术应用中引发的教育伦理困境，并提出切实可行的举措，变得尤为迫切。因而要积极“探索‘人工智能+教育’应用场景新范式，推动大模型与教育教学深度融合”[3]。本研究建立起整体性的分析框架，以数据安全为基点，以主体性强化为核心关切，以价值失范为批判准绳，系统审视

收稿日期：2025-09-25

作者简介：杨伟涛，男，郑州大学哲学学院副院长、教授、博士生导师(河南郑州　450001)。李晨伟，女，郑州大学马克思主义学院博士生(河南郑州　450001)。

生成式人工智能教育应用中的伦理困境,致力于构建兼顾技术赋能与伦理规约的治理路径。

一、教育智能化的伦理范式跃迁:生成式人工智能的赋能潜力重构

生成式人工智能作为依托海量数据训练的深度神经网络模型,凭借其在知识表征、认知模拟、情景构建等方面的技术优势,正深刻融入教育生态的核心环节。与其他领域不同,GAI在教育中的赋能始终围绕育人这一根本目标展开,不仅重塑了知识传递的路径与效率,而且致力于构建适配个体认知发展的学习环境,推动教育从标准化规模供给向个性化精准培育的深刻变革,开拓了教育创新发展的崭新图景。

(一)从标准化到智适应:驱动教学精准化变革

在教育教学场景中,人人都是"特殊的个体,并且正是人的特殊性使人成为个体,成为现实的、单个的社会存在物"[4],个体之间始终存在差异性。生成式人工智能凭借强大的信息处理能力、深度理解与智能交互等功能,可依托用户倾向性特征提供"一对一"专属学习体验,"量身定制"多元化学习资源。便捷高效的智能化终端设备及海量云存储系统,更是为数字囤积提供了技术上的有力支撑,引发了数字教育领域的精准化变革。

第一,刻画学习者数字画像,构建定制化知识网络。用户需求的精准识别、全面认识是教育活动有效开展的重要前提。当前,基于通用模型的教育资源服务难以实现针对用户学情的动态调整,尚存在资源供需不同步、个性化程度不足、预测能力不佳等问题。其固有局限在于静态的知识体系难以精准适配每一位学习者独特的认知节奏、兴趣倾向与情感状态,构成了传统教育中难以兼顾的"现实鸿沟"。数字化生存是当代青年的重要生活方式,任何形式的网络行为都能以数据的形式留下足迹。依托强大的算法技术,用户留存于网络空间的多模态数据,如文本、图像、音频、视频等,都能被快速捕捉,成为系统推荐的关键。生成式人工智能作为海量数据资源的庞大集合体,并不是单向地输出提前"预设"的答案,而是通过对采集到的数据进行相关性及行为偏好分析,如网络留言、作业完成情况等,塑造出符合用户需求与学习节奏的画像模型,进而层层深入靶向聚焦,开展千人千面的个性化学习体验[5]。同时,通过对用户历史学习数据的深度挖掘与比对,该技术能够精准诊断其认知结构与知识盲区,并自动生成与之匹配的适应性教学内容。在此基础上,系统以无所不知、无时不应的智能形象,实时感知动态交互中产生的新问题,快速关联并整合相关领域的知识点,实现多源信息的智能分析与融合,全方位充当用户的智能学伴,促使因材施教成为可能。

第二,依托人机交互技术,强化学习体验感与参与感。在数字教育背景下,学习体验与参与程度已成为影响学习成效的关键变量。用户体验最早由唐纳德·A.诺曼(Donald Arthur. Norman)提出。他强调"以用户为中心的设计",重视使用过程中的全部主观感受、态度和情感[6]。然而,当前多数在线教育平台仍以短视频、微课程等单向输出模式为主,其设计的留言区、讨论区等交互模块往往缺乏即时响应机制,难以为用户提供有效的实时反馈与针对性指导。这种单向度的知识传递模式不仅限制了用户间的深度交流,而且导致在知识内化过程中无法获得必要的学习支持与动态调整。换言之,就是难以适应用户对教育资源服务的即时性、交互性和个性化等核心诉求。"人的意志动机来源于需求,是受教育者主动自觉接受教育的基本驱动力。"[7]基于智能算法的个性化推荐,与用户需求高度适配。这种动态的服务供给,能够有效增强学习投入度与获得感,激发用户从被动接受转向主动建构,从而促使其积极参与到教育教学的活动当中。此外,生成式人工智能凭借其解析复杂文本、实时情境对话、仿真场景创设、生成启发性内容等方面的技术优势,能够实现个性化内容解析,系统完成动态学情诊断,提供适配学习者认知水平的启发性内容,为增强用户与教育资源的交互提供了更多可能。以OpenAI发布的GPT-4o为例,语音交互响应速度显著提升,平均达320毫秒(最快232毫秒)[8]。这一技术突破使得智能教学系统能够以趋近人类自然对话的节奏,实时感知学习者的认知状态变化,并提供精准、及时的教学反馈。这种近乎自然的交互体验,不仅极大地激发了用户的交互意愿,保障了高质量的互动投入,而且为实现"诊断—反馈—优化"的精准化教学闭环提供了关键技术保障。

