生成式人工智能及其教育应用的基本争议和对策

苗逢春[1,2]

（1. 北京师范大学 互联网教育智能技术及应用国家工程实验室，北京 1000875；
2. 联合国教科文组织总部，巴黎 75007）

[摘要] 本文是对联合国教科文组织《生成式人工智能教育与研究应用指南》的系列解读第二篇，着重讨论生成式人工智能及其教育应用引发的基本争议。“基于工作过程”技术缺陷，生成式人工智能已引发加速数据贫穷、技术不透明导致服务辖区内治理缺失、未经许可搜集训练用数据、模型架构不可解释、基础模型不理解真实世界、生成的信息污染互联网、强势价值观投射、助长违法性深伪等多重争议。生成式人工智能会对平等、包容、学习主体能动性、价值观及语言文化多样性、知识建构的多元性等教育核心价值产生直接而深远的冲击，而这些核心价值应被秉承为考证生成式人工智能教育适用性的逻辑基点。决策者和实践者应遵循“优先管制、确保包容、引导应用”的逻辑路径，强化全系统监管法规和执法能力，确保教育生成式人工智能生态系统安全可信、自主可控、本地适用，进而通过能力建设和实践指导等措施引导合理的教育应用实践。

[关键词] 生成式人工智能；内容加工；训练数据集来源；公平、包容及语言文化多样性

[中图分类号] G434 [文献标识码] A [文章编号] 1007-2179(2024)01-0004-12

2022 年 11 月，美国开放人工智能研究中心（OpenAI Artificial Intelligence Research Center INC, OpenAI）发布了第三代聊天生成式预训练转换模型（Chat Generative Pre-trained Transformers, ChatGPT）—ChatGPT-3，开启了生成式人工智能（Generative AI）从研发转向商用和民用的新历史时期。在 ChatGPT 发布近一年时间里，其引发的影响及管制反弹主要体现为四个方面。1）垄断与多元。OpenAI、谷歌公司和 Meta 公司的生成式人工智能平台形成了贯穿基础模型、网络基础设施和文图音视内容加工等领域的垂直垄断。同时，其他大型公司和开源大模型社群等发起了基础模型开源化、平台选择多元化、语言文化多样化的研发追赶与生态布局抗衡。2）应用与替代。生成式人工智能在商业领域迅速推广，引发相关行业工作岗位快速自动化。3）争议与治理。生成式人工智能的安全和伦理威胁从理论忧患浮现为实际法律案例，形成坚信其积极变革潜力和忧虑其潜在人文威胁之间的对立，加速中国、美国、欧盟等国家和经济体的针对性立法。4）愿景与现实。迅速涌现的生成式人工智能正在颠覆和变革教育等社会服务领域，但与该技术对本土学生尤其是未成年人的教育适用性和实用性的理性研判之间存在明显断层。其中，社会各界对生成式人工智能可能引发的安全及伦理忧患众说纷纭、莫衷一是。联合国教科文组织 2023 年 9 月发布的《生成式人工智能教育与研究应用指南》（简称《指南》）（Miao, 2023）首次在对该类技术的工作原理进行溯源的基础上，系统总结了八个有关生成式人工智能的基本争议，进而揭示了争议对生成式人工智能教育应用的根本影响。《指

[收稿日期] 2023-11-16 [修回日期] 2023-11-17 [DOI 编码] 10.13966/j.cnki.kfjyyj.2024.01.001

[作者简介] 苗逢春，研究员，北京师范大学，联合国教科文组织总部部门主任，研究方向：人工智能与教育、数字学习政策、未来数字学校（f.miao@unesco.org）。

[引用信息] 苗逢春（2024）. 生成式人工智能及其教育应用的基本争议和对策[J]. 开放教育研究，30(1)：4-15.

南》针对性地提出应对这些基本争议的公共治理策略、生成式人工智能教育应用的政策和引导主体适用的人机互动应用的实践框架。

本研究是对《指南》的第二篇解读，聚焦于系统总结和剖析生成式人工智能及其教育应用的基本争议，并针对这些争议的起因和责任主体提出治理对策和实践应用建议。本研究的相关解读基于三个相互关联的基本假设：第一，人工智能科技创新、人工智能的安全可信性、包容平等的社会应用不应成为三难悖论（trilemma），人类应追求三维同频共振；第二，生成式人工智能对全社会及教育平等与包容、学习主体能动性、价值观及语言文化多样性、知识建构的多元性等教育核心价值的威胁最为直接和深入，这些核心价值应成为考证生成式人工智能教育适用性的逻辑起点；第三，生成式人工智能的教育应用应遵循“优先管制、确保包容、引导应用”的逻辑。

一、争议的技术起因

对生成式人工智能教育应用争议的讨论须以其工作原理、技术缺陷及其对社会的显性和潜在影响为依据。

（一）生成式人工智能工作原理及其训练用数据来源和语言分布

《指南》从人工智能对人类思维表征符号系统模拟的角度界定生成式人工智能：生成式人工智能是基于人类思维符号表征系统表达的提示工程（prompt engineering）自动生成内容的人工智能技术。生成式人工智能技术对借助各类符号表征系统呈现的内容进行模式识别和内容生产方面的性能日益强大，目前已能贯通文字、语音、声音、图像、视频、计算机编码等格式进行模式识别，并借助上述符合表征系统生成新内容。文本生成式人工智能使用人工神经网络技术的通用文本转换器，通常被称为“大语言模型”（Large Language Model），是一种利用从互联网网页内容、社交媒体对话和其他在线媒体收集数据进行训练的内容生成深度学习模型。文本或语音生成式预训练转换模型，可以对训练用数据集的各类句法模式进行识别和学习，然后经过反复训练、测试和优化，获得根据提示、通过重复执行事先确认的模式生成内容或提供答案的能力。其关键技术环节包括：1）将提示指令分解为人工智能可处理的文本最小单位字节（token）后，输入到生成式预训练转换器中；2）转换器根据从训练数据集中确认的语言模式，预测特定单词或短语在特定语境出现的概率，通过统计模型预测的拟合度组合为连贯反应的词语及其连缀方式（即句法），并借此预测后续最有可能使用的单词或短语；3）将预测产生的单词或短语转化为可阅读的文本（或可理解的声音）。可理解的文本或声音经过“护栏技术”（guardrails）过滤明显违法或不合标准的不良输出，并通过处理技术提高句法的拟人化程度和可理解性。上述过程不断循环重复，直到完成一个完整的响应。

