生成式人工智能所引发的社会争议及其治理*

高　芳[1]　王彦雨[2]　王艺颖[1]

（1.中国科学技术信息研究所，北京　100038；2.中国科学院自然科学史研究所，北京　100190）

[摘　要]从早期基于自然语言处理的聊天系统，到以Transformer为架构的人工智能大模型相继涌现，生成式人工智能已进入商业化、通用化发展阶段，在赋能经济社会发展的同时，正在引发人类社会的深层次变革。作为人工智能的一个分支，生成式人工智能既具有人工智能原有的伦理隐患，也带来了新的社会问题和风险，对现有治理体系提出了新要求。因此，有必要清晰界定生成式人工智能的内涵，基于人工智能60多年发展历程系统梳理生成式人工智能的阶段性特征，研究分析当前阶段其所引发的各类伦理和社会发展争议，梳理主要治理主体的治理实践等。研究发现，因生成式人工智能生成创造性、技术泛化性、现实重构性和类人化等独特属性，其引发的隐私信息泄露、社会不信任、知识产权争议、科研学术造假、社会歧视偏见等各类伦理问题也更具隐蔽性，给人类社会发展带来的挑战正逐渐显现。生成式人工智能引发的社会发展问题，近期包括改变宏观就业结构、加剧全球数字鸿沟、挑战可持续发展等，远期则包括社会形态的变革，甚至人工智能系统“失控”威胁到人类自身的生存等。主要国家和多边国际组织通过机制建设、立法监管、标准制定、以技治技、治理合作等多元实践加强对生成式人工智能的治理，为我国更好统筹人工智能发展与安全、创新与规制，进一步健全人工智能安全监管制度提供了一定经验借鉴。

[关键词] 生成式人工智能　人工智能大模型　伦理风险　社会争议　治理　科技伦理

[中图分类号]D630；TP18　　**[文献标识码]**A　　**[文章编号]**2096-983X（2025）02-0115-13

自2022年ChatGPT发布以来，生成式人工智能快速发展。作为人工智能的一个分支，生成式人工智能表现出强大的生成能力和技术泛化能力，正在加速赋能经济社会发展。然而，其独特的技术属性在带来巨大发展机遇的同时，也加剧了隐私泄露、信息伪造、算法偏见等技术风险，对人伦道德、社会秩序乃至全球命运共同体产生更大范围的冲击，引发人类社会更深层次的变革。现有人工智能治理研究和规制体系已无法有效满足生成式人工智能的治理需求，需要更加客观认识其技术特性，分析其所产生的伦理风险和社会影响，汲取多元主体的良好治理实践和经验，才能构建形成更具前瞻性、长远性和针对性的治理规制体系，以期推动

收稿日期：2024-09-24；**修回日期：**2024-11-20

***基金项目：**国家自然科学基金专项项目“我国开放科学政策研究”（L2124020）；科技部科技强国行动纲要编制专项“培育激励创新的政策环境重点任务研究”（KJQG202309）；国家科技重大专项“新一代人工智能伦理风险评估与应对策略研究”（2023ZD0121701）

作者简介：高芳，研究员，主要从事科技政策与战略、人工智能发展战略、重点科技领域信息分析等研究；王彦雨（通讯作者），副研究员，主要从事科技战略、人工智能风险等研究；王艺颖，助理研究员，主要从事科技政策与战略、人工智能发展战略等研究。

技术的向善发展。

一、生成式人工智能的内涵及独特属性

生成式人工智能（Generative Artificial Intelligence，GAI）是人工智能的一个分支，即基于算法和数据生成创新内容的一类人工智能系统，且是一种强调具有广泛认知能力的智能系统，其目标是实现类似于甚至超过人类的泛化智能，能够理解、学习和解决问题，可以灵活地适应各种任务和环境，具有以下四方面独特属性。

（一）生成创造性

这是核心属性。“生成”意味着从无到有的创造性输出，它与传统单纯的“搜索”模式相对。后者强调的是对已存贮信息进行精确识别与提取，根据已有的数据进行分析、判断、预测。而生成式人工智能则强调“形成新数据、涌现新知识”，更注重在数据学习归纳后进行演绎创造，能够根据训练数据集的模式和规律自主生成全新的、原创的内容或产品。[1]