(二)从虚拟到沉浸:重塑学习环境生态

习近平总书记指出,"要充分发挥人工智能优势",加快发展"更加开放灵活的教育"[1]。生成式人工智能强大的语言理解和生成能力,正推进教育走向智慧化问答的新阶段,不仅能即时答疑解惑,还能提供情境化的学习体验,触发用户的好奇心和探

索欲,从而促进知识习得方式从被动接收转化为主动构建。

第一,构建智能对话系统,深化人机交互体验。对话是教学的重要环节,正如有学者所指出的,在教学方式中起相互作用的对话是优秀教学的一种本质性的标识[9]。相比传统预设路径下的人机对话,生成式人工智能作为一种根据自然语言对话提示自动生成响应内容的人工智能技术,通过借助人类反馈的强化学习、文图对比预训和可扩散模型等机制[10],具备开展多轮对话的卓越能力。用户与智慧平台的交互形式也由传统的菜单式向递进式问答转变。正如苏格拉底式教学法,以高频次的启发式对话,激发学习者的批判性思维与自主探究意识,最大程度地促进了其创新能力的培养。在万物互联的工业4.0时代,生成式人工智能支持下的大语言模型扮演着“智慧导师”的角色,展现出高度的“类人性”,不仅能够对输入文本进行语义分析与上下文感知,快速应答用户委托的复杂任务;还能综合对话情景及行为数据的关联性分析,洞察内在意涵,识别行为意图,锚定学习者深层次学习需求及困惑障碍,进而动态调整学习内容的难度与广度,针对性展开教育内容。高密度、可持续、实时的人机对话及多轮对话反馈,有效克服了响应不及时、供需不对称、适配性不足等局限。这种更具温润度的情感关怀与更高阶的思维启迪,能更好地引导学习者在不断追问和思考中加速其认知图示的自我组织,持续思索“是什么”“为什么”以及“如何可能”,助力学习者从知识的被动接受者转变为知识的积极建构者与意义的主动探寻者,从而在沉浸式交互过程中不知不觉实现知识的内化与重构。除了技术能力之外,在OpenAI最新发布的GPT-5.1模型中,系统引入的个性化与音调控制功能,允许用户从友好、专业、古怪和书呆子等预设风格中自主选择,从而实现对模型对话风格的自定义配置[11],变相增加了交互过程中的代入感、沉浸感与沟通效率。

第二,拓展教育场域,创设拟真多变的学习情境。汇聚数据、算力、算法,生成式人工智能以算法模型为驱动,实现了文本、图像、语音和视频等多模态内容的生成,不仅能根据用户需求高品质输出响应内容,还能通过深度学习场景的构建创设具身、沉浸、真实的教学交互场域,进而增进学习投入,提高用户的学习体验与自我效能感。虚拟现实(VR)与增强现实(AR)技术通过多感官交互和三维动态模拟,重构了传统学习场景,打破了教育在物理场域上的局限,为开展跨时空交互与场景化学习提供了技术基础,有效规避了传统教学中知识与真实应用场景相脱节的“离境学习”困境。具体而言,用户可通过虚拟实验室操作、历史场景重现或微观世界探索等多样化路径,身临其境地参与到学习过程中,真正实现“从做中学”[12]。以Sora构建的“中世纪生活情境”为例,该系统通过模拟古罗马时期的辩论场景,引导学习者以提问、回应、辩论等形式,将认知、情感与行为投射于虚拟场域之中,创设能够激发好奇心与探索欲的探究性学习情境,使其在沉浸式体验不同历史角色与社会情境的过程中,实现对历史知识的理解与建构。

二、隐忧与挑战:教育领域生成式人工智能应用的伦理风险图景

生成式人工智能大模型与教育领域的深度融合,正引发革命性影响,然而在数据治理、算法决策、内容生成及大模型应用的全生命周期中,这一技术范式也不可避免地衍生出一系列深层次的伦理挑战。