图像或音乐生成式人工智能多采用生成对抗网络（generative adversarial networks, GANs）人工神经网络技术，并可与变分自编码器（variational auto-encoders）技术结合使用。也有图像生成式人工采取扩散模型（diffusion models）等无监督生成模型。例如，生成对抗网络模型由两个对抗器组成，即生成器（generator）和判别器（discriminator）。生成器针对提示识别图像或音乐要素组合模式并生成随机图像或音乐片段，判别器对比生成的图像或音乐与真实图像或音乐（或范例）之间的拟合度。生成器随后根据判别器的对比结果调整其使用的参数以便生成更优化的图像。通过千百次不断的迭代训练，生成器创作的图像或音乐越来越逼真。

生成式预训练转换器的功能依赖于模型架构、训练方法和预训练数据集的质量、数量和模型使用的参数。其中，参数是决定人工神经网络系统如何加工输入和产生输出的数值，它通过界定训练中的数据对模型的内容要素进行编码。参数的定义和数量决定预训练转换器的性能和应用表现。GPT-3 使用了约 1750 亿个参数，而 GPT-4 使用的参数据称达 1.8 万亿。从模型架构的成熟度、所用的参数规模、内容处理和生产能力、语言覆盖范围等方面考量，占全球垄断地位的大模型包括 OpenAI 的 ChatGPT 系列产品、Meta 公司的“羊驼”大语言模型（Alpaca）和 Meta 大语言人工智能模型（简称 Llama 大模型）、谷歌公司“诗人”大语言模型（Bard，基于谷歌的 PaLM2 基础模型）和“双子座”多模态大模型（Gemini）。已有生成式人工

智能模型的训练用数据集主要包括通过爬虫软件读取互联网网页信息、社交媒体对话信息、在线图书馆图书资料和互联网百科类平台的百科内容。以 ChatGPT-3 为例，其训练用文本数据（即语料）约 1 TB 左右（（即语料）约 1 TB 左右（Thompson, 2023），主要来源包括：自 2012 年以来持续通过“网络爬虫数据集”（common crawl）从互联网收集的数据，约占数据总量的 61.75%；通过“红迪”电子布告栏（Reddit）收集的点赞数超过 3 个的社交媒体发帖和讨论数据，约占 18.86%；两个在线图书平台（Library Genesis 和 Smashwords）的在线图书，约占 15.9%；维基百科数据，约占 3.49%。目前垄断性生成式人工智能模型的训练用数据集以美国和欧洲国家的语言为主。在 ChatGPT-3 的训练数据集中，英语语料约占 92.65%，欧洲各国语言占比超过 5%，汉语语料占比不到 0.1%。Meta 公司开发的 Llama 2 语料中，英语占比有所下降但仍占 89.7%，其他占比排前 15 的语言几乎没有改变，汉语语料占比为 0.13%（Touvron et al., 2023）。预训练用数据集和参数的几何级数增长要求超算能力同步加速。在超级计算支撑方面，从 2012 年到 2019 年，用于生成式人工智能模型训练的算力的翻倍周期为 3～4 个月（Stanford University, 2019）。

（二）生成式人工智能在内容处理范畴的集成性技术跃迁与潜在技术范畴瓶颈

生成式人工智能在多种深度学习技术中的综合应用、模型架构的优化、所用参数以千亿级为基点的持续细化、训练用数据的跨平台动态挖掘与叠增、处理海量数据和参数所需计算能力的周期倍增等核心技术和支撑技术领域都取得了集成性的突破。这种集成性技术突破在技术和实践领域产生了“逃逸效应”（runaway effect，又译为“失控效应”）。首先，生成式人工智能的近期成果表现为人工智能技术在跨符号表征系统数据加工和呈现方面的突破，提升了人类挖掘技术能力，由此加速了人工智能芯片、超算技术、数据加工模型等全领域的技术研发。鉴于其基础性技术突破和影响，斯坦福大学学者 2021 年提出的“基础模型”（foundation models）概念已被广泛接受（Bommasani, 2021）。其次，生成式人工智能已引发网络浏览器和网络搜索引擎等数字基础设施的全面升级，成为最底层国家数据安全和个人数据隐私保护的核心控制节点，并将引发数字管制政策和数字安全设施的全面升级。再次，生成式人工智能为直接和间接以内容生产和内容综述为目的的经济和社会领域提供了提高生产效率的基础工具，将引发大规模的生产方式变革。但生成式人工智能对教育等不以内容生产为目的的行业的效能提升和行业变革能力会有极大的局限性。

从可知的技术路线分析发现，生成式人工智能采用的人工神经网络技术取得的成就皆属统计曲线拟合，它不同于人类结合时间、地点和因果关系等的推理智能（Pearl, et al., 2018）。如果生成式人工智能所代表的深度学习技术路线是对人类智能问题解决进行可计算性模拟的正确路径，那么其持续的迭代突破将会产生超越内容加工范畴的通用人工智能逃逸效应，即积蓄足够的技术势能后会全面赶超人类智能的奇点并进入通用人工智能，进入相对脱离人类控制的发展轨道。但据目前可知的基础模型工作原理，生成式人工智能的底层技术似乎还停留在内容综述、借助符号表征系统的内容加工和格式转换范畴，尚未进入模拟人类理解力的技术路线，仍属“范畴性错误”（a category mistake）（Bishop, 2021）。目前取得的技术突破是否属于范畴错误瓶颈前的技术性能跃迁有待观察。

二、生成式人工智能的基本争议

生成式人工智能的基本争议本质上属于人机互动引发的人文忧患，本研究从人机互动的技术和人文两个维度解析。其中，技术维度是人工智能系统生命周期的主要环节，主要包含以下向度：数据的产生与保存、数据及数据设备的访问权与控制权、基于数据与算法的预测与决策影响的外显行为、智能人机界面及智能设备等实体人工智能。人文维度即人类借助技术以个体存在、社会交往、国家治理以及人类与生态系统互动等的多层次人文活动，主要包括以下彼此关联的向度：人类个体、人与人互动的群体、以主权国家形式存在的人与人关系体、人与环境及生态系统的互动（苗逢春，2022）。在大面积推广使用该技术前，使用者有必要从其训练数据采集、数据使用、基础技术架构、基于模式识别的内容输出等方面加以分析，研判对

个体、社会和国家的现实威胁和潜在影响。

（一）数据生产力挖掘争议：数据贫穷和数字贫穷恶化

中共中央、国务院（2022）颁布的《关于构建数据基础制度更好发挥数据要素作用的意见》是人类进入数据财产和数据产权保护时代的法律标志。从此视域出发，生成式人工智能对个体和商业数据的免费采集使用并借助基于数据训练的技术产品进行商业谋利，会将原本潜藏的数据生产要素跨国跨行业价值挖掘争议推向前台。

