

生成式人工智能应用于学术研究的风险及其预防机制

周 莎, 张 尚

(陕西师范大学, 陕西 西安 710062)

摘要:生成式人工智能是利用计算机模拟人类的智能行为所开发的相应实践理论与技术,其在各个领域的爆发应用为未来人类的发展提供了无限的想象空间,与此同时,虚假信息、学术伦理、劳动替代等问题也引发全球热议。实践地看,生成式人工智能在当前的学术研究中可能存在人文性缺失、科学性缺失、依赖性成瘾、原创性与批判性缺失等风险。为更好地将人工智能应用于学术研究并使其得到可持续发展,需要规范生成式人工智能使用,助推学术研究的向善发展;优化生成式人工智能参数,鉴别学术研究伪成立性;调控生成式人工智能发展,加强学术研究的前瞻预防。

关键词:生成式;人工智能;学术研究;ChatGPT

中图分类号:G644

文献标志码:A

文章编号:1003-2614(2024)03-0001-07

DOI:10.19903/j.cnki.cn23-1074/g.2024.03.005

生成式人工智能(Generative Artificial Intelligence)是一种“能够模拟人类的创造性思维展现出强大的类人能力”的计算机大型语言模型,它能根据用户输入的指令,通过统计与算法,输出新的文本、图像、音频、视频、代码等内容。ChatGPT(Chat Generative Pre-trained Transformer)是生成式人工智能的集大成者,它由美国人工智能公司 Open AI 于 2022 年 11 月 30 日正式发布,并迅速成为历史上用户数量呈现指数增长的应用聊天机器人。ChatGPT 不断推陈出新,目前已更新到 ChatGPT-4,其更新迭代的迅猛与成功的势态把生成式人工智能逼近通用智能的重大问题推到了前台^[1]。面对迅猛发展的生成式人工智能技术,各国政府与世界机构组织均有监管动作。2023 年 4 月 11 日至 8 月 15 日,我国迅速起草、审议、公布并实施了《生成式人工智能服务管理暂行办法》,提出“对生成式人工智能服务实行包容审慎和分类分级监管”原则^{[2][3]};2023 年 4 月 30 日,西方七国(美国、英国、法国、德国、日本、意大利和加拿大)集团数字与技术部长联合声明对生成式人工智能技术提出了“法治”“正当程序”“民主”“尊重人权”“利用创新机会”等五项原则,力争兼顾开发、利用与监管的效能。这些举措表明:监管部门对待生成式人工智能技术新产品的态度是在可控状态下发展,以规范应用为手段,以促良性发展为目的。

与传统的人工智能不同,生成式人工智能既能将输入数据进行处理和分类,也能将现有数据转化生成新的数据,还具备与人类进行对话的基本能力,展现出极强的类似人类创

造力的功能。简言之,生成式人工智能能够提供“问题提示-算法生成-会话响应”的基本服务^[4]。虽然这种服务在科研领域表现为能够帮助科研工作者在海量信息数据与强大的算法加持下进行深度学习和高效研究,但也因其便捷性和创造性的独特属性,又会引发科研工作者对于学术研究的诸多争议,乃至造成学术失真等严重影响。当前,国内外多家学术期刊、出版机构禁止将 ChatGPT 列为论文合著者,而针对这些问题,学术界目前最主要的应对措施就是政策性封堵、禁止使用或限制性使用。封堵、禁止或限制并不是解决生成式人工智能可能诱发的学术诚信问题的最终办法,而是要以一种辩证的态度审议这一现象,通过生成式人工智能的本质及特征,从其应用于学术研究所造成的可能风险出发,深层次发掘造成这些风险的原因,由此提供一种建基于理性思想方案上的可操作路径,从而为我们正确理解与科学地将生成式人工智能用于学术研究提供解释依据与行动范式。

一、生成式人工智能的内涵特征及其应用于学术研究的风险

(一) 生成式人工智能的内涵

随着 ChatGPT 等人工智能机器人的兴起、普及及其各种深度学习和强化学习之类 AI 算法的快速迭代发展,生成式人工智能已经展示了多领域、多学科的类似人类响应的非凡能力。生成式人工智能的内涵已经从传统的基于规则或统计的学习模型转变为深度学习模型,而这种转变同时也引起