(一)数据隐私与算法治理隐忧

生成式人工智能作为大数据驱动的自动化数据训练与自动化内容生成的技术范式,以海量数据为原料,在教育数据的采集、处理和应用全流程中,普遍存在用户数据过度收集与泄露风险,引发数据安全伦理风险。

第一,数据采集、存储与运用环节可能涉及隐私泄露风险。数据是助力网络教育进步的重要生产要素,高增长、多类别用户数据的收集、分析、挖掘更是网络教育活动有效开展的前提和基础。大模型对用户数据汲取的体量超越了以往任何技术,依托强大的算力与智能算法模型,可以轻而易举地从多元异构数据中实现精准的信息检索与定位。如此庞大规模的数据包括但不限于大量敏感信息,如用户个人身份、兴趣爱好、心理特征、数字行为轨迹等,实质上构成了一种持续性的“数字监视”。在教育活动中,生成式人工智能往往通过网络爬虫、传感器感知等技术途径抓取网络数据,以实现对主体需求的精准识别以及教育资源的精准化、分众化供给,提高教育的智能化水平。理论上而言,大模型采集到的全部数据,其使用范围仅限于教育教学活动或教育研究,而不能将其公之于众。但具体实践中,未经信息主体同意便非法收集,甚至将未经脱敏处理的数据共

享给第三方的现象屡见不鲜。也就是说,当数据访问控制机制存在执行失效或监管缺位时,用户的信息便有可能被非授权人员访问,私人信息一旦“公之于众”,人人都有可能成为“透明人”。这不仅带来安全威胁,也可能引发信任瓦解,教育过程的伦理性也将面临根本性质疑。

第二,信息茧房固化认知,打破数据生态平衡。作为支撑教学精准化的关键技术,算法推荐系统在提升知识传递效率方面展现出显著优势。然而,其基于协同过滤的运作机制,在实现个性化推送的同时,迎合用户偏好的“投喂”可能引发“认知偏食”现象,暗含认知窄化与价值偏见固化的伦理陷阱。教育心理学研究表明,过度依赖AI生成的用户,其知识图谱呈现明显的“蜂巢型”结构——核心概念高度强化而边缘知识严重萎缩。算法为了迎合用户需求,将其喜好无限放大,甚至“一本正经”编造貌同实异的答案来迎合用户的偏好,致使同质化信息扑面而来,形成“回音壁效应”。在主观兴趣的影响下,用户往往选择与自己偏好、观点一致的内容。也就是说,只听得到自己想听的声音,看得到自己想看的信息。个性化推送编织的“信息茧房”,使用户无暇顾及“茧房”外多样化的世界,逐渐迷失在智能算法为其定制的拟态环境中。当用户长期处于AI构建的认知舒适区,会在无意识间强化算法对其行为数据的依赖。此时,携带特定偏好标签的数据通过算法的反馈循环被不断巩固和增强,这不仅会固化内容推送路径,还可能进一步加剧用户的认知偏差。

(二)技术依赖与教育主体性建构危机

教育作为培养人的主体意识、批判思维与自主能力的核心场域,其本质是“人的唤醒”而非“技术的规训”。然而,当前生成式人工智能的应用逻辑——从个性化学习内容的自动生成到教学交互的算法中介化——正悄然重塑学习认知、人际关系与教育实践,这种转变可能导致教育主体能动性的弱化以及情感交流和人际互动的减少。

第一,技术沉溺引发主体能动性弱化。生成式人工智能凭借强大的自然语言交互、多模态数据整合以及逻辑演绎推理等优势,正成为驱动教育创新发展的新动能。当下,生成式人工智能通过深度神经网络对海量人类语料进行表征学习,不仅能够轻松应对多元对话情景,精准把握不同学科领域的语义特征与表达规范,而且可以快速解释概念、总结文章并提出创造性观点,使用率呈爆炸式增长。用户逐渐欣喜于新技术带来的“一站式服务”,倾向于将更多的任务交由智能算法完成,这一新技术甚至成了部分用户解决学习问题的“灵丹妙药”。研究显示,英国大学生在学习中使用生成式人工智能的情况已经达到惊人的92%[13]。然而,“无时不在”的生成式人工智能看似可以“人性”地思考,并生成类似于人甚至超越于人的创造性答复,但究其本质,生成式人工智能只是对已经存在的事实知识的逻辑演算,是以海量数据为支撑的纯计算行为,而非有意识的自主行动,更不具备从无到有的创造性。在智能工具编织的“美丽新世界”的笼罩下,用户极易陷入信息茧房而不自知,算法黑箱的不可解释性使得学生被动接受推送,难以突破“信息偏食”的认知陷阱,很难不受到信息拜物教的控制。依赖的加剧,逐渐模糊了技术工具与智能伙伴的边界,影响了用户自主思考的能力,进而导致批判性思维的受损和认知能力的退化[14]。正如马克思所言,“我们的一切发明和进步,似乎结果是使物质力量成为有智慧的生命,而人的生命则化为愚钝的物质力量”[15]。而一旦独立思考让位于算法推荐,创新思维屈从于数据拟合,用户自由全面发展的创造性特征可能会被生成式人工智能所消弭,价值理性和独立思想会受到压制和束缚。在此过程中,现代科技逐渐沦为工具理性支配下的“压抑人性的异化力量”[16],导致人的主体性地位受到工具理性的僭越而消解,进而陷入主体性失落的困境。