访问和应用高质量数据、随时生产高质量在线数据和转化数据的能力已成为人工智能时代支撑国家经济发展和个体获得数字发展机会的基本条件。故而，缺乏数据访问机会、不具备数据挖掘所需的技术能力和超算能力的国家或不具备数据应用支付能力的个体将处于“数据贫穷”（data poverty）（Marwala, 2023）的境地。生成式人工智能提供商基于免费数据训练基础模型和借助训练成熟的模型提供有偿服务的数据剥削生产方式会加剧数据贫穷的恶化。生成式人工智能对数据生产要素的挖掘依赖于三个必要条件：人工智能架构设计和训练方法的迭代创新、海量数据集和超级计算能力。目前全球仅美国、中国和欧盟或极少数超大型数字技术公司同时具备参与基础模型竞争所需的必备条件，数据贫穷国家在生成式人工智能领域的差距迅速拉大并被排斥在基础模型核心研发圈外。生成式人工智能的跨领域普及加快了人工智能领先国家和公司数据生成和技术迭代的速度，成为加速数字鸿沟恶性循环的底层技术成因。

逆转数据贫穷恶性循环的当务之急，是从国家层面解析和补足转化数据要素生产所需的各层次短板，基于下述“数据贫穷成因分类目录”解构和配给转化数据生产所需的各类生产要素：大数据生产所需的互联网普及率、全民数字素养普及率、数据流量成本可承受性、人工智能创新人才储备和创新激发机制、人工智能芯片及超算能力的可及性和性能、借助本地或国际可信数据训练本地模型的能力等。针对该争议的延伸问题是：如果跨国生成式人工智能提供商从低收入国家搜集使用的数据达到一定规模，是否应通过征收数据使用税等国际立法措施平衡数据生产要素剩余价值的分配机制？在具体立法方面，如何界定和追踪数据要素的使用量、如何计算数据生产剩余价值及其税收标准、如何在鼓励技术创新和保护数据贫穷人口基本利益方面取得平衡等问题，都将是国际数据要素治理的前沿难题。

（二）服务辖区内治理争议：生成式人工智能服务辖区内治理失控

生成式人工智能系统的跨境服务应接受其服务覆盖区域当地政府治理机构的管制，但生成式人工智能基于技术不透明的跨境服务已引发治理领域的多重争议。首先，垄断生成式人工智能系统提供商拒绝向独立学术机构提供基本的透明性资料并接受基本学术评估（Bommasani, 2023）。其次，生成式人工智能的基础性技术多受以美国为主的提供商所在国知识产权保护而不向其服务覆盖的国家开放，导致已有用户所在国家在管制技术系统和应用实践安全性方面面临极大挑战（Lin, 2023）。第三，尽管有专家呼吁暂缓生成式人工智能的研发并谋求与公共治理机制同频共振，但资本驱动的人工智能研发迭代节奏远超各国监管法规的起草速度，对各国治理机构应对相关法律和伦理忧患提出了技术不对等的巨大挑战。

各国生成式人工智能的治理呈现梯度性制度缺失和滞后：1）通用数据隐私保护法尚未形成覆盖全球的完整图谱。截至 2023 年 7 月，全球只有 137 个国家制定并颁布数据隐私保护的法律框架，近三分之一的国家无基本数据隐私保护法（UNCTAD, 2023）。2）整体性国家人工智能战略缺失。《指南》颁布前，约 67 个国家制定了国家人工智能战略规划。《指南》颁布后，卢旺达和多米尼加共和国发布各自的人工智能战略。3）国家人工智能伦理治理框架制定滞后。调研发现，截至 2023 年 7 月，全球仅约 40 个国家制定了针对人工智能伦理治理的相关政策。4）应对生成式人工智能的立法无力。调研发现，截至 2023 年 7 月，针对生成式人工智能技术合成内容作品能否受知识产权保护进行论证并提出明确管理意见的只有中国、美国和欧盟三个国家或经济体。在《指南》发布前，只有中国制定并发布了《生成式人工智能暂行管理办法》。此后，美国政府 2023 年 10 月底发布了《关于安全有保障和可信地开发与使用人工智能的行政命令》

(The White House, 2023)。欧洲议会 2023 年 7 月启动《人工智能法案》的起草和谈判程序，并于 2023 年 12 月通过全球最具有法律绑定效力的人工智能管制法律(European Parliament, 2023)。

(三)预训练数据版权争议：未经许可使用内容训练模型

生成式人工智能在搜集和使用训练用数据集方面存在未经许可使用个体或机构数据以及版权保护内容的巨大争议。

如前所述，生成式人工智能模型主要基于数据爬虫软件从互联网爬取的文本、声音、计算机代码、图像等数据集训练。已有垄断性大模型在爬取数据时大多未事先取得个体和机构的许可，易引发广泛且深刻的知识产权争议和法律纠纷。这种行为已被控违反了包括欧盟《公用数据保护条例》(European Union, 2016)在内的数据保护法，已进入法律诉讼的案例集中在新闻媒体行业。2023 年 10 月底，代表 2200 多个新闻出版个体和组织权益的美国新闻媒体联盟指控 ChatGPT 借助爬虫软件，爬取数以百万计的付费新闻报道和报告作为训练语料，但未征得版权拥有者许可，并通过法律程序要求 Crawl Common 删除非法搜集的内容(Robertson, 2023)。生成式预训练模型未经数据拥有者许可使用网络数据的做法被进一步质疑侵犯了用户的“数据遗忘权”，即数据拥有者有权要求有关产品和平台删除未经许可搜集的数据。但在基础模型研发领域，一旦用户数据被基础模型用作训练转换器，已生成的模型从技术上不存在反学习(unlearning)的可能性，不可能从平台输出中删除基于用户数据的深度学习结果，包括反映数据拥有者观点、语言文化习惯等特征的应答(Zhang, 2023)。

(四)模型架构解释性争议：使用不可解释的模型生成内容输出

生成式人工智能采用的人工神经网络模型一直存在模型架构“黑盒”的缺陷，在人工神经节点的节点数和节点层数、参数定义及其计算方法等方面不可解释，这一争议在生成式人工智能的近期突破中得到放大。尽管生成式人工智能包括算法在内的总体技术路径具有一定的可解释性，但难以解释具体模型尤其是模型的具体参数及其在决定内容输出中的权重。GPT-4 等基础模型，通过数以十亿级的参数及其权重界定复杂学习过程识别模式并决定基于模型识别的应答输出，导致难以解释某一预训练模型为什么生成特定的输出。基础模型的主要迭代路径仍依赖参数数量和模型架构复杂性的增加，其不可解释性问题会更加严重。

