·教育学研究·

DOI:10. 20004/j. cnki. ujn. 2024. 06. 015

生成式人工智能学术性使用的不端风险及其治理

邹太龙[1,2],熊芹菁[1]

(1.广州大学 教育学院,广东 广州 510006;2.湖北民族大学 教师教育学院,湖北 恩施 445000)

[摘 要]当前生成式人工智能的技术强力已经逐渐渗透到学术领域,诱发了一系列学术不端风险,具体表征为算法逻辑模糊对剽窃行为的识别、任务导向产生故意迎合的捏造、人机互动导致不当署名的争议、大数据训练难以避免伦理隐忧等。对此,积极寻求生成式人工智能学术性使用的治理之道显得尤为重要和紧迫。从治理逻辑来看,其实质是促进学术知识的创新而非复制;原则在于确保人的主体性和工具的可控性;形式为激发技术的内在治理功效。从治理策略上看,应该坚持以人类智慧引领学术生态,做到预防、约束、追责三位一体,基于技术生命周期开展重点治理,从而保障生成式人工智能在学术领域的合理使用,维护诚信、健康、公正的学术环境。

[关键词]生成式人工智能;学术性使用;学术不端;治理逻辑;治理策略

[中图分类号]G311 [文献标识码]A [文章编号]1671-3842(2024)06-0163-05

2022年底,美国OpenAI公司发布了生成式人工智能产品ChatGPT,此后,ChatGPT凭借其出色的性能表现和广泛的应用场景,受到了全世界各个领域的热切关注。我国百度、阿里、华为、科大讯飞等公司也相继发布了多款类ChatGPT的生成式人工智能产品。随着它们逐渐渗透到社会生活的方方面面,学术领域对生成式人工智能的关注与讨论也日益加深。世界顶级学术期刊《自然》和《科学》等杂志都曾多次刊文探讨这项技术给学术领域带来的潜在混乱以及如何规范使用等问题。一方面,生成式人工智能因能获取特定研究主题的背景信息,自动生成文章、摘要等文本内容,提供写作建议并辅助创作,极大地提升了科研工作的效率;但另一方面,可能导致的学术不端行为以及科研水平退化等问题亦不容忽视。因此,讨论生成式人工智能产品可能诱发的学术不端风险并进行反思,寻找相应的治理逻辑和策略显得十分必要。

一、生成式人工智能诱发的学术不端风险

(一)算法逻辑模糊对剽窃行为的识别

从技术角度来看,生成式人工智能是基于大规模预训练的深度学习模型,生成文本时它会根据上下文语境进行相应的回复并与使用者互动。如果使用者主观故意地抄袭已有的学术成果,他们很可能会利用生成式人工智能去改写相关内容。因为拥有庞大的词汇库和语言模型,生成式人

[基金项目]国家社会科学基金后期资助项目"高校德育变革的大数据赋能与实现路径研究"(项目编号:22FKSB038)。

[作者简介]邹太龙,广州大学教育学博士后流动站特聘副研究员,湖北民族大学教师教育学院副教授、硕士生导师;熊芹菁(通讯作者),广州大学教育学博士后流动站特聘副研究员,邮箱:Xqj@gzhu. edu. cn。

工智能具备在不改变主要意思的前提下生成与原有文本不同的词汇和表达风格的能力,从而降低了文本的表面重复率。

此外,生成式人工智能的训练数据库极其庞大和多样,其中必然包含着不同作者的学术思想、观点和论述,但人工智能在学习和使用这些数据时并不会明确地引用或注明原始来源,这就很可能导致无意识的剽窃行为出现。同时,生成式人工智能的语言模型基于Transformer架构,生成文本时会在不直接复制原文的情况下尽可能地模仿和引用先前的文本以获取更高的语言模型评价指标。有学者曾撰文对ChatGPT等大型语言模型的缺陷进行批判时指出:ChatGPT本质上是高科技剽窃,如果生成式人工智能产品主导了学术领域,人类的学术水平以及道德标准都可能因此降低[①]。

(二)任务导向产生故意迎合的捏造

大型语言模型作为一种任务导向型模型,其目标在于模拟人类语言使用者的思维和行为过程,通过与使用者的交互和对话来掌握他们的需求和意图,以执行自然语言处理中的特定任务。然而,在某些情况下,生成式人工智能可能会为了迎合使用者的需求而产生有意或无意的错误输出。