（二）技术泛化性

在代理任务上通过预训练所形成的生成式人工智能模型，通过微调即可适配到具体的下游任务，以快速适应不同的领域和场景需求，甚至完成跨文本、语音、视觉等多模态、多场景、多任务处理，显示出极强的技术泛化和通用能力。[2]以ChatGPT、文心一言等自然语言生成工具为例，基于预训练阶段习得的模式和统计规律，可以完成聊天、翻译以及文案、脚本、代码等撰写任务，并且能够处理语音、图像等多模态输入，在一定程度上已经具备了通用人工智能的一些核心技术和特征。

（三）现实重构性

生成式人工智能通过自主学习能够生成与真实数据非常相似的新数据，展现出对现实世界的强大理解能力和物理建模能力。例如，以Sora为代表的人工智能文生视频大模型能够根据用户文本提示创建数字视频，深度模拟真实物理世界，生成具有多个角色、包含特定运动的复杂场景，还具有静态图生成视频、视频扩展与缺失帧填充、连接视频、3D一致等诸多功能，充分显示出生成式人工智能在解析真实世界场景方面的巨大潜力。

（四）类人化特征

生成式人工智能在落地应用的过程中逐步显现出类人化智能的特点。以ChatGPT为代表的系列生成式人工智能产品与服务具备很强的“人性”，可以按照类似于人类的自然语言表达方式，模仿人类的行为习惯和浅层思维模式，自然地与人类的生产生活环境进行交互。[3]这种融合类人智能“强拟人化”的技术特征，可以帮助智能技术与人类社会进行深度交互。

二、生成式人工智能的源起及发展历程

关于生成式人工智能，有一个错误的观念，即认为它是近几年才出现、以ChatGPT为标志的新生事物。然而实际上，生成式人工智能在人工智能发展初期便已出现，如具有新信息生成功能的自然语言对话程序等。只不过在很长一段时间里，早期的生成式人工智能主要限于实验层面、没有进入市场，且主要依赖已设定的逻辑规则，不具备学习能力。整体看，生成式人工智能的发展经历了如下几个阶段。

（一）生成式人工智能的诞生：基于启发式规则的聊天程序（1960s—1970s）

第一个具有生成功能的人工智能程序，是1966年由Joseph Weizenbaum所发明的聊天程序ELIZA。ELIZA通过运行一个名为“医生”的脚本，可以模仿心理治疗师的说话方式与用户聊天互动，但它并没有考虑上下文等情景因素，不具备语义“理解”及复杂对话功能。另一个类似的聊天程序是PARRY，它于1972年由斯坦福大学的Kenneth Colby研发，模拟的是具有偏执型性格的精神分裂症患者，同样基于特定的逻辑

规则进行语言输出。在1972年的ICCC会议上，ELIZA通过阿帕网与PARRY进行了模拟对话。[4]早期的生成式人工智能，主要依靠预先设定的简单规则，但有诸多局限性，如：一是依赖已有规则，只能处理明确编程到其中的任务，泛化性及灵活性差；二是学习能力较弱，难以通过持续的信息输入进行主动学习并生成新的知识；三是只能处理简单的任务，可扩展性有限。[5]

（二）生成式人工智能的功能扩展：从自然语言对话到多样化领域（1980s—2010）

20世纪七八十年代起，生成式人工智能的功能由单纯的人机对话，逐渐扩展至散文诗创作、绘画、科学知识生成等领域，代表性的生成式人工智能成果包括三种。第一，绘画生成式人工智能。1971年，Harold Cohen开发了绘画生成系统AARON，可生成原创艺术作品。1979年在旧金山现代艺术博物馆展览上，展出了一幅由AARON绘制的长约100英尺的壁画；1980年，AARON为美国计算机博物馆创作了艺术作品《春天的春花》（Primavera in the Spring）。[6]第二，散文生成式人工智能。1983年，William Chamberlai开发出散文诗生成程序Raacter，并基于这一程序，于1984年生成一本长达120页的书籍《警察的胡子是半成品》。[7]第三，音乐生成式人工智能。20世纪80年代，David Cope研发出可生成原创音乐作品的人工智能算法EMI，它首先通过分析将音乐中的结构编码化，并基于特定的规则对其进行解析、重组，从而生成新的、有趣的音乐架构。EMI制作了几张专辑，包括《虚拟莫扎特》和《虚拟拉赫玛尼诺夫》，甚至创作了一部莫扎特风格的完整交响曲，并于1997年在圣克鲁斯巴洛克音乐节上演出。[8]