基础模型客观存在的技术不可解释性与提供商不愿公开必要技术指标的主观行为相交织，给监管机构和独立研究人员检测模型的有意风险和无意危害造成难以克服的障碍。斯坦福大学针对基础模型的核心技术要素研制了基础模型透明性指数(Bommasani, 2023)。该大学基于该指标体系对十多个主流基础模型的透明性作了综合评估。其中，三个主要垄断性生成式人工智能基础模型的透明性得分见表 1。基础模型的不可解释性和风险的不可检测性导致其产生错误时无法追溯原因且无法通过透明机制评估和防范风险。为此，有专家建议不能将生成式人工智能用于高风险任务。

表 1 生成式人工智能平台透明性得分

垄断性模式例举	模型透明度重点指标		
	Llama2(%)	ChatGPT-4(%)	PaLM 2(%)
数据(数据量、来源分布、数据生成者、数据整理与增强)	40	20	20
人力使用(数据工人的数量、薪酬、权益保护、地域分布等)	29	14	0
计算(超算硬件、算力、电耗水耗、环境影响)	57	14	14
方法(训练的主要阶段和目的、软件架构等)	75	50	75
模型基础(模型的技术类型、参数级别、架构等)	100	50	67
模型的可访问性(模型是否对第三方开放共享)	100	67	33
性能(模型的主要技术性能)	60	10	80
风险(模型存在的有意和无意风险及其评估机制等)	57	57	29
模型风险管控(模型层面的风险管控技术包括效能评估及优化机制)	60	60	40
推广渠道(发布过程、发行渠道等)	71	57	71
使用政策(可接受和不可接受的应用场景，包括用户限制等)	40	80	60
反馈(用户反馈报告机制)	33	33	33
影响(对下游不同行业、不同用户群体和地区等的影响)	14	14%	0
平均得分	57	47	39

（五）基础模型理解力争议：生成式人工智能不理解语义和真实世界

生成式人工智能借助概率对文本上下文进行模式识别，根据句法规则生成文本内容。但因其不理解语言的语义（semantics），容易生成关于事实性、史实性甚至科学性错误输出内容幻象（hallucination）。根据代码托管平台 GitHub（2023）基于英文问答的测算，ChatGPT 系列平台的出错率在 3%～3.5% 区间，Llama 系列平台出错率为 5.1%～5.9%，谷歌 PaLM 平台出错率为 12.1%。基于中文等其他语言的出错率应显著高于这一范围。缺乏坚实基础知识的未成年学生通过与生成式人工智能平台的独立对话开展学习，会将学生置于一种基于不确信内容开展学习的争议境地。这一局限意味着基于目前技术的生成式人工智能不能被用作可靠的教学内容来源。此外，生成式人工智能也不能借助句法理解文本和图像等格式背后的现实世界、物体及其关系、人类和社会关系、人与物体的关系或人与技术的关系的真正意义。迄今为止，人类主要的科学发现方法主要是基于对真实世界的观察、科学实验和科学推理。生成式人工智能主要基于对已有文献的综述生成新内容，除非用户基于自身能动性并借助人工智能辅助发现知识，否则生成式人工智能不能输出新知识。依据目前各主要国家的版权保护法，生成式人工智能生成的内容并不被认可为“知识”。与此关联，现有基础模型为现实世界的具体复杂挑战提供有针对性或创新性的解决方案方面表现不佳（Candelon, 2023），更不能作出社会价值判断。故而，目前生成式人工智能尚不能脱离人类教师成为引导学生复杂知识学习和结构不良问题解决的独立导学系统。

上述技术局限会限制生成式人工智能变革教育的正面支持价值。生成式人工智能的现有技术性能在基础性教育内容提供、高阶思维和复杂问题解决过程导学、价值观引导或育人实践等领域可提供的变革性影响有限。目前，生成式人工智能对教育变革的作用似乎更多体现在通过逆向挑战学习结果和评价方式倒逼教育改革：生成式人工智能提高了内容加工的自动化程度和防伪难度，降低了低阶内容综述和作品制作作为核心学习结果的必要价值。处于低水平思维阶段的基本拼写和句法、文献综述报告、演示文稿制作、低阶艺术作品制作等在形成性评价和低利害性考试中的占比降低，将会倒逼教育系统重新界定学习结果的侧重点和相应的评价方式。

（六）生成性信息污染争议：技术合成内容污染互联网

生成式人工智能输出和传播的内容对互联网的污染体现在以下两方面。

一方面，生成式人工智能存在通过生成和传播不良内容、污染互联网信息的争议。目前基础模型训练均从互联网提取训练用数据集，充斥互联网的有害信息、错误信息、歧视信息、憎恨言论等会被转换成有害信息再次输出并通过互联网二次传播，会对不同年龄学习者造成难以逆转的污染。

另一方面，被机器合成内容污染的互联网会影响后续基础模型的培训。优质的深度学习模型依赖于人类产生的高个性化数据，它通过从人类创造和使用的差异性表达方式中识别和学习高差异化的句法和模式，以维持机器的深度学习进程并生成带有模式差异的输出。生成式人工智能大规模生成的内容经互联网的二次传播导致后续的基础模型不可避免地从其先前生成的内容中学习。在基于技术自身生产的数据开展训练的进程中，已有统计模型的高概率事件会被过度高估、低概率事件会被过度低估，导致训练用数据集中低概率事件的（统计曲线）长尾逐步消失。而训练数据的小概率长尾在提高内容输出的准确性和差异性方面具有重要价值，其消失导致的模型过于强化先前识别的模型会引发模型的性能衰退，主要表现为越来越多地生成与机器合成内容趋同而非拟合现实内容的同质应答，出错率上升，最终可能导致模型坍塌。这对后续的基础模型开发者提取优质互联网训练数据提出更大的挑战（Lutkevich, 2023）。