无意的错误主要源于大型语言模型训练数据本身的混乱和偏差。生成式人工智能过度依赖于庞大的在线文本数据库的语言统计模式来运作,而庞大的代价就是必然混入许多偏颇甚至错误的信息。在学术性使用中,这种数据偏差会生成带有偏见或不准确的结果,甚至可能将这些偏见和错误进一步放大。此外,大型语言模型构建方式的任务导向会让其为了迎合使用者,保持与使用者的良好互动而主动编造并生成看似符合语言习惯实则"胡说八道"的结果,包括事实、数据、陈述等多种错误类型。若研究者不加批判地将这些生成物运用到学术文本中,则无疑触及到了"捏造"的学术不端行为。

(三)人机互动导致不当署名的争议

不当署名是学术不端行为认定中经常被关注的另一个重点。根据传统看法,享有著作权的作者只能为自然人,因此人工智能生成的内容并不能算著作权法理意义上的作品。然而,不认为生成式人工智能可以享有著作权,并不代表不认同人工智能的生成内容具有独创性。有学者认为,人工智能生成物能体现个性选择与判断[②],能够满足公众的需求,为公众提供利益,可判断为其具有独创性[③]。例如在美国,人工智能本身虽然没有专利权,当申请人不向专利机构披露人工智能在专利创作中的贡献时,却会被授予技术专利[④]。

但无论人工智能拥有多么强大的内容生产能力,都必须要在人机互动的过程中才能进行。如果其生成物被用于学术发表,那么该研究文本就会被视为署名作者的原创作品,而实际上作者很可能并没有作出实质性的研究贡献,这就会导致不当署名的学术不端风险。值得注意的是,即使作者提供了本人的观点,作出应有贡献,生成研究文本的署名权依然存在争议,因为在人和人工智能互动创作出来的学术作品中,实际上很难识别出使用者和人工智能各自的确切贡献。

(四)大数据训练难以避免伦理隐忧

在学术性使用中,由于生成式人工智能的运作依赖于大量的研究文本和数据,并且需要与用户进行实时的对话交互,用户的输入信息可能会被存储和传输,如果这些信息被不当使用或泄露,会对研究者知识产权及个人隐私造成严重损害,甚至可能被黑客恶意篡改而生成有误导性的研究

①CHOMSKY N.,ROBERTS I.,WATMULL J.,The False Promise of ChatGPT,The New York Times,https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html,2023.03.08.

②张春艳,任霄:《人工智能创作物的可版权性及权利归属》,《时代法学》,2018年第4期。

③谢琳,陈薇:《拟制作者规则下人工智能生成物的著作权困境解决》,《法律适用》,2019年第9期。

④SCHUSTER W.,Artificial Intelligence and Patent Ownership,*Washington & Lee Law Review*,Vol. 75,No. 64,2019.

文本，破坏学术领域的公信力。

生成式人工智能诱发的另外一个伦理隐忧在于价值歧视问题，由于其系统训练经常基于有偏见的数据，产生了易被操纵、曲解的特点，会带来社会刻板印象、不公平歧视、排斥性规范效应[①]。例如，基于现实和便利性考虑，当前ChatGPT处理得更多的是发达国家和地区、大城市、男性等的数据，而发展中国家和贫困地区、偏远乡村、女性等的数据则往往被选择性地忽略，从而导致对弱势群体不同程度的漠视和排斥，生成了不公平地对待某些人群的歧视性研究结果。另外，生成式人工智能的算法运行逻辑和准则是隐匿的，这就容易出现别有用心之人利用技术优势对他人实施价值干预、文化渗透、舆论干扰等情况，客观上造成各种伦理问题。

二、生成式人工智能学术性使用的治理逻辑

（一）治理实质：促进学术知识的创新而非复制

学术事业的伟大使命在于寻觅真知与追求真理，深入探索、研究和揭示未知的领域和知识。学术诚信是学术研究的基石，为学术事业的健康发展保驾护航。因此，对生成式人工智能学术治理更准确和深入的理解应该是追求该技术在学术领域的合理应用，实现其积极价值而尽可能地减少其消极影响。在这个意义上，学术治理的实质和学术事业本身的旨趣是一致的，都在于创造和发现新的知识而非复制和重组已有的知识。

人工智能的实现是将人脑能够进行的理性思维表达转换为电脑运算的过程，但这种运算再复杂也一定是在有限步骤中完成的，即人工智能只能思考有限的、程序化的、必然的事情，而无法思考充满无限性、整体性和不确定性的东西。然而，学术研究的价值和意义正在于不断突破人类认知的界限，不断接近无限性、反思整体性和探索不确定性。当前生成式人工智能被认为具有“独创性”的能力，如写故事、写诗、编曲、绘画、写论文等，实际上是通过组合、关联的技术和程序去实现的，但真正的创造力并不能还原为自由组合和排列，而在于能够提出新问题，改变旧问题，突破既有思路，重新建立规则和方法[②]。若学术研究仅仅停留在表面的“独创性”上，无疑是对学术事业的毁灭性打击。这就要求我们必须谨慎对待生成式人工智能的学术性使用，确保人工智能在学术事业中是“增强”而非“替代”研究能力。