（三）基于人工神经网络的专用型生成式人工智能：范式的转换（2010—2020）

2010年后，基于神经网络的生成式人工智能开始兴起，可通过数据训练，习得符号信息中所隐藏的模式与结构，重点是分析符号间的上下文关系或强关联网络，从而生成或预测复杂、连续的内容。2014年古德费罗（Ian Goodfellow）提出生成对抗网络，并广泛应用于图像、视频生成。这一时期的生成模型，主要是专用型的，仅限于特定领域，典型系统包括：一是稿件编辑。2012年，《华盛顿邮报》发起名为“Truth Teller”的实时新闻核查项目，机器人开始介入新闻写作领域。此后，一系列新闻撰稿机器人程序被研发出来，如《纽约时报》的Blossom、《卫报》的Open001、路透社的Open Calais等。二是电影制作。2016年，电影导演Oscar Sharp和人工智能专家Ross Goodwin合作，发明电影创作算法“Benjamin”，2018年，Benjamin算法在48小时内完成了电影《Zone Out》的创作。[9]三是文学作品创作。2017年，携程上线了一款名为“携程小诗机—为你写诗”的智能程序，用户通过上传风景照，便可一键成诗。2017年5月，微软小冰出版了其原创诗集《阳光失了玻璃窗》，成为人类历史上第一部100%由人工智能创造的诗集。[10]四是声音/歌曲合成。2023年5月，免费人工智能语音转换项目So-vits-svc，合成了歌手孙燕姿的音色，以及上千首基于其音线所形成的歌曲。五是图像及视频生成。2022年，人工智能绘画模型DALL-E2、Midjourney、Stable Diffusion相继问世，美国设计师杰森·艾伦借助Midjourney创作的油画《太空歌剧院》，获得了美国科罗拉多州博览会美术竞赛“数字艺术”类金奖。[11]

（四）生成式人工智能的商业化及通用化：人工智能知识生成走向普通大众（2018至今）

2017年，随着Transformer算法的推出，大型生成式模型得以迅速发展。2018年，OpenAI和谷歌分别推出大模型GPT-1及BERT。2023年3月，OpenAI推出参数量达1.76万亿的GPT-4。此后，一系列大模型陆续出现，生成式人工智能快速进入商业化、普用化、多模化时期。这一时期的生成式人工智能，其功能日益多样化，且操作过程呈现“自然语言化”特征。如ChatGPT可完成写作辅助、摘要生成、智能问答、机器翻译、文本生成、歌曲编写、代码自动生成等多种任务，且所有的需求、操作均可化约为自然语

言。同时，人工智能模型的多模态化特征日益突出，知识互转换与互生成趋势愈发明显，生成式人工智能日益成为一个通用型的知识生成平台，各类符号与信息之间（如图片、文字、视频、公式等），可实现共融、共通、共译、共生。尤其是2023年以来，一系列多模态大模型陆续推出，包括OpenAI的GPT-4V，Anthropic的Claude3，以及谷歌的Gemini等。

三、生成式人工智能引发的社会伦理问题

随着人工智能技术的广泛应用，工具理性不断膨胀，人的道德主体性不断削弱。其伦理困境的根本原因在于，人工智能可以不依赖人类进行自主决策，却没有既定的道德标准来指引其决策。[12]在这种困境下，人工智能技术及应用因不符合道德规范或伦理准则，而对人类社会层面造成的负面影响、危害乃至破坏，即为其社会伦理风险。这种风险往往是涉及人伦道德层次的根本性、原则性的问题，具体又可以分为基于技术本身的风险，如数据隐私泄露、算法歧视；基于技术开发的风险，如知识产权争议；以及技术在具体场景下的不当应用所产生的风险，如深度伪造、科研造假等。

而生成式人工智能因其决策的自主性，其内容生成的多模态性、功能的通用性、应用的普遍性，进一步放大了人工智能所引发的伦理问题。

在我们讨论生成式人工智能伦理议题时，应注意以下几点：第一，问题形成的非源生性，即生成式人工智能所引发的一系列问题（如隐私、歧视等），并非其所独有，此前的一些人工智能产品，已内含此类的相关争议；第二，争议生成的广泛性，即生成式人工智能的普及化、平民化、商业化特征，加剧了此前各类人工智能伦理和风险问题形成的频繁度，形成了争议的“放大效应”与“叠加效应”；第三，争议制造的简单性，生成式人工智能在操作上更为简单，普通用户可基于自然语言指引其内容生成，不需要复杂的专业技术操作，从而造成约束困难、治理失效。整体看，当前生成式人工智能所形成的伦理争议问题主要包括以下几个方面。