（七）生成性价值观投射争议：同质化输出内容的价值观和语言文化压制

生成式人工智能输出内容的趋同性价值观投射会压制数字弱势群体和教育领域知识建构的多元性和多元化观点表达。

如果一个字符串在训练用数据集中频繁出现，转换器倾向于在其输出中重复这些字符及其连缀成的语句。ChatGPT 等垄断性基础模型采用的欧

美数据集中表达的共识性观点、主流信念或主流媒体主导性观念等都会被识别为与这些价值观和语言文化习惯拟合的“标准答案”输出，从而形成越是互联网强势价值观和语言文化习惯越会在生成式人工智能输出中得到强化的反馈闭环。如果不辨析生成式人工智能平台所用训练数据集的文化和语言来源，大量盲目采用当前垄断性基础模型，会强化美欧价值观及其文化刻板印象。例如，医学领域对 ChatGPT-3、ChatGPT-4、诗人大模型和 Anthronopic 公司的 Claude 大模型问答结果的大量反复检测发现，这些基础模型在回答有关肺活量、估算肾小球滤过率、皮肤厚度、脑容量等客观医学问题时，均会生成基于黑人和白人种族刻板印象的偏见性答案（Omiye, 2023）。相反，数据贫穷群体包括边缘群体中的在线数字化“足迹”稀少，在基础模型训练数据集中的占比很小，其价值观和语言文化习惯无法被基础模式加工、识别和强化。如果没有突出语言文化多样性的本地模型的强势出现，欧美基础模型的全球垄断会危及土著语言和文化的可持续发展和弱势群体的合法利益（苗逢春，2023）。

另外，过分依赖生成式人工智能寻求“标准答案”或问题解决方案，会导致观点的趋同性，削弱多样性创新观点的建构。波士顿咨询集团针对 750 多名被试的研究发现，借助 ChatGPT-4 寻求创新方案的被试所形成观点的多样性比不使用者低 41%，而且被试收到 ChatGPT-4 提供的建议后多缺乏增加观点多样性的意愿。同时，70% 的被调查者认为长期使用生成式人工智能寻求答案会导致人类创造能力的退化（Candelon, 2023）。

（八）助长违法性深伪争议：助推“更深”的深伪的生成与传播

生成式人工智能极大地降低了生成违法性深伪（deeper deepfakes）内容的技术和成本门槛，提高了识别深伪的技术难度，助推了违法性深伪的合成与传播。

生成式人工智能可支持新闻编辑能力低、零音乐和视频制作的用户获得零基础、零成本制作和发布高仿真深伪内容的能力，包括模仿真人的语言风格生成虚假新闻或网络消息用以传播虚假信息、宣传憎恨言论或诋毁他人，或通过修改和操纵已有图像和视频生成难辨真假的伪造视频非法牟利或达到其他不法目的等。换言之，生成式人工智能也许尚不能为解决人类面临的公益问题提供有效解决方案，但已为别有用心者借助深伪内容产品达到违法目的提供了低成本便捷工具。根据“2023 深伪状态报告”（Home Security Hero, 2023）的统计和分析，借助生成式人工智能，只需一张清晰的面部照片，平均不到 25 分钟就可零成本生成一段 60 秒长的色情深伪视频；由于生成式人工智能的助推，2023 年新增在线深伪视频达近一百万段，是 2019 年的 5.5 倍；在所有深伪视频中，色情类深伪视频占 98%，其中借助女性肖像生成的占 99%。

三、管制建议

为应对生成式人工智能引发的上述基本争议并挖掘其教育潜能，《指南》提出了以人为本的人工智能开发和应用指导原则，并遵循“优先管制、确保包容、引导应用”的逻辑路径，确保合乎伦理、安全可信、公平包容和富有意义的应用。

（一）以人为本的人工智能治理与应用取向

自 2019 年以来，联合国教科文组织一直倡导以人为本的人工智能研发和应用取向，并通过《人工智能伦理问题建议书》（UNESCO, 2022）、《关于人工智能与教育的北京共识》（UNESCO, 2019）、《人工智能与教育：决策者指南》（Miao, 2022）等多份文献对以人为本的人工智能的应用取向进行了系统深入的界定。要义如下：人工智能的应用与治理须以人为本，确保人的基本权利、尊严和文化多样性，并追求人与环境和生态系统协调发展的生物中心主义（bio-centred）发展观；人工智能的开发应以技术服务于人为目的，确保人工智能致力于增强人类进行有效人机协作所需的能力；人工智能系统设计、开发、应用、迭代的全生命周期应以确保人机互动的人类主体能动性为原则，确保人工智能系统及其应用的安全可信性（trustable）、主体和领域适用性（proportional）、可解释性（explainable）、人类可控性（human-controlled）、人类问责（human-accountable）（苗逢春，2022）。

（二）生成式人工智能的治理路线图

《指南》建议制定和实施政府一体化、明确跨

领域和行业的主体责权和义务的协同共治策略。

1. 协商跨国通用数据保护法等国际法规

协商和制定跨国互认的通用数据保护法和跨境治理机制是应对辖域内治理失序的必要国际共治前提。欧盟的《通用数据保护法》(简称《保护法》)为针对跨境数据服务开展辖域内治理提供了国际法律框架的先例。

2. 制定政府一体化的跨部门、跨领域人工智能发展战略和伦理共治机制

政府一体化的人工智能发展战略是保证本国各领域和各部门协同治理人工智能的关键机制。其中,人工智能伦理治理机制需清晰界定国家数据主权、机构和个体数据拥有权与隐私等核心权益、合法的数据要素生产关系、人工智能技术开发和应用的核心伦理原则以及基于公共和他人数据生产的人工智能产品剩余价值分配关系等。

3. 研制针对生成式人工智能的专门管理办法

2023 年 7 月,中国发布了《生成式人工智能暂行管理办法》(简称《管理办法》),是全球首部对生成式人工智能进行服务辖区管制的正式法规。该法规可进一步借鉴《指南》以及欧洲与美国近期的立法举措,修改完善其中的诸多法律和伦理要点。其中与教育应用息息相关的要点有以下方面:

1)对生成式人工智能的安全风险进行技术分类和分级监管。欧洲议会颁布的《人工智能法案》草案将人工智能系统对人类安全和基本权利等的风险分成不可接受的风险(包括引诱未成年对人工智能聊天平台产生依赖的技术、情感识别技术、智力和行为预测人工智能等)、高风险、有限风险和轻微风险四类,并针对不同类型采取禁用(禁止开发和投入市场)、重点监管(教育领域属重点监管领域)、审查监督和行业自律的分级监管措施。

2)未成年人独立使用生成式人工智能聊天的年龄限制。基于聊天预训练模型会生成不适合未成年人的输出等安全隐患,《指南》建议各国将 13 岁设为未成年人独立使用生成式人工智能聊天服务的年龄下限,并考虑 16 岁的更严格年龄限制。最近,生成式人工智能的技术服务方式开始实现从基于平台的聊天服务向手机等个人终端应用软件的延伸,监管部门和成人将更难监督未成年人的独立聊天安全隐患。《管理办法》的修订需深入研判相关风险,明确对成年人独立使用生成式人工智能聊天类软件的多方监管责任。