第二,过度依赖人机交互,可能导致人际互动与情感交流的淡漠。在教育活动中,知识的交流和情感的互动都是不可或缺的。然而,现行教育大多过分强调成绩的至关重要性,常常忽视对学生情感需求的及时回应,师生之间缺乏必要且深入的情感联结。生成式人工智能的介入正在重塑人际交往的方式,拓宽情感表达的空间。当学生在困境中精疲力竭而无法得到鼓励时,高度拟人化的AI能够及时作出符合预期的情感共鸣,互动沟通中所展现出的“人情味”会给予用户力量,一定程度上缓解了受教育者深层次情感需求的紧迫性。随着用户与AI互动频率的增多,他们越来越倾向于向“无时不在”的人工智能寻求情感慰藉。但值得注意的是,AI所提供的情感支持并非真实的生命体验与情感投入,而是一种无主体、无责任的情感“拟像”,只能暂时弥补教师“不在场”的缺憾。当人机交互的工具理性凌驾于人际交往的价值理性之上,互动中蕴含的情感价值与共情理解便面临被消解的风险。也就是说,AI若仅作为效率工具被无反思地采纳,则可能

将教育推向一个去主体化、去关系化的危险境地。长此以往,甚至会衍生出虚拟世界高谈阔论、现实世界却沉默寡言等问题,无形中可能会导致教育主体的客体化,削弱现实中的人际互动和情感交流。

(三)算法偏见与价值导向的伦理困境

依托精准化的算法技术和自适应人机交互模式,生成式人工智能正在重塑意识形态的内容和传播方式。通过对多模态信息的处理以及持续的用户互动,不仅能够输出与用户喜好和习惯、认知结构、价值追求相匹配的内容,还能通过个性化、交互式的传播方式,潜移默化地影响用户的认知框架和价值判断,从而提升意识形态传播的渗透力和持久性。然而,技术的不当使用也为传播内容失真、价值渗透、舆论操控带来了可乘之机,为网络意识形态的斗争提供了更加隐蔽的智能化载体,严重影响了网络生态。

第一,算法偏见加剧错误价值观念的隐性渗透。生成式人工智能文本生成的结果受其自然语言模型的影响,这一模型本质上依靠算法选择和用于训练的大规模数据库。人们常常认为,以强算法、大数据为基础的机器决策可以克服人类因主观认知或知识局限而导致的偏见,是绝对客观的、公正的。然而,这种客观性往往掩盖了其内在的价值负载性。有学者提出,“丰饶论者所依据的判断结果都是以自动化方式生成的,因此很有可能是错误的、偏颇的或者具有破坏性”[17]。算法并非价值无涉的纯粹工具,而是社会结构与权力关系在技术层面的再现。算法设计中甚至“存在隐蔽的种族主义”[18]。研究发现,西方发达国家凭借其在数字技术领域的结构性优势,能够通过更为隐蔽的方式助长涉华虚假信息的生成与散播,制造虚假的共识表象,进而直接或间接地操控国际舆论[19]。同样,在教育场景中,看似合理、客观的人工智能应用却存在技术导致的偏见与歧视,暗藏着数据处理者本身对数据的价值判断与选择,所谓“价值中立”的算法本质上是一种认识论迷思。算法偏见所导致的刻板印象输出与文化忽视,会潜移默化地影响学习者的价值判断和文化认同,在实质上构成一种数字时代的暴力。这不仅与教育所追求的多元包容、客观理性的价值导向形成冲突,甚至可能助长文化霸权主义的蔓延,偏离技术向善的伦理旨趣。