（一）加剧信息泄露忧虑

数据是人工智能的基础，人工智能大模型的训练需要大量的数据。一方面，生成式大模型在信息抓取、处理方面更具侵略性、隐蔽性。一些生成式大模型如ChatGPT，既可通过爬虫工具从公共语料库或网站自动抓取信息，也可基于人机交互实时提取敏感数据，数据层面的不当使用或将造成用户隐私信息泄露。另一方面，生成式人工智能产品通常会对其所有数据信息进行加密加噪处理，但由于自身层面的不当技术处理，或是储存管理过程中的不规范，一些包含指纹识别、人脸识别等敏感信息的个人数据仍有可能被恢复和窃取。ChatGPT就像飞机上的黑匣子，可记录、处理用户对任何主题的相关信息。这种情况下，用户往往不再能够控制自己的信息，并引发普遍的隐私及数据安全隐忧。2023年，三星设备解决方案部的内部员工由于使用ChatGPT，导致了多起公司内部数据泄露事件，最终三星决定限制员工通过公司电脑使用ChatGPT。[13]与之类似，摩根大通也因担心泄露财务信息，严禁内部员工在办工场所使用ChatGPT。此外，韩国SK海力士、日本软银集团、美国银行、花旗集团等企业也采取了类似措施。

（二）引发社会信任危机

生成式人工智能使内容生成的方式、形态和传播渠道更加逼真和高效。在网络化时代，智能新技术在信息传播与新媒体行业的误用或滥用，加剧了虚假信息、深度伪造等问题的发生和严重程度，将逐步侵蚀社会信任系统。其引发的社会信任危机主要包括三方面。一是结果的可错性问题。由于机器学习算法尚不完全成熟，算法“黑箱”等问题具有不可解释性和不可控性，生成式大模型无法确保在各类复杂场景下的决策正确性。特别是数据源的缺陷以

及训练和推理阶段的技术局限性，极易引发人工智能“幻觉”问题，使人工智能系统生成毫无意义或完全不符合源内容的结果，完全信任ChatGPT等会非常危险。如2023年6月，纽约律师史蒂文·施瓦茨，利用ChatGPT生成了6项虚假案例和裁决并用于法庭陈述，最后他向法官道歉，表示自己不知道ChatGPT会编造虚假信息。[14]二是社会欺诈问题。具有强大的内容生成能力的人工智能系统若被主观恶意使用，或将助长传播虚假信息、社会欺诈等犯罪行为，成为网络犯罪分子设计更高效战术的推动者。据Sumsub研究显示，从2022年到2023年第一季度，美国基于深度伪造的社会欺诈事件数量翻了一番，且仅在2023年第一季度，深度伪造在美国所有社会欺诈事件中的比例从0.2%上升到了2.6%。[15]三是政治生态破坏问题。虚假信息的产生具有主观因素，在国际政治中常被作为一种斗争手段，即利用人工智能生成并传播政治人物的虚假视频、语音等来进行心理战、信息战，进行意识形态渗透，从而维护部分主体的既得利益。如2024年1月，新罕布什尔州的选民接到了一个据称来自美国总统拜登的自动电话，敦促他们放弃参加该州的总统初选，最后发现这是一则合成电话。[16]

（三）泛化知识产权争议

生成式人工智能作品的知识产权归属一直饱受争议。[17]围绕生成式人工智能是否可以作为法律实体承担法律责任，其生成的内容是否享有版权保护，在训练及生成过程中引用的素材是否构成侵权等问题，法律层面尚未形成明确规定。特别是随着算法模型规模的增长，大量互联网文本或图片都被用来训练模型，而该类操作都是在没有得到作者明确同意的情况下进行的，目前已经引发了大规模的争议和声讨。2023年12月，美国《纽约时报》宣布将起诉OpenAI和微软，指控其未经允许使用了该报数百万篇文章。[18]同时，生成式大模型因具有新知识、新作品涌现能力，而引发一系列与知识产权相关的争议问题，如作者署名优先权、原创衍生作品的著作权等，这一争议在AI for Science大背景下讨论显得尤为急迫而重要。特别是在科研领域、文化领域等生成式人工智能应用最为普遍的场景下，智能工具引发的知识产权争议尤为普遍。例如，《自然》杂志针对科研人员广泛使用生成式工具协助论文撰写所导致的产权纠纷，制定了两条原则，规定任何大型语言模型工具都不会被接受作为研究论文的署名作者；使用大型语言模型工具的研究人员应该在方法或致谢部分记录相关使用情况。[19]2022年底，三位艺术家联合起来，对各种生成式人工智能平台提起集体诉讼，指出其在没有获得许可的情况下使用了他们的原创作品。[20]