3)明确其他责任主体的责任和义务。《管理办法》的具体条款主要针对的是生成式人工智能提供商,而美国的“行政命令”则较为全面地对各类应用机构的集体责任和义务作了界定。《管理办法》的修订应明确集中采购和部署生成式人工智能系统的机构在协同审核数据、工具和内容服务合法性等方面的共治职责,并协助开展对用户尤其是弱势群体影响的动态监督和评估;在明确机构治理责任的基础上,考虑个体用户应用相关技术应履行的法律和伦理责任,包括个体对其潜在威胁和自我安全保障的基本意识和技能、合乎法律法规地使用相关工具和生成内容的知识和技能等。

4)对人工智能生成内容的版权识别和应用范畴界定。《管理办法》要求“提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识”,欧盟的《保护法》要求明确标注说明人工智能生成的内容。然而,上述三份最有针对性的法律法规均未对人工智能生成的内容产品的合法使用作出清晰界定。“如何有效限制非法性深伪内容的生成和传播”“付费生成内容产品的用户对内容的拥有权如何界定和如何保护”“借助生成式人工智能文献综述支持人类主导的科研活动与完全违背学术道德或教学纪律的生成式人工智能作弊之间如何划定界限及如何识别”等都是有待深入研判的监管难题。

四、确保生成式人工智能有效教育应用的政策和实践建议

《指南》对确保生成式人工智能有效教育应用的政策和实践提出了详细建议,要点如下:

(一)确保包容、公平、语言与文化多样性

生成式人工智能与包容、公平、语言与文化多样等人本原则的双向关系包含以下政策和实践意蕴。1)包容公平的技术使用权和使用机会是借助人工智能促进教育公平包容的前提:应确保无论何种性别、种族、能力水平、社会经济地位或有无固定居住地的人群都有能包容性地使用人工智能的机会。为此,各国应借鉴前述“数据贫穷成因分类

目录”精准确定缺乏使用人工智能机会的人群及其成因,并采取专项措施补齐短板,缩小数字鸿沟。2)面向不同能力水平和年龄阶段终身学习者的包容:通过专项经费开发和推广有针对性的人工智能技术和工具,满足有特殊需要和不同能力水平的群体和不同年龄段的学习者的终身学习需要。3)确保数据和技术去偏:生成式人工智能评估体系应该重点考核数据来源和数据预处理、模型设计和输出内容中的性别偏见、特殊群体刻板印象、边缘群体歧视、憎恨言论等。4)倡导开发具有语言文化多样的基础模型:制定和实施确保生成式人工智能语言文化多样性的指标体系,确保预训练数据集、模型架构和训练方法对少数或土著语言文化的包容性,严禁提供商有意或无意从预训练数据中剔除少数族群的语言和添加带有语言文化偏见的数据过滤或输出的后处理技术。

(二)保护主体能动性

由于生成式人工智能性能日益精细并在一定程度上部分替代人类的初级内容加工活动,其教育应用存在压制人类能动性的可能。借助生成式人工智能工具撰写论文和提交基础艺术作品带来的便利,会诱使学生对借助工具加工外部内容形成依赖。对外部创作的长期依赖会使学生失去锻炼心智和形成基础知识、基本技能的机会。为此,保护和增强人类的能动性应是设计和采用任何人工智能技术的核心原则,应坚守生成式人工智能可被用于挑战和拓展人类的思维,但决不能用来篡越人类思维活动的底线原则。《指南》建议从以下方面界定人机互动的主体性,保护生成式人工智能教育应用的师生能动性:1)明确告知生成式人工智能会搜集和使用的学生数据类型、数据将被如何使用以及相关数据应用会对其教育和社会生活的影响;2)保护学生成长和学习的内部动机,强化人类在基于日益复杂的人工智能系统开展教和学中的决策和行为自主性;3)防止生成式人工智能的使用剥夺学生通过观察现实世界、实证方法(如实验、与他人的讨论等)和逻辑推理发展其认知能力和社会技能的机会;4)在学习活动中确保学生有足够的社会互动和接触人类创作作品的机会,防止学生过度依赖生成式人工智能或成瘾;5)在审核并决定是否大规模采纳生成式人工智能工具前,充分咨询研究人员、教师和学生的意见;6)鼓励学生和教师批判和质疑生成式人工智能背后采用的技术方法、输出内容的准确性、隐含的价值观以及对教学方法和过程的潜在影响等;7)师生在借助人工智能作决策和选择时,应避免将人类的决策责任让渡给生成式人工智能系统。

(三)审核监控教育生成式人工智能工具与开发本地适用性教育基础模型

教育生成式人工智能工具的开发和部署在遵循“设计伦理”指导原则的基础上,应从人工智能系统的全生命周期出发,避免具有潜在技术和伦理风险或不具备对教学适用性的人工智能技术对师生和各类教育主体关系的影响。针对已有生成式人工智能的教育应用,《指南》建议通过以下机制开展审核、准入和全程监控:在准入审查中强化伦理检测,考核生成式人工智能系统是否有去除偏见尤其是性别偏见的技术和机制、是否采用代表语言文化多样性的训练数据集;在给予准入权限前,确保被审查的系统不会对师生产生可预测的伤害,提高教育有效性、针对不同年龄和能力学生的适用性并符合教育机构确认的教学原则(如适用于相关的知识技能类型、预期的学习结果和价值观培养目标等);采取有效措施解决数据使用和服务许可授权等难题。例如,针对未成年人和残障人士等不具备完全刑事能力的主体,如果其被告知数据隐私和安全等风险的前提下被授权使用数据和接受服务的难题等;审查生成式人工智能输出内容中是否含有深伪图像、虚假新闻或憎恨言论等不良或非法信息;结合当地的环境影响评估结果,分析生成式人工智能系统教育应用的环境成本,尤其是模型训练的电耗和水耗等因素。

现有垄断性大模型对欧美之外其他国家的语言文化适用性较低,在保障对合法国际竞争的前提下,应采取积极自主的开发策略提高教育生成式人工智能的本地适用性;在鼓励版权自主、安全可控的本国基础模型开发基础上,支持基于本地价值观、语言文化多样性和本国课程标准的教育生成式预训练转换器(EdGPT)或教育大模型的研制、试用与迭代;通过激励机制鼓励开发基于本国基础模型的面向探究性学习和多样化学习选择等需求的生成式人工智能教育平台和应用插件,培育基于本国