收稿日期:2023-07-12

作者简介:周 莎,陕西师范大学教育学部博士研究生,研究方向:学校课程与教学;张 尚,陕西师范大学教育学部博士研究生,研究方向:学校课程与教学。

了学术研究领域的重大演变,意味着计算机模型能够生成新的东西,而不再局限于分析现存的东西。

《生成式人工智能服务管理办法(征求意见稿)》将生成式人工智能定义为“基于算法、模型、规则生成文本、图片、声音、视频、代码等技术”^[5]。国外学者 Weng、Marc 和 Lim 认为,生成式人工智能的内涵包含三个方面:一是利用深度学习模型;二是生成如图像、文字等类人内容以响应;三是复杂多样的提示(如语言、指令、问题等)技术^[6]。此外,自然语言处理(Natural Language Processing, NLP)是一种特殊的生成式人工智能技术,其价值在于让计算机理解、处理、生成自然语言,其核心算法是 Transformer,这是一种基于自注意力机制的深度神经网络结构,具有较强的序列建模能力和表示学习能力。

总之,生成式人工智能是利用计算机模拟人类的智能行为,并开发相应的实践理论与技术,这些技术包括搜索技术、推理技术、文本表示、机器学习与生成式人工智能语言等。生成式人工智能是围绕智能活动而构造的人工系统,涉及计算机科学、信息论、控制论、自动化、仿生学、心理学、数理逻辑、语言学、医学、教育学和哲学等多门学科的研究与开发,能够仿真、模拟、延伸和扩展人类智慧。质言之,生成性人工智能具有独特的能力,不仅能够提供回应,而且可以生成回应中的内容。

(二) 生成式人工智能的特征

1. 大模型的预训练

生成式人工智能基于大规模的语料库进行预训练,可以自动学习并掌握大量的语言知识和语言规律。这种对自然语言进行高度理解和处理的 ChatGPT 技术,在回答问题和生成文本方面均表现出很高的准确性和流畅性,因而其智能表现和效果非同一般,但是这种准确性和流畅性是在技术加持下的一种物化特征,极易消解学术研究的人文性特质。

2. 人工标注数据

生成式人工智能可根据输入的问题自动回答,能够支持多种类型的问答,并在 GPT 系列技术路线上发生了颠覆式迭代。如 ChatGPT-4 在 ChatGPT-3.5 大规模语言模型的基础上开始依托大量人工标注数据,通过专业标注人员为 ChatGPT 提供人类指令的高质量答案,优化 ChatGPT-3.5 无法理解人类指令含义、无法判断输入的情况。尽管 GPT 系列产品以其海量数据的容纳从而实现自身的迭代升级,并以此为人类提供所需求的高质量答案,但是这种服务质量和满意度都是相对的,是带有个人主义倾向的一种偏见选择,而这又极容易导致学术研究的科学性消解。

3. 可扩展性和可定制性

生成式人工智能采用来自人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)进行训练,它是深度学习的产物^[7],其架构、预训练模型结构和参数可以

通过不断迭代和调整进行优化。这充分表明了生成式人工智能具有以适应不同自然语言处理任务的需求和应用场景来增强用户体验的可扩展性和可定制性。在与 ChatGPT 的互动过程中,人们发现它会承认错误、修复自己的答案,这是它从 RLHF 训练中获取的强化学习并重新思考的能力。不过这种认识观极易夸大生成式人工智能的自修功能,同时也会因此弱化作为能动主体的批判能力与行动意识,乃至形成一种惰性精神。

4. 智能的类人属性

生成式人工智能通过掌握特殊标记和经验性学习能给出一定的情感反应,从而形成了对人类的“超模拟”^[8],意即生成式人工智能使用无监督训练和监督微调的组合来生成类似人类的查询响应,并提供类似人类的自然语言进行回应,由最初对人类外形的模拟,发展至如今对人类智能与情感的内在模拟^[9]。同时可以模拟人类对话,使用自然语言处理技术,使得对话更加自然流畅,让生成式人工智能具有类人性。生成式人工智能是包含于机器学习当中,反映了人脑在学习和生成数据、信息以及回应^[10]的类人性智能,它可以通过分析用户的对话和反馈来不断改进和提高自己的服务质量。人的智性嵌入使生成式人工智能具有人的某些属性,以致一部分拥趸相信通过智能产品的智慧升级可以实现人的完全替代,这不仅会给科学研究的思想原创带来现实梗阻,而且会动摇人的主体地位,消解人的存在论价值。

(三) 生成式人工智能应用学术研究的风险表现

人工智能对于人类来说,既充满诱惑,又蕴含许多未知之谜,因而其存在很强的不可预知的风险性^[11]。生成式人工智能虽然能引发人类与数据信息的互动范式重构,但同时也会导致学术研究模式的结构变革。生成式人工智能既能充当科研工作者的助手,为科研机构提供额外资源和创新方案,也会在不经意间为其带来难以预估的风险。