第二,潜藏舆论操控风险,冲击主流意识形态影响力。生成式人工智能在教育领域的运用呈现出自我悖谬的特征:其高效的多模态内容生成能力在赋能教育创新的同时,也可能异化为制造和传播虚假信息、煽动性言论的技术工具,直接威胁到个体认知自主权与社会公共理性的培育。在人机交互过程中,基于对用户行为轨迹与偏好特征等海量异构数据的深度整合与智能分析,能够实现高度适配不同数据特点和需求的个性化定制。然而,其在发挥工具理性效率优势的同时,也不可避免地吸收人类语言与行为数据中潜藏的种种偏见与价值倾向,进而被内嵌于某种意识形态与权力框架之中。这使得人工智能有可能从教育的辅助工具演变为强化特定叙事、助推认知操控的基本载体。如此,在教育活动中,当生成式人工智能被用于制造虚假共识、煽动对立情绪或消解主流价值时,它不仅破坏教育作为文化传承与价值引领的公共职能,更可能危及社会团结与意识形态安全,造成价值共识的“离心效应”。

三、迈向善治:生成式人工智能教育伦理风险的纾解路径

生成式人工智能是教育数字化转型的新引擎,规避生成式人工智能应用于教育的风险,需要从技术治理、主体性强化、价值引领等方面共同发力,筑牢风险监管和治理的“堤坝”,在趋利避害中实现生成式人工智能对教育的有效赋能。

(一)优化算法治理,强化风险防控伦理规约

基于用户个人信息与学习行为数据面临泄露风险及虚假信息泛滥的安全考量,强化隐私保护机制与数据安全监管十分必要。针对这一问题,需要从制度规范、技术优化、算法治理等多维度构建治理框架,确保教育数据在赋能教学的同时,始终处于安全可控的边界之内。

第一,建立基于伦理规范的智能规则库,规范教育数据全生命周期管理。从数据正义的视角看,公平的数据治理要求保障所有教育参与者对其个人数据享有基本的知情权、访问权与控制权。为平衡数据开发与隐私保护之间的关系,需建立覆盖数据全生命周期的防护机制,将学生视为数据关系中的权利主体而非被动的数据客体。在数据采集环节,贯彻“最小必要”原则,明确数据采集的范围、目的和边界,确保数据采集获得充分的知情同意,避免形式化的“点击同意”。在数据存储阶段,应建立教育数据分类分级管理制度,构建数据安全存储的“防护网”。针对高敏感数据,采用加密存储+独立服务器隔离,仅限“人脸+密码”双重验证访问,防止信息泄

露、丢失或非法篡改；针对日常学习数据，需经严格的脱敏处理后方可进行云端存储，既充分保护隐私，又支持学习分析报告的生成。在数据处理与应用阶段，通过差分隐私技术，对采集的原始数据进行去标识化处理，移除或加密那些直接标识个人身份的信息。同时，设置通畅的用户反馈与响应渠道，对数据处理过程中可能存在的敏感性、偏见性或不当信息进行动态监测、预警与拦截，鼓励报告发现的问题，形成人机协同的内容治理机制。

第二，运用联邦学习技术，筑牢教育数据安全屏障。联邦学习作为一种分布式机器学习范式，能够在不集中原始数据的前提下实现多方协同建模。在这一模式下，教育平台、智能终端等主体仅需上传加密处理的模型参数，而无须共享学习成绩、行为轨迹、身份画像等敏感信息的原始数据。这种“数据不动，模型动”的技术路径，显著降低了因数据集中存储与流转而引发的隐私泄露、违规滥用等伦理风险。与此同时，通过增强差分隐私技术，对敏感学习行为数据添加噪声处理，确保用户数据在收集、传输与生成环节的匿名性与不可篡改性。这一做法既能防止算法对个人偏好的过度挖掘和固化输出，又能有效破解“数据孤岛”与“隐私保护”的两难困境。

（二）唤醒主体意识，构建人机协同的伦理秩序

教育作为一种以人的发展为根本指归的实践活动，其本质在于促进主体在知识架构、能力素养、价值观念等维度的全面发展与完善。但生成式人工智能的工具理性扩张却可能导致“人的异化”，使用用户面临被技术逻辑支配的伦理困境。为此，亟待通过主体性复归重构以人为中心的人机伦理秩序，发展和提高用户的主体性地位。

