全球高等教育应对生成式人工智能风险的举措与思考

生成式人工智能技术的蓬勃发展为高等教育创新与变革增添了新动力，但也带来了新的风险挑战。本文分析高等教育应用生成式人工智能技术时面临的几个主要风险挑战，梳理全球高等教育界形成的主要应对举措，进而提供防范化解可能风险、推动负责任的人工智能正向赋能高等教育的建议参考。

文丨**沈锦璐 李建慧 冯国楠** 中国电子信息产业发展研究院

一、高等教育面临的风险挑战

人工智能技术迭代加速，尤其是在生成式人工智能技术广泛渗透各行各业的当下，其安全性议题已成为国际社会关注焦点。高校作为科技创新与人才培养的前沿阵地，在积极推进人工智能融入教学、科研与管理的过程中，亦衍生出一系列深层次挑战。

（一）师生创新思维出现技术依赖

生成式人工智能技术的实质是基于数据训练与概率预测生成内容，其产出内容往往是同质化的创意启发与案例集成，难以体现多元化表达和原创性突破。因而，当人机协作异化对技术的路径产生依赖时，高等教育的“教与学”也将异化为标准化的工厂生产，衰减师生的创新能力与批判性思维。在教师层面，若惯于借助人工智能技术撰写课程方案、生成教学材料，将会逐渐丧失教学实践的主动性，使得教学思维与内容趋于僵化，最终沦为智能技术的被动执行者；在学生层面，如果求学期间面对困难时习惯等待技术给出答案，其将逐渐缺失学习自主性、弱化批判性思维，再难以突破主流思维框架实现真正的创新表达。

（二）伦理价值体系受算法“黑箱”挑战

受限于人工智能模型的训练数据偏

差性、有限可解释性与决策过程的复杂性，高校在教学评估、学术评价、科研辅助等应用场景中，无法判断算法是否存在偏见、数据是否被合理使用，导致公平、公正等伦理价值难以落实。例如，当部分群体数据被过度代表或不足代表，则会由算法在学习过程中转化为模型的“认知偏好”，进而在生成结果中表现出对特定群体的偏见，造成不公平甚至歧视性的输出结果，可能影响教育评价的公正性、学术发表的可信性等。

（三）传统学术评价体系面临新型考验

在生成式人工智能重塑学术生态的当下，虚假科研信息制造、论文快速代写等新型学术失信行为开始出现，真实可信知识和未经验证的信息杂糅其中，挑战知识科学性以及学术合规性。而面对人工智能生成内容的批量产出，传统查重工具与评审机制在识别“人机协同创作”成果时力不从心。所谓的“智能创作能力”模糊了学术成果原创与技术辅助的界限，导致学术评价标准与质量把控体系陷入被动调整的困境，也使得“单一作者承担学术责任”的传统责任追溯模式难以简单复用。

（四）数据隐私安全防护体系压力增大

人工智能技术对海量数据进行深度挖掘与高频调用，过程中可能形成训练数据的不当采集、模型运行中的数据滥用、云服务提供商的数据隔离失效等风险链条，而生成式人工智能技术的“黑箱”特性与数据处理的自动化流程，使得隐私泄露风险更具隐蔽性与复杂性。

高校对云数据的监管存在一定的难度，大学师生提交的输入数据可能含有未适当保存的个人隐私信息，均存在被不当留存、处理并以隐性方式传递给后续用户的风险，可能造成数据泄露和隐私侵犯。

二、全球高等教育应对风险挑战的主要举措

为有效应对生成式人工智能技术带来的风险挑战，全球高等教育界积极行动，相继发布一系列指南、建议及规范文件，为生成式人工智能的应用划定边界、明确标准，提升人工智能技术应用的安全性与合规性，以期推动高等教育和人工智能技术的深度融合与健康发展。

（一）引导学生正确应用人工智能开展创新活动

面对学生创新性与批判性思维可能因过度依赖人工智能而衰减的潜在危机，高校对教师如何开展课堂教学提供指导建议，鼓励师生加深讨论并协作建立应用人工智能的规则，引导学生提升人工智能应用素养，将人工智能转化为实现创意构想的“加速器”。