语言文化多样性和课程标准、基础模型与中下游应用软件同步的教育生成式人工智能生态系统。

(四)培养师生的人工智能能力

培养学生尤其是中小学生的人工智能能力对确保学生安全、符合伦理和有意义地应用人工智能至关重要。截至2022年中,全球只有约15个国家制定并通过国家认可的中小学人工智能课程(UNESCO, 2022)。随着生成式人工智能的迅速推广应用,培养面向所有人的基础性人工智能素养(AI Literacy)的需求更加迫切。联合国教科文组织正在研制的“中小学生人工智能能力框架”,从“能力表现”和“能力层面”两个维度界定可通过课堂教学和课外课程结合的方式培养的中小学生人工智能能力。能力表现维度包括“人工智能观念”“人工智能伦理”“人工智能底层技术与应用”“人工智能系统设计”;能力层面维度则从对相关知识、技能和情感态度的“理解”“应用”和“创造”三个方面界定学生的能力表现(见表2)。

与此同时,联合国教科文组织在组织研制中小学教师人工智能能力框架,旨在引导各国制定相关标准和教师培训课程以支持教师做好合理应用生成式人工智能的能力准备(见表3)。该框架倡议从以下能力层面界定教师的能力:人本人工智能观念、人工智能伦理、人工智能基础与应用、人工智能与教学整合和人工智能支持教师专业发展,并建议从“获取”“深化”和“创造”三个水平划分不同能力背景的教师通过培训可以达到的能力表现。其中,“获取”水平是预期所有教师经过培训均能达到的能力表现,包括缺乏人工智能知识技能准备的教师以及技术条件贫乏地区的教师;“深化”水平是预期具有中等先前知识基础的骨干教师经过培训可以达到的能力表现;“创造”水平针对在人工智能教育应用方面有深厚的知识技能储备的教师,预期经过培训后可以达到专家的能力表现。

表2 联合国教科文组织中小学生人工智能能力框架(研制中)

能力表现	能力层面		
	理解	应用	创造
人本人工智能观念	人机交互中的主体能动性	人机交互中的人类问责	人工智能时代的公民素养
人工智能伦理	批判性反思	安全负责的人工智能应用	设计伦理
人工智能底层技术与应用	人工智能基础	应用技能	借助人工智能的创造
人工智能系统设计	问题界定与抽象	架构设计	系统实现与迭代

(五)倡导创造性教学应用并强调多元观点建构及观点的多元化表达

教育担负着维持和促进文明延续和持续繁荣的历史使命。实现这一教育使命的支柱性原则是确保教育过程中语言文化的多样化、鼓励个性化观点的建构和多样化表达。生成式人工智能通过

表3 联合国教科文组织中小学教师人工智能能力框架(研制中)

能力层面	能力进阶		
	获取	深化	创造
人本人工智能观念	教师基于对人的基本权利、社会正义和基本价值的理解,能意识到人工智能给教育带来的机遇与挑战	教师能在其教育实践中安全、负责地整合应用人工智能工具,充分考虑国家和当地的相关政策,并能考虑学生的安全性、隐私和基本权利	教师能批判地评价、反思人工智能的教育应用并能适度推动相关工作的开展;能展示出对人工智能教育应用社会影响的深刻理解,有参与应对人工智能教育挑战所需变革行动的积极态度
人工智能伦理	教师能意识到并理解遵守人工智能伦理基本原则的重要性,认识到人工智能开发和应用的关键环节应由人类主导的根本特征	教师能基于对人工智能工具伦理影响的理解批判性地评价和应用相关工具,能坚守和宣传公平、包容和多样性的核心价值,能理解人工智能创建者的设计,取舍既可能支持也可能损害人工智能的伦理应用	教师能通过自身的模范作用引领对人工智能工具的合乎伦理的应用,能在人工智能应用中倡导关爱和共情等伦理观念,能适当参与所在机构或社会伦理监管机制的制定
人工智能基础与应用	教师意识到人工智能基本概念的重要性,并能展现出对日常使用的人工智能基本工作过程的初步理解	教师能基于具体的教育背景熟练地辨别、评价、选择和应用人工智能工具	教师能熟练地通过改造、修改开源工具或组合多种工具,设计适合其特定需要的人工智能解决方案
人工智能与教学整合	教师可以识别特定人工智能系统为教学带来的益处,展现出将具体人工智能系统融入学科教学所需的教学策略	教师能结合教学法原则指导人工智能应用过程并能确保以人为本教学活动的开展	教师能批判性地评估教学实践中人工智能的作用,能设计人工智能支持的创新性教学方法
人工智能支持教师专业发展	教师能意识到人工智能在支持自身持续专业发展方面的潜能,并具备借助人工智能积极开展终身专业学习的动机	教师能应用合适的人工智能工具参与专业发展社团,并开展合作以应对持续变化的环境所需的专业发展	教师能批判性地改进、合成或修改人工智能工具,以满足自身或所在团体专业发展转型的需要

复制或强化训练数据拥有者的世界观和语言文化观念，压制多样性观点的形成和多元表达，并对教育的文化多样性和观点多元化使命造成直接威胁。为此，借助生成式人工智能查询或深化某个(些)主题教学的根本前提是：无论围绕任何主题开展人机互动，均不能将生成式人工智能作为权威的知识来源，应引导教师和学生坚守对生成式人工智能输出内容进行批判性思维的正确定位，即明确生成式人工智能可用于快速查询信息、支持文献检索和格式转换，但会含有不可靠内容的信息来源。在具体教学中，教师应基于主体适用原则设计教学活动，激励和辅助学生评价和批判其输出内容对价值观和语言文化观念的投射，并借助生成式人工智能在文献综述和数据加工方面的优势支持探究性学习；与此同时，为学生提供足够的不依赖人工智能工具的试错学习机会、实证性实验和对真实世界的观察机会。

(六)跨领域跨学科审视生成式人工智能对教育的长远影响

尽管本研究对目前可知的生成式人工智能工作过程进行了概括和追溯，但面对迅速进化的基础模式和快速裂变的中下游应用软件，以及人类教育活动的互动场景千差万别并充满不确定性，不能仅基于现有技术缺陷或单纯从消极怀疑论的视角出发短视地分析其教育影响。《指南》建议教育决策者仍应与人工智能研究者和提供商、教育教学理论研究者、认知神经科学等学科研究人员以及教师、学生和家长等合作，跨领域、跨学科评判生成式人工智能对知识生产和学习过程、版权与科学研究、课程与评价、人类协作和社会动态发展等领域的深远影响，并以此为基础审慎地反思迅速迭代的人工智能技术对课程框架、教学目标界定和考试评价方式的影响，从而作出相应的系统调适。

面对一项在内容加工功能上出现代际跃迁的人工智能技术，论证其教育潜能的逻辑起点不应限于关于其技术性能的商业宣传，也不应始于其在内容创作等商业领域的生产力提升表现，仍应坚守以人为本的基本原则。以此为出发点，教育者辩证分析生成式人工智能系统全生命周期的关键技术环节可能引发的根本争议，系统考证和应对其对公平、包容、价值观培养、语言文化多样性等教育核心价值的影响，唯此才能避免作为教育主体和教育服务对象的人类个体和人类群体成为商业驱动的人工智能技术的实验品和仆从者。

［参考文献］

[1] Bishop, J. M.(2021). Artificial Intelligence is stupid and causal reasoning won't fix it[J]. Frontier in Psychology, 2021(11): 1-18.