1. 学术研究人文性缺失的风险

第一,失去人文情感和人情味。生成式人工智能可能缺乏人类的情感和情感智能,引发人文情感缺失^[12],由于其缺少情感交流机制,从而导致生成的内容缺乏人情味,无法真正反映出人类的情感和情感需求,无法真正与个体产生情感联系,进而会在社会科学领域的研究中产生冷漠和零情感交互。第二,文化和历史偏见。生成式人工智能可能受到其训练数据的影响,存在文化和历史偏见的风险,人工智能系统有可能巩固或放大文化和历史偏见^[13],导致生成的内容带有特定文化和历史的偏见,缺乏客观性和全面性。第三,缺乏文学创意和审美品位。尽管人工智能写作作为文学创意活动的重要证据^[14],但是生成式人工智能是通过预训练来模仿人类的神经和智能,而文学创意则需要“乌托邦”式的想象力以及优秀的审美品位,由于生成式人工智能缺乏这种文学创意和审美品位,导致所生成的文学作品或艺术品仅仅是

输入指令的产物。

2. 学术研究科学性缺失风险

学术研究有其科学性^[15],但是生成式人工智能是人工标注数据,极易产生错误和缺漏,这就会潜在地消解学术研究的科学性,进而反噬学术研究者的声誉^[16]。这种消解与反噬的可能风险具体表现在:第一,数据偏见。生成式人工智能可能会受到训练数据的偏见影响,人工智能算法会加剧现有的数据偏见,并产生新的偏见,不同算法会产生不同类型的算法偏见。“如果人工智能被黑客攻击和操纵,提供看似客观但实际上是隐藏得很好的有偏见的信息或扭曲的观点,那么人工智能可能会成为危险的宣传机器”^[17],换言之,如果人工标注数据存在缺陷或偏见,算法生成的内容也会受到这些问题的影响,数据偏见影响公众对信息的客观全面认知^[18],从而影响学术研究的科学性和客观性。第二,统计推断风险。生成式人工智能可能无法进行充分的统计推断,导致生成的结论缺乏科学依据或可靠性。在学术研究中,统计推断的准确性和可信度至关重要,生成式人工智能的不足可能引发科学性风险。第三,逻辑推理能力不足。生成式人工智能可能在逻辑推理和推断上重在“自圆其说”,无法进行系统性的逻辑推理。这可能导致生成的结论缺乏科学合理性和逻辑性,使得后续研判可能出现严重偏差。第四,模型不完善和错误。生成式人工智能的模型可能存在不完善或错误,导致生成的内容不符合科学规范,缺乏科学性。

3. 学术研究依赖性成瘾风险

随着生成式人工智能技术应用的加深,必然造成其可扩展性和可定制性的广泛使用,部分科研工作者可能出现依赖成瘾的现象,进而衍生一些负面影响:第一,破坏思维习惯。生成式人工智能的输出结果是基于现有语库而来的,无法复现人类最基础的活动。正因如此,如果成果构成缺乏真实的情感体验和价值输出,就无法解决复杂问题,可能还会产出某些偏见和陈旧观点;如果科研工作者过度依赖生成式人工智能,就会固化人类思维,进而阻碍人类批判性思维和创造力的发展。第二,沉迷虚拟交互。生成式人工智能虽然模拟人类对话交互,但仍是一种人机虚拟交互。使用者可能沉迷于对人工智能技术的崇拜,形成一种技术至上的错误理念^[19],如果科研工作者过度依赖生成式人工智能、沉迷于虚拟交互中,则会导致自身社交能力下降,进而影响学术研究能力的提升。第三,导致技能退化。生成式人工智能掌握一些类似人类的智能和技术,能自动产生自然语言和学术文本,可代替科研工作者完成一定的研究任务。如果过度依赖生成式人工智能,导致人为机器所“奴役”而缺乏个人思考和人为矫正,只会为技术所反噬,逐步丧失某些领域的技能退化和知识缺位,进而影响个人的发展以及对社会的贡献。