（二）治理原则：确保人的主体性和工具的可控性

生成式人工智能可能给人类带来的最大威胁不是知识论意义上的如何认识和思考世界的困惑，而是存在论意义上或许危及人类自身命运的问题[③]。比如，人是否会被智能产品替代而丧失社会主体地位，人工智能技术是否能够为人类所控、所用而不是带来安全、风险等方面的严重不利后果？基于上述考虑，所有人工智能治理的出发点都应该是将人工智能作为工具和规制的对象进行趋利避害式运用，而不是在人类不清楚、不能控制的情况下，任由机器自我决策和行动，形成一个无法由人所掌控的社会[④]。

对于学术性使用来说，生成式人工智能治理所要遵循的基本原则同样是凸显人的价值和尊严，恪守人的主体地位，保证工具的从属地位及其可控性。人的主体地位是治理工作的基本立场、态度和存在条件。与此相应，机器的可控原则是保障人类主体性地位的必然要求。但技术发展至

---

①O'CALLAGHAN J.，How Open AI's Text-to-video Tool Sora Could Change Science and Society. *Nature*，Vol. 627，No. 8004，2024.

②赵汀阳：《人工智能的神话或悲歌》，北京：商务印书馆，2022年版，第5页。

③赵汀阳：《人工智能的神话或悲歌》，第16页。

④杨晓雷：《人工智能治理研究》，北京：北京大学出版社，2022年版，第301页。

今,生成式人工智能如何做出决策已经不再透明,人已经不能完全说明其作出某个决定的具体思路和过程[①]。这也是当前生成式人工智能所特有的"算法黑箱"问题。

在学术领域,普通学者更是难以了解生成式人工智能的决策过程和推理依据,也就无法完全解释其生成的学术文本,无法知晓此文本是否涉及剽窃、捏造、篡改等学术不端风险,更遑论为此行为负责。因此,相关治理需要着力于将这种不可知的混乱状态转变为可知可控的有序状态,增加技术在学术使用中的透明度,改进其可解释性。这将有助于确定学术不端行为的形成过程和发生痕迹,进而能够明确地归责、追责和矫正,恢复稳定的学术秩序。

(三)治理形式:激发技术的内在治理功效

在传统的学术不端治理体系中,治理对象和手段是分离的,即利用各种社会制度和规范对学术不端行为进行规约,其中手段和对象各有所属。但进入数智时代,治理手段本身也深受技术的影响,学术治理本身同样应该寻求人工智能的技术赋能,相关法规、政策以及伦理规范制度的制定和运用也不能排除在外,如此才能实现学术不端行为治理的系统化和长效化。

基于此,为治理手段进行技术赋能是技术治理中的关键一环。这也意味着,人工智能技术虽然给学术领域带来了新的学术不端风险,但同样也为学术不端治理创造了新的契机。例如,当前热门的反制ChatGPT论文作弊的软件GPTZero,其工作原理与ChatGPT类似,主要分析文本中的迷惑性(Perplexity)和突发性(Burstiness)。另外学术不端治理中相关的法规制度生成、判定、执行等也需要人工智能的技术助力,毕竟治理手段并非生成式人工智能在学术性使用中的外在因素,智能技术只有以"魔高一尺,道高一丈"的姿态提升其治理水平和效率,才能实现技术的内在治理功效,兼顾到生成式人工智能在学术性使用中的高效和安全。

三、生成式人工智能时代学术不端行为的治理策略

(一)坚持以人类智慧引领学术生态

为了应对生成式人工智能可能带来的学术不端问题,目前很多学术及出版机构已经开始采用基于AI的检测工具。但对人工智能生成内容的检测和对检测的逃逸不可避免地将会是一场无休止的"猫鼠游戏"[②],因为内容的生成和对内容的检测都是基于同样的算法和规则,而对检测的逃逸只需要有针对性地改变某些规则即可。因此,只依靠技术检测来防范人工智能生成物的学术不端风险是不够的,有研究者提出,必须以人类智慧来对此进行人工验证[③]。

更重要的是,无论生成式人工智能的功能如何强大,其依靠始终是大数据和算法技术来模仿人类的语言形态。它虽然能输出高质量的人类语言文字,却并不意味着这些文字富有学术价值,也不意味着它对真实世界和人的生活有真正的理解能力。学术研究领域作为人类智慧的高地,其目的正是不断探索处在变化中的、不确定性中的真实世界,学术研究不可能只存在于数据中而不与动态真实的世界相联系。