美国东北大学提出，教师可以要求学生根据人工智能的响应内容进行核查，讨论“工具能充分回答哪些问题？”“局限性是什么？”“这种反应在哪些方面是完全错误的？”等问题。也可以选择让学生使用 ChatGPT 开发连续的工作草稿，记录改进每个草稿的提示并分析提示的优点或局限性，由此提升批判性思

维能力和信息素养。

威斯康星大学麦迪逊分校建议教师引导学生们以作业形式制定一个使用人工智能的计划，阐明哪些步骤使用人工智能是有价值和适当的、何时需要原创的想法和创造力及其原因。

杜克大学认为标准化的、“一刀切”的人工智能规则对高等教育而言并不可持续，鼓励教师深思熟虑地使用人工智能并与学生就这一主题进行讨论，进而共同制定一套清晰的人工智能规则。

（二）融入伦理教学，实现负责任的人工智能

生成式人工智能的深度渗透与广泛应用不可避免地带来系列伦理问题，高校积极破局，将人工智能伦理相关模块有机融入课程教学、师资培训及学术实践等环节。

哈佛大学提出无论人工智能如何继续发展，道德需要处于对话的最前沿并融入教育，采用“嵌入式伦理”计划将哲学和伦理模块编入计算机科学课程中。

米兰大学推出“人工智能素养计划”，在 2025 年面向教职工提供将计算机与技术技能同伦理、法律及职场相关技能相结合的培训项目。

中国科学技术大学和中山大学联合完成了人工智能伦理和治理领域教学资源共享服务平台知识图谱建设与《人工智能与技术伦理》课程建设，其课程特别面向人工智能大模型和生成式人工智能的新进展与新挑战，开展人工智能伦理和治理的系统性培训。

此外，高校还关注到人工智能生成可能存在的内容偏见，比如哥伦比亚大学要求用户考虑生成式人工智能工具的数据输入和输出是否会根据种族、民族、国籍、年龄等个人在适用法律下的，受保护分类生成对个人造成不同影响的决策，并提醒用户勿依赖任何有潜在偏见的输出。

（三）维护学术诚信，促进人工智能工具向善

面对生成式人工智能技术带来的学术诚信新挑战，各高校制定规则清晰界定学术活动中使用人工智能工具“可为”与“不可为”的边界，形成抵御学术失范风险的防线。

英国罗素大学集团在其“关于在教育中使用生成式人工智能工具的原则”中指出，其 24 所大学均已审查学术行为政策和指导方针，在政策中明确告知学生和教职员工在哪些情况下使用生成式人工智能是不合适的，并将持续监测生成式人工智能工具融入学术生活的有效性、公平性和伦理影响。哥伦比亚大学指出“生成式人工智能可以生成不存在的论文引用，并且还被用来为从未实际进行的实验生成图像”的虚假科研信息生成问题，要求研究人员负责研究输出中包含的人工智能创建内容的准确性，并且必须谨慎使用研究中的人工智能输出。

复旦大学发布《复旦大学关于在本科毕业论文（设计）中使用人工智能工具的规定（试行）》，明确 4 个“允许使用范围”和 6 个“禁止使用范围”，并提出将根据人工智能技术的发展不定时地进行修订。

（四）开发特定人工智能工具，保护数据隐私安全

站在抵御数据安全风险的关键节点，高校在发布“人工智能工具可接受使用范围”的指导文件之余，已开始自主开发校园内专属的生成式人工智能工具，以期平衡教育信息数据私有化与共享之间的矛盾，为师生营造安全可信的数字化教育环境。

得克萨斯大学奥斯汀分校信息安全办公室明确禁止人工智能工具被用于专有或未发表的研究、法律分析或建议、招聘或人事决策、教师不允许的学术工作、创建非公开教学材料或评分非公开的产出。

密歇根大学开发了专属的封闭生成式人工智能工具，承诺师生收集和使用的数据将依托“信息资源负责任使用”“机构数据资源管理政策”等多种机制得到保护。

哈佛大学采用“人工智能沙盒”向用户提供了使用生成式人工智能的安全环境，确保师生输入的数据不会被用于训练任何企业的大语言模型，但用户能够通过专门界面访问来自 OpenAI 和 Meta 等最新的大语言模型。

三、对推动人工智能赋能高等教育发展的建议

全球范围内的高等教育实践为防范和化解生成式人工智能带来的诸项风险提供了有益参考。为进一步引导技术正向赋能高等教育，建议在以下几个方面持续发力，形成系统化的应对策略。