4. 学术研究原创性与批判性缺失的风险

由于生成式人工智能的类人属性特征,其对学术研究原

创性和批判性的影响值得关注。尽管生成式人工智能在创造内容方面具有巨大潜力,但也存在一些风险。一方面,对于学术研究的原创性而言,“ChatGPT能够跨越学科疆界生成极富知识含量的文本,但它做不到彻底原创性地生成新知”^[20]。首先,生成式人工智能可以生成源自已有的数据集或文献的内容,并在此基础上进行创作,这可能导致生成的内容缺乏独创性。换言之,生成的作品可能从已有的文献、数据中提取信息,而缺乏真正的创造性。这对于学术研究的原创性会构成极大的挑战,因为这些生成的内容可能只是在已有文献基础上进行的形式变化,而非真正的实质创新。其次,由于生成式人工智能可能会使用大量的数据样本进行训练,存在可能不够全面或不够准确的数据选择,导致生成的结果可能缺乏全面性或准确性,进而降低生成内容的原创性和学术价值。最后,生成式人工智能对于原创性的定义也可能存在争议,因其生成的内容可能是基于大量的数据分析和模式识别而来,并非真正的思考和创造力。另一方面,批判性思维是人工智能无法取代的技能^[21],但这并不影响某些社群相信智能机器能够拥有人的批判性思维,甚至为之付诸行动,也正是这些未经科学证明的主张,极易给学术研究造成失真风险:第一,可能缺乏思辨和批判性思维。生成式人工智能可能缺乏人类的批判性思维和思辨能力,导致生成的内容缺乏深度和逻辑上的一致性,无法真正进行批判性的分析和思考。第二,控制和操纵。生成式人工智能可能受到其训练数据的影响,存在被操纵和控制的风险,导致生成的内容受到特定方面的影响,缺乏客观性和中立性。第三,缺乏反思和自我审查能力。生成式人工智能可能缺乏反思和自我审查的能力,无法对生成的内容进行评估和调整,容易产生未经深入思考的内容。第四,信息过载和误导性。由于生成式人工智能可以大量生成内容,存在信息过载的风险,而且一些生成的信息可能是误导性的,导致学术研究产生不准确或误导性的结论。

二、生成式人工智能造成学术研究风险的原因剖析

(一)生成式人工智能的“拟主体性”导致学术研究伦理道德的迷失

“拟主体性”赋予生成式人工智能具备了特定的“拟伦理角色”^[22]。虽然生成式人工智能是一种具有创造能力的人工智能,它在学术研究中应用具有巨大的潜力和前景,但生成式人工智能在给科研工作者带来条件支持的同时,也会挑战其伦理道德。

首先,伦理道德模糊导致科研工作者无法正确理解生成式人工智能。由于生成式人工智能缺乏统一、权威的定义标准,这就造成了不同科研工作者对其内涵和特征有着不同的理解,使得生成式人工智能使用缺乏一致性和参照性,因此

给学术研究带来了诸多困扰。其次,伦理道德模糊影响了科研工作者的生成式人工智能的开发观与使用观。在学术研究中,面对生成式人工智能可能产生的潜在风险和伦理问题,科研工作者往往缺乏明确的规范和指导,这使得他们在使用、发布和分享生成式人工智能时面临很大的困惑和不确定性。同时,某些研究可能又会涉及社会伦理问题,如人体试验、动物实验或相关敏感性调查等,这些研究需要一定的伦理审查,以此确保研究过程中的参与者和对象的权益受到尊重与保护。最后,伦理道德模糊也给生成式人工智能在学术研究中的道德责任和社会责任带来了挑战。生成式人工智能作为一种具有自主学习和创造能力的智能系统,其行为和决策具有一定的道德后果。由于伦理道德标准的不确定性,因此很难对生成式人工智能的行为和决策负责或进行有效监管。

(二)生成式人工智能的主观偏见导致学术研究价值立场的争议

价值立场争议下的生成式人工智能会造成研究者对于生成结果的主观偏见、信息扭曲和道德责任避让考虑。首先,生成式人工智能模型的训练数据和算法可能受到研究者的主观偏见和信息扭曲的影响。带有研究者主观色彩的价值观和基本信念可能会导致模型在生成结果时出现特定的认识偏见,如政治立场、社会偏好或人种歧视,从而引发生成式人工智能的误用及其观念的错误建立。其次,生成式人工智能模型通常是复杂的黑盒系统,难以解释其生成结果的内在原因和逻辑,这使得研究者不但难以审查和纠正模型的偏见或错误,还会因缺乏解释性进而加剧公众对人工智能系统的质疑,累积不信任感。最后,生成式人工智能模型极易受到对抗性攻击的影响。恶意用户不仅可以有意设计样本或篡改输入,以操纵模型的生成结果来误导、侮辱、违法或伤害他人,还可以利用生成式人工智能传播虚假信息、实施网络诈骗或制造混乱等,进而对社会造成难以预估的负面影响。

生成式人工智能的“能为”与科研工作者的“当为”在价值立场上的不匹配也加剧了学术研究的风险。一方面,生成式人工智能的“能为”与科研工作者的“当为”之间尚缺乏可以评定的价值标准或行为规范;另一方面,生成式人工智能的“能为”价值取向是完全显性呈现,而科研工作者的“当为”价值取向是部分显性呈现,其涵盖范围与作用能力的差异也会造成价值立场的争议。