第一，构建人机协同的新型教育模式。有学者指出，“技术是我们在环境中以各种方式使用的那些物质文化的人工物”[20]。该学者提出的“人—技术”关系理论开始关注人与技术之间的关联，超越了传统主客间的二元对立。技术既非纯粹被动的工具，也非自主塑造社会的力量，而是与人相互塑造的关系性存在。因此，强化教育场景的人机协作力，应发挥生成式人工智能和教育主体的各自优势，使算法的高效性和人的能动性形成互补，共同推动教育的创新发展。一是发挥生成式人工智能赋能教育的技术优势。技术是人类感知与行动能力的具象化延伸，生成式人工智能凭借其多模态感知与自然语言理解的技术优势，能够实时捕捉覆盖教学全场景的动态数据。这使得教育规律的发现从局部样本的归纳推断迈向基于全域数据的建模推演，有效缓解了信息过载引发的认知负荷，从而为创新思维的激发与学习效率的跃升注入了持续的技术动能。二是理性审视生成式人工智能的工具属性。“机器文明的一切机制都必须服从于人的目的，人的需求。”[21]在人机交互的教育实践中，明确生成式人工智能只是一种更为高效与便捷的技术手段，其核心价值在于拓展而非替代人类认知活动。因此，在具体学习场景中，既要充分发挥技术的赋能价值，将其作为熟练且智能的思考伙伴，也需保持独立思考的批判意识，最终在人机优势互补中推动教育效能的提升。

第二，立足学习者发展，提升主体认知能力。康德指出，“不论是谁在任何时候都不应把自己和他人仅仅当作工具，而应该永远看作自身就是目的”[22]。他强调，“人就是创造的终极目的”[23]。生成式人工智能强大的“类人性”表达使用户知识获取的途径从检索式向生成式转变，有效缓解了用户在信息处理、初步分析等基础性决策环节的投入。然而，过度的技术依赖也可能让用户陷入“决策闭环”，将本该属于用户的创造、创生能力让渡给了智能机器，认知判断能力、逻辑辨析能力、创新构想能力被不断削弱。面对这一危机，一是需要强化人的主体意识，提升自我调节学习能力。自我调节学习作为一种强调学习者主体性的认知策略，其核心在于通过系统化的过程监控与动态调整机制，实现对学习活动的自主管理。生成式人工智能通过辅助用户设定学习目标、选择和调整学习策略、管理和监督学习过程、评估和反思学习效果，不仅显著提升了学习任务的执行效能，更有效激发了学习者的内在学习动机和认知投入度，为学习者主动探索新知、深入发现问题、开拓认知边界提供了动力支持，充分激发了学习者的创造性潜能。究其根本，人的存在是一种无限的自我创造，而技术工具的价值，正是在于通过接管程序化任务，为个体能动性与创造性的发展提供赋能与支持。二是需要加强数字素养教育，理性对待智能算法技术。数字素养是指利用数字技术、数据资源，发现、分析和解决教育问题的意识、能力和责任。在认知维度，需开展网络安全和人工智能伦理教育，帮助学习者培养数字意识，从而有效识别并抵御算法可能带来的认知偏差与潜在偏见。在道德维度，培养学习者的数字责任感，使其在面对算法生成的内容时，能够自主进行价值判断与理性决策，审慎评估信息来源的可信度、内容的真实性与价值导向。在实践层面，鼓励学习者从被动的内容消

费者转变为能动的技术参与者与创造者，通过编程开发、智能系统设计等具身实践，深入理解生成式人工智能的技术边界，并在与之协作中不断提升自主性和创造力。

第三，强化情感教育，促进和谐人际交往。大语言模型通过注意力机制形成的复杂神经网络，能够针对用户的困惑输出较为详尽的应对策略，营造近乎真实的沉浸式交往体验，语言表达高度贴近人类自然语言的表征方式。但面对用户表露情绪时，系统的回复依然无法做到“动之以情”，在情感共鸣上缺乏一定的共情深度，难以复制或替代现实生活中的经验感知及人际互动。为解决这一问题，可以将情感计算技术引入教育领域。情感计算技术能够实时捕捉用户的面部表情变化、语音语调特征以及交互行为数据，从而准确识别其当前情绪状态。基于这些数据，生成式人工智能可动态调整交互策略，根据用户情绪状态提供相应的反馈和支持。例如，当系统识别到用户呈现焦虑状态时，将自动切换至更温和、鼓励性的话语模式；当感知到用户表现出浓厚兴趣时，则会适时推送更具挑战性的问题以维持其投入度；当监测到用户严重的心理波动信号时，将自动生成评估报告并向教师发送预警信息，帮助他们制定更具针对性的关怀策略与干预方案。这种智能化的交互优化，使人机对话更具温度，能够有效提升学习体验。