（四）隐秘化科研造假

生成式人工智能在科研领域的快速广泛应用，对科研活动和科研评估中的诚实、尊重、责任、公平和可信原则造成了冲击。ChatGPT上线两个月后，众多学者就宣称其为一种文化轰动，对科学界和学术界造成严重影响。[21]其引发的科研造假行为，如生成虚假科学图片、捏造虚假科研证据等事件正在与日俱增，且往往难以检测，主要包括以下几类：一是生成欺诈性论文。据科技新闻网站404 Media的一份报告，网上已出现许多由ChatGPT生成的欺诈性论文，涵盖主题包括脊柱损伤、电池技术、农村医疗、细菌感染等。来自捷克查尔斯大学的研究人员利用ChatGPT，制作了一篇完全虚构的神经外科文章，包括摘要、方法、数据、结果和讨论，且仅用1个小时完成。[22]二是制造虚假参考文献。根据Athaluri S. A的一项研究，发现ChatGPT所生成的178篇参考文献中，69篇没有DOI，其中28篇根据不存在。[23]而另一项关于ChatGPT所生成的医学文章参考文献真实性的研究发现，在生成的115篇参考文献中，47%为捏造，46%真实但不准确，只有7%是可靠的。[24]三是数据及证据造假。2023年11月发表在《美国医学会眼科杂志》上的一篇论文，发现GPT-4可在几分钟内生成“看似真实的数据库”。[25]德国海德堡FEBS出版社图像完整性分析师Jana Christopher

谈道："我见过人工智能刚刚生成的假显微镜图像，但如何确凿证明图像是由人工智能生成的，仍然是一个挑战。"[26]

（五）放大歧视及偏见

人工智能的道德决策在很大程度上是人类预制的，确保人工智能算法合乎伦理是一个复杂的工程，需要人类在智能系统中嵌入并权衡不同的道德、宗教、政治、性别、民俗等因素权重。[27]然而，人类认知的局限性或将导致训练数据或算法设计中植入历史偏见，进而使得模型训练产生的结果带有歧视和偏见，道德普适性难以实现。生成式人工智能技术的强大，或将放大已有的社会歧视或偏见，如性别歧视、种族主义，或是对特定类型群体权益的忽视或无视等，并随着广泛应用对大批用户和受众产生潜移默化的影响。例如，通过检查来自Stable Diffusion的5000多张图片，发现它放大了种族和性别不平等，如将白人男性描绘成首席执行官，而女性则扮演从属角色，把黑皮肤女性定型为从事卑微的工作等。另据Cem Dilmegani的研究，与2018年开发的1.17亿参数模型相比，新近所创建的2800亿参数模型，其"毒性"或偏见水平增加了29%，且随着更大规模的人工智能模型的开发与部署，偏见风险有可能会继续增加。[28]在造成生成式人工智能偏见的多重因素之中，训练数据本身的不完善性以及由此所导致的统计偏差是最为突出的因素。如某些类型的数据占比过多（白人相对于黑人，身体正常人群相对于残疾人群，英语人群相对于非英语人群等），社区擦除（如根据安全规定，导致某些社区的数据被消除），或群体边缘化（一些群体的数据由于缺乏代表性而被边缘化）。此外，"毒性"信息（如含有种族歧视信息、暴力和仇恨言论）也会被模型所习得。生成式人工智能系统不可能是中立或客观的，也不可能包含真正的普世价值。如何确保智能系统遵循的道德准则适用于各类应用场景，如何调和不同群体间的价值紧张关系，这始终是人工智能伦理面临的一大难题。