(三)生成式人工智能的技术指导导致学术研究行为依赖的成瘾

在人工智能的介入下,人机二元对立的思维正朝向人机协同的共生思维转变^[23]。生成式人工智能颠覆性地变革了人类的传播范式,创造了人类传播的“奇点假说”^[24],该假说指出使用生成式人工智能会陷入一种持续且难以控制的依赖之中。一方面,具备信息全面性和使用便捷性的生成式人

工智能极易满足科研工作者的基本心理需要(自主性、相关性和能力感),因此科研工作者对生成式人工智能的依赖逐步增强;另一方面,科研工作者可能会受到社会认同、奖励和观察学习等因素的影响,对生成式人工智能产生强烈的技术依赖。“奇点假说”并不适用于所有人,因为个体差异和环境因素也会对学术研究行为依赖成瘾产生影响。

生成式人工智能之所以导致学术研究行为依赖成瘾,主要原因在于人机问答互动的服务效率和频率。从服务效率上讲,生成式人工智能具备“省时省力”的特征,它能为科研工作者提供诸多便利,直接产出“作品”。生成式人工智能是能够引起科研工作者发散思维的有力工具,可让大语言模型做一些半成品的语义搜索工作;从服务频率上讲,生成式人工智能具备“随时随地”提供服务的特征,能够根据科研工作者的个人选择偏好,提供个性化反馈。与此同时,科研工作者随着自身对生成式人工智能的依赖增强,自身的批判性思维和创造性思维便会逐渐钝化。过度沉溺于生成式人工智能的交互世界,科研工作者并不完全能够与周围的人或事物建立真实且沉浸式的联结,从而失去对现实学术世界的感知能力,也会失去学术敏感性和表达学术观点的能力。

(四)生成式人工智能的技术隐蔽导致学术研究生成文本的缺陷

目前,在技术创变的转折点上,人文社会科学正面临第四次工业革命的冲击,将迎来与新一轮信息技术革命的交叉融合趋势,这种趋势会推动人文社会科学学科研究范式等的演进与迭代^[25]。在海德格尔看来,技术是一种现象学意义上的“现象”,正是由于作为“现象”的技术蔽而不显,使得通常关于技术的诸多议论远离技术的本质。对此,他进一步指出,现代技术的本质是人类的技术性生存方式,是人类不可逃避的“天命”^[26]。科研工作者如果对技术的本质视而不见,那么技术发展有可能将人类命置于无法控制的风险之中^[27]。归根结底,技术的本质就是对知识的运用。

一是生成式人工智能主要通过运用大量的训练数据来生成文本,但训练数据当中可能已经存在偏见和不公平性。如果这些偏见和不公平性在生成的文本中被反映出来,将会对个人、群体或社会产生负面影响。二是生成式人工智能可以生成高度逼真的文本,但这种技术赋能也带来了信息可信度与造假的风险。一些使用者可以借助生成式人工智能生成虚假信息,并利用虚假信息来渲染情绪、搬弄是非、颠倒黑白、造谣生事、伪造文件等,进而给学术研究中的意识形态安全带来隐患。三是生成式人工智能需要大量的训练数据,可能包括敏感学术信息或者原始数据资料。如果这些数据内容被滥用或泄露,会对个人隐私和安全构成威胁。四是生成式人工智能几乎可以生成任何类型的文本,可能产生操纵和滥用他人学术成果的风险,引发科研工作者对技术的信任危机。由于技术本身缺陷,对于生成的文本的内容和后续影

响,不仅很难明确划定责任归属,因而会引发技术层面的争议,也会导致对公众、社会和政治环境的错误认识,破坏沟通和信任。

三、生成式人工智能应用于学术研究的预防机制

生成式人工智能在学术研究中的应用呈现出广阔的前景,但同时也面临一些学术失信问题。为确保生成式人工智能科学合理应用于学术研究,有必要建立系统性的预防机制,维护学术研究生态。

(一)规范生成式人工智能使用,助推学术研究的向善发展

生成式人工智能应用学术研究面临人文性、科学性、依赖性、原创性和批判性风险,如何规避相关风险,减少学术研究的内耗,实现可持续发展,已引起国际社会的广泛关注。规避人工智能技术难以预测的危害,可从顶层设计、标准建立、监管机制等方面着手考虑。