[2] Bommasani, R., Hudson, D. A., Adeli, E., *et al*(2021). On the opportunities and risks of foundation models[DB/OL]. [2023-10-16]. https://crfm.stanford.edu/report.html.

[3] Bommasani, R., Klyman, K., Longpre, S. et al.(2023). The foundation model transparency index. [DB/OL]. [2023-10-10]. https://arxiv.org/abs/2310.12941.

[4] Candelon, F., Krayer, L., Rajendran, S. et al.(2023). How people can create and destroy value with generative AI[DB/OL]. [2023-10-12]. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai.

[5] European Parliament(2023). MEPs ready to negotiate first-ever rules for safe and transparent AI. [DB/OL]. [2023-10-18]. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

[6] European Union(2016). General Data Protection Regulation [DB/OL]. [2023-08-09]. http://data.europa.eu/eli/reg/2016/679/oj.

[7] GitHub(2023). Hallucination Leaderboard. [DB/OL]. [2023-11-08]. https://github.com/vectara/hallucination-leaderboard.

[8] Home Security Hero(2023). State of deepfakes: Realities, threats, and impact [DB/OL]. [2023-11-11]. https://www.homesecurityheroes.com/state-of-deepfakes/#key-findings.

[9] Lin, B.(2023). AI is generating security risks faster than companies can keep pp. [DB/OL]. [2023-08-25]. https://www.wsj.com/articles/ai-is-generating-security-risks-faster-thancompanies-can-keep-up-a2bdedd4.

[10] Lutkevich, B.(2023). Model collapse explained: How synthetic training data breaks AI [DB/OL]. [2023-10-28]. https://www.techtarget.com/whatis/feature/Model-collapse-explained-How-synthetic-training-data-breaks-AI.

[11] Marwala, T.(2023). Algorithm Bias: Synthetic Data Should Be Option of Last Resort When Training AI Systems. [DB/OL]. [2023-07-31]. https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems.

[12] Miao, F.(2022). AI and education: Guidance for policy-makers[M]. Paris, UNESCO.

[13] Miao, F. et al(2023). Guidance on generative AI in education and research[M]. Paris: UNESCO 2023.

[14] 苗逢春(2022). 教育人工智能伦理的解析与治理：人工智能伦理问题建议书的教育解读 [J]. 中国电化教育, 2022 (6): 22-36.

[15] 苗逢春(2023). 数字文明变局中的教育数字化转型 [J]. 电化教育研究, (2): 47-63, 91.

[16] Omiye, J. A., Lester, J. C., Spichak, S. et al.(2023). Large language models propagate race-based medicine[J]. Digital Medicine,

2023(6): 195: 1-4.

[17] Pearl, J., & Mackenzie, D.(2018). The book of why: The new science of cause and effect [M]. NY: Basic Books.

[18] Robertson, K.(2023). News group says A. I. Chatbots heavily rely on news content [EB/OL]. The New York Times. https://www.nytimes.com/2023/10/31/business/media/news-artificial-intelligence-chatbots.html.

[19] Stanford University(2019). The 2019 AI Index Report [DB/OL]. [2023-1018]. https://hai.stanford.edu/ai-index-2019.

[20] The White House(2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence [OL] The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[21] Thompson, A. D.(2023). Contents of GPT-3 & the pile v1[EB/OL]. [2023-1018]. https://lifearchitect.ai/models/#gpt-3-top-10.

[22] Touvron H, Martin, L, Stone, K et al.(2023) Llama 2: Open foundation and fine-tuned chat models[DB/OL] [2023-10-16]. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models.

[23] UNCTAD(2023). Data protection and privacy legislation worldwide [DB/OL]. [2023-07-18]. https://unctad.org/page/data-protection-and-privacy-legislation-worldwide.

[24] UNESCO(2019). Beijing consensus on artificial intelligence and education[S]. Paris, UNESCO.

[25] UNESCO(2022). Recommendation on the Ethics of Artificial Intelligence[S]. Paris, UNESCO.

[26] Zhang, D., Finckenberg-Broman, P., Hoang, T., et al(2023). Right to be forgotten in the era of large language models: Implications, challenges, and solutions [DB/OL]. arXiv: 2307.03941v3 [cs.CY] 22 Sep 2023.

[27] 中共中央、国务院(2022). 关于构建数据基础制度更好发挥数据要素作用的意见 [EB/OL]. [2023-10-23]. https://www.gov.cn/zhengce/2022-12/19/content_5732695.htm.

(编辑: 李学书)

Generative AI and its Uses in Education: Foundational Controversies and Responding Strategies

MIAO Fengchun[1,2]

(1. *National Engineering Laboratory for Cyberlearning and Intelligent Technology*, *Faculty of Education*, *Beijing Normal University*, *Beijing*, 100875, *China*;
2. *Headquarters of UNESCO*, *Paris*, 75007)

Abstract: *The paper is the second edition of a serials of interpretative articles on the Guidance on generative AI in education and research released by UNESCO, and the main aim of this edition is to examine the controversies around generative AI and its use in education. Based on an analysis of how it works, the development and deployment of generative AI have triggered the following controversies: worsening digital poverties, outpacing domestic governance, the collection of pre-training data without consent, the unexplainability of the architecture, the foundational models can't under the real world, the outputs of generative AI is polluting the internet, the outputs are projecting dominant values, amplifying the creation and dissemination of illegal deepfakes. Generative AI poses direct and most profound threats on the core values of education including inclusion, equity, cultural and linguistic diversities, pluralism in knowledge construction and expression, and therefore, those core values should be upheld as the logic base when examining the proportionality of generative AI to service purposes of education. Policy-makers and practitioners should follow the logic roadmap of "prioritizing regulation, ensuring inclusion, and guiding applications" to enhance the system-wide governance regulations in order to ensure trustable, self-automatic, and locally relevant eco-systems of generative AI for education, followed by developing capacities and guidance to steer proper uses of generative AI in education.*

Key words: *generative AI*; *content processing across symbolic representations of human thinking*; *sources of pre-training datasets*; *equity, inclusion, cultural and linguistic diversities*