四、生成式人工智能引发的社会发展问题

生成式人工智能因其强大的赋能潜力，成为大国竞相投资的热点，旨在推动新技术加速与经济社会深度融合。根据斯坦福大学所发布的《2024年人工智能指数报告》，2023年世界各国对生成式人工智能投资激增，由2022年的30亿美元上升至252亿美元，增长近8倍。[29]然而，当生成式人工智能充分商业化，成为经济社会发展的核心驱动力时，它所带来的人类社会发展争议问题更不容忽视。在底层伦理问题之外，生成式人工智能技术快速发展所带来的资源消耗，及其赋能应用之广之深，将对社会系统的良序运行造成更为深远的影响，这种影响不仅涉及人类个体、机构、集体，更涉及社会发展与人类生存等长远性、全局性的问题，需要全社会合作应对解决。[30]生成式人工智能引发诸多社会发展问题，包括：第一，生成式人工智能对人类脑力劳动的替代性更强，大量重复性甚至是一些创造性的脑力活动面临被机器替代的可能；第二，生成式人工智能的架构及训练成本更高，依赖巨量的投资、需要大规模的算力芯片，这使得生成式人工智能成为少数国家的游戏，从而加大国际发展的不平衡；第三，生成式人工智能对能源的消耗量极大，从而引发了人们关于可持续发展的忧虑；第四，生成式人工智能日益成为社会发展的基础性底座技术，围绕数据算法所形成的生产要素垄断化、社会运转智能化，以及生活方式的"虚拟—现实"融合化，深刻改变着人类社会的存在与发展形态；第五，生成式人工智能的自改进、自演化能力日渐突出，社会各界愈发关注"人工智能是否会引发人类的生存危机"问题，等等。

（一）改变宏观就业结构

马克思曾提出，劳动资料一作为机器出现，就立刻成了工人本身的竞争者。[31]历史上蒸汽技术、电力技术、信息技术等每次技术变革都引发了劳动力结构的变化。以蒸汽技术为

例，蒸汽机的发明和应用使传统纺织、采矿、运输等重体力劳动实现机械化替代，随着城市化进程加速，大量人口向工业中心集聚，又催生了车间工人、机器操作员、火车司机等新的职业群体。以生成式人工智能为代表的新一轮人工智能技术革命，也将推动宏观就业结构发生深刻变化，具有类人化智能水平的生成式人工智能将取代大量基础岗位，并对传统的生产方式和工作内容产生颠覆性影响，引发劳动需求和劳动市场结构的变动。[32]刘益东指出，人工智能造成“短期失业、长期增业”论不能成立，通能革命势必引发大规模失业，即“通能塔诅咒”。[33]世界经济论坛研究指出，ChatGPT、Midjourney、Github Copilot等生成式人工智能的飞速发展，对全球经济和劳动市场产生巨大影响。未来5年内，全球23%的就业岗位将发生巨大变化。例如，信贷授权员、电话营销员、统计助理、出纳员、记者等可能会被人工智能替代实现自动化，同时也会创造出一些全新的职业，比如人工智能内容创建者、人工智能数据管理者等岗位。[34]总体上看，人工智能特别是生成式人工智能的发展，使得受技术影响的劳动力类型，从体力任务向认知任务转变。国际劳工组织研究发现，文职工作受生成式人工智能的影响最大，24%的岗位受到高度影响，58%的岗位受到中度影响；而在管理类、专业类和技术类等其他职业类别中，只有1%—4%的岗位受到高度影响，不超过25%的岗位受到中度影响。针对不同国家，高收入国家总就业岗位的5.5%可能受到技术自动化的影响，而这个数字在低收入国家只有约0.4%。[35]

（二）加深全球数字鸿沟

历史证明，新技术可以消除基于旧技术的不平等，但也可能产生更新更大的不平等。生成式人工智能的研发与应用对硬件设施、数据资源基础，以及劳动力数字素养有着较高的要求，经济发展水平特别是数字经济发展的差异，可能加剧不同国家、不同群体之间的发展鸿沟。[36]特别是在生成式大模型研发方面，面对高昂的开发成本和高水平的技术要求，部分发展中国家难以自主研发本土模型，而是要依赖于其他发达国家的开源架构和云计算服务等来发展本土技术，从而无法建立起真正的主权人工智能，进而加大全球数字鸿沟，导致国家间的权利失衡。联合国教科文组织在《生成式人工智能教育与研究应用指南》中指出，生成式人工智能依赖于大量的数据和强大的计算能力，这些创新大多只适用于国际大型科技企业和少数经济体（主要是美国、中国、欧洲国家）。如根据《2024年人工智能指数报告》显示，2023年世界上大部分的基础模型主要来自美国（109个），其次是中国（20个）和英国（18个），且自2019年以来，美国一直是基础模型的最大产出国。生成式人工智能的训练及维护，均需巨量的资源和资金投入，如GPT-4在训练过程中使用了价值近7800万美元的计算资源，而谷歌的Gemini Ultra的计算成本则更是高达1.91亿美元。[29]那些无法获取足量数据、难以购进大模型发展所需的芯片和算力资源，以及无法承担大模型训练所带来高昂成本的国家或地区，将陷入“数据贫困”及“算法荒漠”的境地。[37]