1. 制定并完善生成人工智能使用指导方针和相应规范

生成式人工智能模型尤其是在海量数据集上的训练模型,有可能引入偏见或产生误导性信息。我们应该研究和分析生成式人工智能的最新发展以及已有的指导方针和法律法规,参考国际经验集思广益,设立由专业人士、学者、行业代表和政府官员组成的机构或委员会,负责制定和完善相应的指导方针和法律法规。通过听证会、征求意见、调查研究等方式,广泛听取群众、行业代表、学者和人工智能专业人士的建议和意见,制定明确的使用原则,并在法律法规中明确相应的义务和违法行为的处罚条例。此外,建立相应的监管机构,明确其职责和权力,加大执法力度,定期评估和更新生成式人工智能使用指导方针和法律法规。在已有的基础上,为相关学术研究部门提供培训和指导,确保生成式人工智能研发不仅考虑工具性价值,而且关注复合性的科技-伦理探究^[28],加强对生成式人工智能的使用和合规要求的认知和理解。

2. 建立生成式人工智能输出内容识别与归因技术标准

其一,建立生成式人工智能输出内容的技术识别库,包括使用自然语言处理、计算机视觉、声音处理等相关技术来进行内容分析,在此基础上设计与开发能够对生成式人工智能输出内容进行有效识别和归因的算法和模型。这些算法和模型可以基于机器学习、深度学习、图像处理等技术,并结合大规模数据集对模型进行训练和改进。其二,完善数据集构建和标注。这些数据集应该包括学术研究的内容样本,涵盖各种研究主题、风格和方法形式的生成内容。建立大规模、多样化的数据集,并进行标注,用于训练和验证生成式人工智能输出内容的识别和归因模型。其三,跨领域合作和共享。促进学术界、产业界、政府机构和非营利组织之间的合

作与共享,共同制定生成式人工智能输出内容的识别和归因标准。这样能够获得更多的专业知识和经验,并在不同领域达成共识,促进学术研究对于生成式人工智能使用的自律。

3. 监管生成式人工智能输入指令的信息安全

其一,使用机器学习和自动化技术来监测和筛选输入指令的内容,建立实时监测系统,检测生成式人工智能系统输入指令的异常行为,及时采取响应措施,防止恶意指令的执行,以便更快地识别和处理对学术研究具有潜在污染的违规信息。另外,引入独立的第三方机构或专家对生成式人工智能算法和内容进行审核,确保平台的输入指令符合法规和道德标准。要求数据驱动网络平台向用户和监管机构提供相关信息的透明度,如平台会员的审核流程、标准和监管措施。其二,加强关于数据驱动网络平台使用的公众教育,提高用户意识和责任意识,同时向用户提供适当的工具和机制来报告和限制不当的输入指令,对生成式人工智能系统的输入指令进行安全审计,记录和监测指令的来源、内容、时间等信息,以便追溯和审查。加强对使用生成式人工智能的合理合规教育和思想意识的普及,提高科研工作者对其潜在风险和影响的认知,并引导其合理使用生成式人工智能。其三,建立机制鼓励科研工作者参与到生成式人工智能输入指令的监管过程中,集思广益,不断改进。监管生成式人工智能输入指令的数据驱动网络平台是一个动态的过程,需要不断关注技术发展和社会变化,应及时进行调整和更新监管措施。

(二)优化生成式人工智能参数,鉴别学术研究的伪成立性

生成式人工智能生成的文本内容可能存在事实性错误、知识盲区、概念误用甚至捏造信息等,尤其涉及某些前沿领域和冷门知识,生成式人工智能输出内容的错误率较高,这些不准确的信息如果不加甄别,就会对研究者产生严重误导^[29]。为了进一步优化生成式人工智能参数和鉴别学术研究的伪成立性,可以考虑以下措施:第一,把关数据来源和质量。优化生成式人工智能参数需要基于高质量、真实、多样的数据集进行训练。对于学术研究的鉴别,需要确保所用数据来自可信的学术出版物、学术机构或合法的研究数据集。在进行参数优化之前,对数据进行预处理和清洗是必要的,这包括去除无效、重复、错误的数据以及进行数据格式转换、标准化等操作。第二,参数调整和模型选择。通过参数调整和模型选择,优化生成式人工智能的参数,提升其生成内容的品质和可信度。在学术研究的鉴别中,需要选择合适的模型和参数,使其能够在检测学术论文的真实性和学术规范性方面有较高的准确性。第三,构建评估指标和标准。确定合适的评估指标和标准参数,用于评估生成式人工智能输出内容的质量和可信度。对于学术研究的伪成立性,可以考虑使用领域专家评审、引文分析、重复实验等方法来进行评估。第四,采取基准测试和对比。进行基准测试和对比,比较不