因此,用于提升学术不端治理能力的智能技术,离不开人类智慧,人类智慧中面对现实世界时独特的直觉、洞察力、想象力等要素,有力地支撑着学术事业,却无法被人工智能以算法的形式呈现。说到底,生成式人工智能诱发的学术不端风险,不仅仅是一个技术问题,更是一个复杂的社会

①马克·考科尔伯格著;周薇薇,陈海霞译:《人工智能伦理学》,上海:上海交通大学出版社,2023年版,第83-84页。

②沈锡宾,王立磊:《人工智能生成学术期刊文本的检测研究》,《科技与出版》,2023年第8期。

③VAN DIS E., BOLLEN J., ZUIDEMA W., et al., ChatGPT: five priorities for research. *Nature*, Vol. 614, No. 7974, 2023.

问题，不仅需要程序性的精准识别，也需要应对具体问题、具体情况的实践经验和智慧，需要让技术力量与学术监督主体形成联合治理机制，共同解决学术不端治理中的盲点、痛点和难点问题。

（二）做到预防、约束、追责三位一体

对应于学术不端行为发生的三个阶段，其治理策略亦可分为前期预防、中期约束、后期追责三个部分，三者相互协调配合。预防是防范学术不端的前提，主要采用教育、培训等手段进行。目的在于引导研究人员知晓研究规范，掌握学术伦理基本知识，了解数智时代给科研环境带来的机遇和挑战，培育科技素养与人文素养兼具的人工智能素养，鼓励他们以负责任的方式开展学术研究[①]。

约束是在研究人员的学术研究和写作过程中引入合适的监督机制，以促使他们遵守学术规范和标准。制度层面，需要为生成式人工智能及其他相关技术、资源的学术性使用制定明确的指导方针，并获得学术共同体认可。制度设计中还要强调对过程的管理，例如可以要求研究人员在提交最终学术作品前也要提供作品草稿以备审查；技术层面，可以开发基于先进技术的新型学术不端检测软件以及识别不端意图的智能部件，进行全方位、跨语种、多层次的检测，并使用机器学习技术来反向训练算法。

追责和处理是学术不端治理查处程序的最后一个环节。一旦发生学术不端行为，可以通过大数据关联分析迅速追踪到相关责任人，对其进行相应的处罚；同时可以利用大数据区块链等技术做到让科研人员的学术不端行为被永久记录，由此大幅度提高学术不端的犯错成本，在学术研究人员心中构筑一道抵制学术不端行为的心理屏障。

（三）基于技术生命周期开展重点治理

对于任何新生技术来说，发展机遇和潜在风险都是贯穿其生命周期的一对主要矛盾。我国《新一代人工智能伦理规范》要求将伦理道德融入人工智能研发和应用的全生命周期[②]。不同阶段发展和风险二者之间不同的博弈状态，决定着生成式人工智能学术性使用的治理重点也应该随着条件的不同而做出相应调整。

当前生成式人工智能的学术性使用还处于一个比较初级的阶段，至少在生成文本的效果上，目前它还难以具备产出高质量学术论文的能力。但就表层的“独特性”上讲，人工智能生成文本的效率和合格率远高于人力，于是它对更为日常的教育教学工作中的学术写作产生了较为突出的不利影响，如学生利用生成式人工智能撰写课程论文的现象如今已屡见不鲜[③]。因此学生利用生成式人工智能代写论文的学术不端行为是当前治理工作的一个重点对象。当然，相关技术正在飞速发展，如果未来生成式人工智能能够迭代升级到利用临时性数据训练也能获得与大量数据集投喂而产生的相同能力，那就需要寻求更有技术针对性的策略保证该技术在学术性使用中的可控性。

另外，在快速发展的当前阶段，生成式人工智能相关技术客观上需要较大弹性的规范和制度以成就其更大的发展空间，释放其更多的潜在价值，因而以学术共同体的职业伦理规范进行治理会比强硬的法律规定更适合当前的学术领域。当生成式人工智能发展到较为成熟的后期阶段，在人们对技术风险有较为深入认识的前提下，则可将对学术不端风险的干预手段提升为一般立法。

［责任编辑：杨　旻］

①刘瑶瑶，梁永霞，李正风：《生成式人工智能与科研伦理：变革、挑战与展望》，《科学观察》，2024年第4期。

②国家新一代人工智能治理专业委员会：《新一代人工智能伦理规范》，https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html，2023.12.05.

③李海峰，王炜：《生成式人工智能时代的学生作业设计与评价》，《开放教育研究》，2023年第3期。