（三）坚守价值引领，筑牢风险管控的伦理防火墙

习近平总书记强调，要积极“探索将人工智能运用在新闻采集、生产、分发、接受、反馈中，用主流价值导向驾驭‘算法’，全面提高舆论引导能力”[24]。在当今全球化浪潮中，多元文化相互交织、彼此碰撞，对主流意识形态的传播和引导带来了巨大挑战。基于此，必须牢牢把握舆论导向，强化主流意识形态在生成式人工智能领域的价值引领，筑牢主流意识形态防线，以实现生成式人工智能对主流意识形态的正向赋能。

第一，构建算法风险动态监测与响应机制，筑牢教育伦理防线。在网络空间，人人都有麦克风，人人都是发声者，已重构为全民参与的传播生态。“每一行代码、每一个界面，都代表着选择，都意味着判断，都承载着价值。”[25]人工智能技术已不是价值无涉的纯粹工具，其设计逻辑天然嵌入了开发者的认知范式与价值预设，承载着特定文化基因与意识形态，可能会诱发一定的意识形态风险，冲击主流价值认同。因此，需建立覆盖数据输入、模型训练、内容生成与传播反馈全链条的“气象雷达站”，及时捕捉网络舆情，从而实现算法风险的前瞻性研判与自适应调控。通过实时语义分析和多模态风险感知，自动标记涉及错误思潮的表述，筛查是否存在西方价值观偏见、历史虚无主义倾向及危害国家安全、煽动歧视、故意抹黑等意识形态风险点，有效做到对与主流价值认同背道而驰的错误舆论的早发现、早处理，有序引导舆论走向，确保思想价值观念传播的正确导向。

第二，掌握议题设置主动权，增加主流意识形态教育内容供给与传播效力。习近平总书记指出，“在信息生产领域，也要进行供给侧结构性改革”[26]。算法作为生成式人工智能的重要要素，凭借逻辑化数据运算、自适应推理及智能化决策生成等特性，可以实现庞杂数据中关键内容的高效输出。鉴于此，需要从以下几方面着力：一是优化内容供给。要深入挖掘主流意识形态教育资源，并借助自然语言处理技术将其编码为机器可读的结构化数据，形成“意识形态素材库”；在算法架构中内嵌中华优秀传统文化的价值编码，主动设置议题，利用智能算法推动主流意识形态数据的精准投递和智能传播，加大优质内容“喂养”，构筑起植根中华文明基因的意识形态传播体系，增强主流价值认同。二是创新议题设置策略，增强主流意识形态的吸引力。通过“议题融合”与“议题再造”两种路径，将主流价值观融入用户关注的热点话题，在互动中实现主流价值的自然浸润。同时，结合重大历史节点或社会事件，主动设置“红色传承基因”“新时代奋斗者”等议题，通过算法生成短视频、互动 H5 等新媒体产品，提升意识形态教育的传播力。三是构建常态化议题效果评估与反馈机制。通过量化分析点击率、停留时长、评论倾向等用户行为数据，精准评估主流意识形态内容的传播广度与认同深度，进而为议题设置策略的持续优化提供反馈与指引。具体而言，当系统监测到特定类型议题参与度持续低于预期阈值时，可借助算法模型分析学习者的兴趣图谱，动态调整内容传播策略或叙事方式。同时，将师生在议题讨论中产生的优质观点纳入素材库，由此形成“生成—反馈—优化”的良性循环，持续提升议题的传播效能与受众覆盖广度。

结　语

生成式人工智能技术带来的美好图景好似“电

子革命的神话”[27],引发了教育领域的数智化转型,也提出了前所未有的挑战。在教育实践中,既不能对技术盲目崇拜而偏离教育本质,陷入技术依赖的误区;也不能因其潜在风险而因噎废食,否定其作为革命性技术所蕴含的赋能潜力。作为推动生产工具革命性变革的关键动力,迅猛发展的人工智能技术的确是焕新育人空间、重塑人机交互、拓展认知边界的赋能技术。因此,在促进教育发展的过程中,还需规避技术宰制引发的教育本质的异化,从而确保生成式人工智能与教育现代化发展同频共振、和谐共生,与构建高质量教育体系的战略进程偕行。

参考文献

[1]习近平向国际人工智能与教育大会致贺信[N].人民日报,2019-05-17(1).