同参数和模型在生成内容质量和学术研究鉴别准确性方面的表现。这样可以找到最佳的参数配置和模型选择。第五,持续监测和改进。建立持续监测机制,对生成式人工智能的参数进行实时监测,并根据监测结果进行及时的参数调整和模型改进。对于学术研究的伪成立性,也需要持续关注新的伪造手段和方法,不断改进鉴别技术。第六,积极合作与共享。促进学术界、产业界和政府机构之间的合作与共享,共同应对生成式人工智能参数优化和学术研究伪成立性的挑战。可以建立共享数据集、合作研究项目,共同制定最佳实践和标准,提高参数优化和鉴别技术的可靠性和普适性。

(三)调控生成式人工智能发展,加强学术研究的前瞻预防

生成式人工智能表征了人类科技进步的“智慧之果”^[30]。任何事物客观上都有两面性,生成式人工智能也不例外,其具有本身的技术“盲区”。生成式人工智能技术系统不是为特定的环境或使用条件而构建的,它们的开放性和易操作性引发前所未有的大规模使用。随着生成式人工智能持续渗透到我们的日常学术研究中,在积极拥抱生成式人工智能的同时,也要考虑到该技术带来的新挑战,有效预防和减少人工智能带来的风险。学术界既需要更加谨慎地评估和监管,以确保生成的内容真正具有学术价值,也需要与相关机构进行深入的讨论和探索,明确生成式人工智能在学术研究中的角色和限制。

为正确引导生成式人工智能助力学术研究的行动路向,应该建立健全生成式人工智能风险监测和预警机制,做到早发现、早介入,努力在萌芽状态消除各种风险,避免风险扩大化。首先,需要对生成式人工智能的应用领域进行全面的风险评估、分类和监测,明确其在学术研究领域可能存在的潜在风险和挑战。建立跨部门的合作机制,包括学术界、产业界和政府部门,共同参与风险监测预警体系的建设,确保多方面的专业知识和视角得到充分整合,以确保生成式人工智能在学术研究中的合理和安全应用。其次,提倡风险共担理念,鼓励相关机构和企业对人工智能系统的风险承担责任和建立相应的风险补偿机制,从制度层面着手,分担社会和个人在使用人工智能时所面临的潜在风险。建立定期的审查和评估机制,监测生成式人工智能在学术研究领域的应用情况,及时发现和应对潜在的风险和问题。此外,加强信息共享和公开透明,建立风险信息数据库和平台,促进各方及时交流和共享风险预警信息。最后,制订应急预案和风险管理政策,针对可能出现的风险情况提前制定有效的预警和防范措施,以最大限度地减少风险对社会和个人的影响。需要加强伦理意识和法律法规的制定,明确人工智能在学术领域应用中可能出现的风险和挑战,并建立相关的法律和伦理规范以约束人工智能的发展和应用。

科研工作者需要在这一科技迅猛发展的时代下静下心来,

跳出当前的系统,重新预估人类发展、前进、奔跑的方向究竟在何处^[31]。因为我们面临的技术挑战不仅来自生成式人工智能,也来自心智健全的人,所以需要我们关注和更新“后生成式人工智能时代”下的学术研究需求,回到学术研究的最初目的,增长学术研究给学习者带来的福祉。

参考文献:

- [1] 涂良川.“生成式人工智能”逼近通用智能的哲学叙事——ChatGPT 追问智能本质的哲学分析[J]. 东北师大学报(哲学社会科学版),2023(4):40-47+93.
- [2][5] 中华人民共和国司法部. 国家互联网信息办公室关于《生成式人工智能服务管理办法(征求意见稿)》公开征求意见的通知[EB/OL]. (2023-04-11) [2023-05-23]. http://www.moj.gov.cn/pub/sfbgw/lfyjzj/lflfyjzj/202304/t20230411_476092.html.
- [3] 中华人民共和国国家互联网信息办公室等七部门. 生成式人工智能服务管理暂行办法[EB/OL]. (2023-07-13) [2023-08-15]. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.
- [4] VAN DIS, E. A. M., BOLLEN, J., ZUIDEMA, W., VAN ROOIJ, R., & BOCKTING, C. ChatGPT: five priorities for research[J]. Nature,2023(614):224-226.
- [6] Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators [J]. The International Journal of Management Education, 2023(21):100790.
- [7] SAHOO, S., KUMAR, S., ABEDIN, M. Z., LIM, W. M., & JAKHAR, S. K. Deep learning applications in manufacturing operations: A review of trends and ways forward [J]. Journal of Enterprise Information Management, 2023(1):221-251.
- [8] 王建磊,曹卉萌.ChatGPT的传播特质、逻辑、范式[J]. 深圳大学学报(人文社会科学版),2023(2):144-152.
- [9] 王治东.人工智能研究路径的四重哲学维度[J]. 南京社会科学,2019(9):39-47.
- [10] PICHAI, S. An important next step on our AI journey[EB/OL]. <https://blog.google/technology/ai/bard-google-ai-search-updates/>.
- [11] 陈爱华.人工智能实然—应然逻辑之思[J]. 逻辑、智能与哲学,2023(1):85-101+247-248.
- [12] 谢娟.如何识别与判定人工智能教育应用伦理问题——基于伦理策略形成的视角[J]. 电化教育研究,2023(1):36-41.
- [13] The Committee on Standards in Public life. Artificial intelligence and public standards[EB/OL]. (2020-02-10). https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
- [14] 张永禄,刘卫东.人工智能写作:创意写作的新景观[J]. 探索与争鸣,2021(3):132-140+179-180.