（一）探索“师－生－机”协作的良性互动模式，提升人工智能素养

在生成式人工智能深度融入高等教育的背景下，高校亟须权衡好“人”与“机”之间的交互协同关系，引导师生共同制定服务于特定教育教学阶段的人工智能使用规则，探索形成“机”服务于“人”的协作模式。

在教师端，厘清技术与教师之间的任务分工，既发挥人工智能在教学资源整合、学情分析等方面的效率优势，又凸显教师在价值引领、情感沟通、创造性教学设计中的不可替代性，设计人机协作的教学任务，服务知识传播与能力的提升。

在学生端，利用人工智能的个性化分析能力为学生量身定制教学辅导、职业生涯规划方案等，向学生推送适合其创新能力发展的学习内容。鼓励学生在与人工智能的交互过程中对生成内容主动提出质疑，并由教师给予指导、评估、反馈，借助技术工具与师生互动有效训练学生的批判性思维，拓宽思维边界。

（二）推动“教学－管理－开发”多端协同，降低伦理风险

面对人工智能引发的伦理风险挑战，需要教师和学生、高等教育管理者、技术开发者等多元主体凝聚共识、协作发力，共同维护人工智能发展的公平公正性。教师将人工智能伦理目标纳入课程整体框架，收集真实发生的人工智能伦理事件，组织案例分析与讨论，设置模拟场景，引导学生思考如何避免因数据偏差导致的歧视问题，培养学生在人工

智能技术开发与应用中主动遵循伦理规范的意识与责任感。高等教育管理者可以主动参与人工智能模型的开发和数据的训练，突破技术瓶颈，提升模型可解释性，确保技术服务于教学与学术发展，避免因技术失控引发伦理危机。可对参与模型构建、工具开发的技术人员进行针对性培训，开展关于人工智能伦理和公平性的培训课程，使其能够了解不同群体的权益和需求，提升对算法偏见和歧视问题的认识，进而在技术开发时能对公平性问题予以考虑。

（三）构建“制度－技术－平台”多维防护，筑牢学术诚信防线

生成式人工智能已深度渗透学术研究场景并带来新型学术失信风险，要求从细化管理制度规范、强化检测技术研发、优化共享平台搭建等方面维护学术诚信。在制度规范层面，高校应着手制定专门的人工智能学术规范，明确生成式人工智能工具在文本撰写、数据分析、插图创作、已有学术成果引用等方面的使用边界。

在技术研发方面，持续加强人工智能论文识别与反识别的检测研究，结合语义分析、语法模式识别等技术，深入剖析生成式人工智能的语句结构机械性重复、论证逻辑程式化、引用风格刻板化等文本生成特征，开发更具针对性的检测算法。

在平台搭建方面，基于全国或区域高校的资源整合，可探索搭建通用的课程作业与学术成果人工智能识别系统，实现跨校数据共享与检测标准统一。结合最新人工智能技术趋势发展，定期更新系统的检测规则与数据库，将新出现的人工智能写作特征纳入监测范围，形成动态化、全覆盖的学术诚信防护网络。

（四）实现“访问－使用－传输”全链可信，守护数据隐私安全

面对人工智能技术引发的数据隐私安全风险，高校需要在保障人工智能模型训练效果的同时，结合高等教育数据应用场景的动态变化，持续提升数据隐私安全防护能力。

在数据访问方面，科学划分数据访问权限，严格限制敏感数据的接触范围，定期对数据脱敏算法和访问控制策略进行更新优化；在数据使用方面，针对人工智能训练数据开展定期合规性审查，重点排查数据偏差、来源合法性等问题，防止因数据质量缺陷引发隐私泄露；在数据传输方面，高校应制定专门的人工智能伦理准则，要求师生在使用人工智能工具时严格遵守数据最小化原则，不得上传包含个人敏感信息的教学、科研数据，并可利用数字水印等技术为数据添加不可篡改的溯源标识。有能力的高校可以自主研发本地化部署的生成式人工智能工具，构建封闭、可控的运行环境，确保数据只能在内部可信系统中流转，有效保护师生个人隐私和学术数据安全。

责任编辑：孙姗姗　投稿邮箱 zhouhl@staff.ccidnet.com