[2]中共中央　国务院.教育强国建设规划纲要(2024—2035 年)[EB/OL].(2024-01-19)[2025-01-19].http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202501/t20250119_1176193.html.

[3]教育部等九部门关于加快推进教育数字化的意见[EB/OL].(2025-04-11)[2025-04-11].https://www.gov.cn/zhengce/zhengceku/202504/content_7019045.htm.

[4]马克思恩格斯文集:第 1 卷[M].北京:人民出版社,2009:188.

[5]白雪梅,郭日发.生成式人工智能何以赋能学习、能力与评价?[J].现代教育技术,2024(1):55-63.

[6]诺曼.设计心理学 1:日常的设计[M].小柯,译.北京:中信出版社,2015:8.

[7]韩职阳,曹洪君.分众化思想政治教育:生成逻辑、内涵要求与推进进路[J].思想政治教育研究,2024(2):63-69.

[8]Hello GPT-4o[EB/OL].(2024-08-08)[2024-08-08].https://openai.com/index/hello-gpt-4o/.

[9]钟启泉.对话与文本:教学规范的转型[J].教育研究,2001(1):33-39.

[10]温旭.新质态与新境遇:生成式人工智能赋能思想政治理论课内涵式发展论析[J].思想教育研究,2025(2):94-101.

[11]GPT-5.1: A new era for ChatGPT[EB/OL].(2025-11-12)[2025-11-12].https://openai.com/zh-Hans-CN/index/gpt-5-1/.

[12]杜威.学校与社会·明日之学校[M].赵祥麟,任钟印,吴志宏,译.北京:人民教育出版社,2005:274.

[13]Kortext & The Higher Education Policy Institute.Student Generative AI Survey 2025.[EB/OL].(2025-02-26)[2025-02-26].https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/.

[14]DENG X J, YU Z G. A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education[J].Sustainability,2023(4):29-40.

[15]马克思恩格斯选集:第 1 卷[M].北京:人民出版社,2012:776.

[16]马尔库塞.单向度的人:发达工业社会意识形态研究[M].刘继,译.上海:上海译文出版社,2006:14.

[17]帕斯奎尔.黑箱社会　控制金钱和信息的数据法则[M].赵亚男,译.北京:中信出版社,2015:26.

[18]GIBNEY E. Chatbot AI makes racist judgements on the basis of dialect[J].Nature, 2024(627):476-477.

[19]黄日涵,姚浩龙.被重塑的世界? ChatGPT 崛起下人工智能与国家安全新特征[J].国际安全研究,2023(4):82-106.

[20]伊德.技术与生活:从伊甸园到尘世[M].韩连庆,译.北京:北京大学出版社,2012:1.

[21]芒福德.技术与文明[M].陈允明,王克仁,李华山,译.北京:中国建筑工业出版社,2009:377.

[22]康德.道德形而上学原理[M].苗力田,译.上海:上海人民出版社,2005:53.

[23]康德.康德著作全集:第 5 卷[M].李秋零,译.北京:中国人民大学出版社,2007:454.

[24]习近平.加快推动媒体融合发展构建全媒体传播格局[J].求是,2019(6):4-8.

[25]叶蓁蓁.主流媒体引导力,可否这样实现?[J].新闻战线,2018(15):13-15.

[26]中共中央党史和文献研究院.习近平关于社会主义精神文明建设论述摘编[M].北京:中央文献出版社,2022:89.

[27]凯瑞.作为文化的传播[M].丁未,译.北京:中国人民大学出版社,2019:103.

Ethical Risks and Countermeasures for Generative AI Applications in Education

Yang Weitao　　*Li Chenwei*

Abstract: Generative AI, exemplified by ChatGPT, is triggering profound changes in science and technology as well as the humanities. Its application in education demonstrates immense potential and value, including advancing precision teaching reforms and driving the digital-intelligent transformation of teaching scenarios. Generative AI undoubtedly serves as a vital engine for the future of education. However, as this technology remains in its nascent development phase, its practical application in education still faces limitations, giving rise to multiple ethical risks such as data security breaches, erosion of agency, and ideological infiltration. To mitigate risks and maximize technological efficacy, it is imperative to fortify data security defenses, restore human agency, and reinforce the guidance of mainstream ideology. By pursuing a three-pronged approach of "technological regulation, humanistic care, and value-driven leadership," the educational application of generative AI can be standardized to boost the high-quality development of education.

Key words: generative AI; education empower; ethical risks

责任编辑:思　齐