- [15] 刘跃进. 国家安全学理论中概念及其定义的几个问题[J]. 中共中央党校(国家行政学院)学报, 2023(4): 29 - 40.
- [16] 王 少. ChatGPT 与学术不端治理: 挑战与应对[J]. 科技进步与对策, 2023(23): 103 - 110.
- [17] JIM CHILTON. The New Risks ChatGPT Poses to Cybersecurity[J/OL]. Harvard Business Review Digital Articles, April 21, 2023, <https://hbr.org/2023/04/the-new-risks-chatgpt-poses-to-cybersecurity>.
- [18] 郭小平, 秦艺轩. 解构智能传播的数据神话: 算法偏见的成因与风险治理路径[J]. 现代传播(中国传媒大学学报), 2019(9): 19 - 24.
- [19] 肖 峰, 来 宁. 生成式人工智能关联的主体地位问题[J/OL]. 山西师大学报(社会科学版), 2023: 1 - 8[2023 - 12 - 07].
- [20] 吴冠军. 面向大语言模型的知识实践[J]. 人民论坛·学术前沿, 2023(21): 45 - 56.
- [21] 刘宝存, 易学瑾. 英国人工智能人才培养政策变迁研究[J]. 比较教育学报, 2023(6): 12 - 23.
- [22] 段伟文. 人工智能时代的价值审度与伦理调适[J]. 中国人民大学学报, 2017(6): 98 - 108.
- [23] 朱永新, 杨 帆. ChatGPT/生成式人工智能与教育创新: 机遇、挑战以及未来[J]. 华东师范大学学报(教育科学版), 2023(7): 1 - 14.
- [24] 全 燕. 基于“奇点假说”的人机传播范式建构[J]. 南京社会科学, 2021(12): 109 - 117.
- [25] 蒋华林. 人工智能聊天机器人对科研成果与人才评价的影响研究——基于 ChatGPT、Microsoft Bing 视角分析[J]. 重庆大学学报(社会科学版), 2023(2): 97 - 110.
- [26] 本刊编辑, 张 绒. 生成式人工智能技术对教育领域的影响——关于 ChatGPT 的专访[J]. 电化教育研究, 2023(2): 5 - 14.
- [27] 赵汀阳. 人工智能提出了什么哲学问题? [J]. 文化纵横, 2020(1): 43 - 57.
- [28] 赵 鹏. 科技治理“伦理化”的法律意涵[J]. 中外法学, 2022(5): 1201 - 1220.
- [29] 钟秉林, 尚俊杰, 王建华, 等. ChatGPT 对教育的挑战(笔谈)[J]. 重庆高教研究, 2023(3): 3 - 25.
- [30] 令小雄, 王鼎民, 袁 健. ChatGPT 爆火后关于科技伦理及学术伦理的冷思考[J]. 新疆师范大学学报(哲学社会科学版), 2023(4): 123 - 136.
- [31] 顾小清, 胡艺龄, 郝祥军. AGI 临近了吗: ChatGPT 热潮之下再看人工智能与未来教育发展[J]. 华东师范大学学报(教育科学版), 2023(7): 117 - 130.

Risks of Generative Artificial Intelligence Applied to Academic Research and Its Preventive Mechanisms

ZHOU Sha, ZHANG Shang

(Shaanxi Normal University, Xi'an 710062, China)

Abstract: Generative Artificial Intelligence (AI) is the corresponding practical theories and technologies developed by using computers to simulate human intelligent behavior, and its explosive application in various fields provides infinite imagination for future human development, while at the same time, issues such as false information, academic ethics, and labor replacement have triggered heated debates around the world. Practically speaking, generative AI may pose risks such as lack of humanity, lack of science, addiction to dependency, and lack of originality and criticality in current academic research. In order to better apply AI to academic research and make it sustainable, it is necessary to regulate the use of generative AI to promote the good development of academic research; optimize the parameters of generative AI to identify the pseudo-establishment of academic research; and regulate the development of generative AI to strengthen the forward-looking prevention of academic research.

Key words: generative; AI; academic research; ChatGPT