（三）挑战可持续发展

生态共同体是可持续发展伦理理念的体现，人工智能等新技术发展改变了人与自然之间的关系，生成式人工智能的产业发展模式对生态共同体的冲击和影响是无法忽视的。模型训练的计算与环境成本和模型规模成正比，算力支撑不仅意味着大量的电力和水资源消耗，也带来碳排放量的增长，从而对全球可持续生态产生影响。斯坦福大学人工智能研究所披露数据表明，OpenAI公司的GPT-3单次训练即需要1287兆瓦时（相当于128.7万度）的电力。[29]国际能源署《2024年电力》报告提到，谷歌搜索一次平均耗电量为0.3瓦时，而ChatGPT请求一次平均耗电量为2.9瓦时，如果将ChatGPT应用到每天谷歌90亿次搜索中，则将需要近10太瓦时（相当于100亿度）的额外电力。[38]随着生成式人工智能技术广泛应用于社会生活的各

个领域，智能技术的资源消耗量将继续增长，对全球生态环境造成的危害是难以预计且不可逆转的。有研究提出，假设人们对人工智能技术的兴趣持续增长且芯片供应充足，那么到2027年，生成式人工智能所消耗的能源可以为荷兰大小的国家提供一年的电力，相当于约850亿–1340亿度电。[39]《2023年人工智能指数报告》提到，BLOOM模型的训练过程产生的碳排放量是单人从纽约飞往旧金山单程航班碳排放量的25倍，而GhatGPT–3的碳排放量是BLOOM的20.1倍。[40]而随着大模型热度的不断升温，谷歌、微软、百度、腾讯等众多科技公司陆续投入ChatGPT类模型研究，也增加了同质化重复性的模型训练。

（四）变革社会存在形态

未来，随着技术的日渐成熟与应用领域的日益广泛，生成式人工智能可能成为整个社会发展的基础性底座技术，人们的生活方式和人类社会的存在形态等将发生深刻变革。第一，数字资源的寡头化控制。未来生成式人工智能的垄断化趋势将更为明显，主要表现在两个方面，一是市场的垄断化，由于生成式人工智能的发展，强调模型大型化、数据资源巨量化，以及计算能力规模化等重投入方式，只有少数企业或机构能够负担这样的成本，因此少数人工智能企业将主导未来市场；二是功能的垄断化，即随着功能的多模态化、通用化，少数生成式大模型将侵吞、整合分散的模型应用（APP），各类特定功能算法将会统一到生成式大模型的软硬件生态之中；第二，社会资源的智能化聚合。未来的生成式大模型将具有更强的社会资源吸纳能力，工业生产、交通、科学研究、教育、医疗、娱乐等领域可能都被聚合在大模型之中，生成式大模型成为社会生产与生活效率提升的关键引擎；第三，生活方式的“虚拟—现实”融合化。未来的生成式大模型，融合6G、VR、机器人、脑机接口等技术，形成与现实物理世界的映射、甚至相互补充，人类生活方式的“虚拟—现实”融合化将成常态。这一生活方式，将使得人工智能所引发的伦理问题更加普遍化、多样化和复杂化。关于智能机器的人文社科研究也愈发受到社会关注。

（五）引发人类生存忧虑

人工智能对人的主体性带来了更大的挑战。生成式人工智能模型所具有的涌现能力，甚至是自改进能力，增加了人们关于其是否会形成自我意识、难以控制其演化路径的忧虑，即人工智能是否会导致人类的生存性风险。在很长一段时间里，机器取代、威胁人类生存的相关论述，主要是哲学式的思辨，以及艺术的幻想，与人工智能的现实技术发展关联不大。然而就在本轮人工智能发展热潮之初，霍金便提醒“人工智能的研究与开发正在迅速推进，也许所有人都应该暂停片刻，把研究重点从提升人工智能的能力转移到最大化人工智能的社会效益上面”，并强调“我们的人工智能系统须要按照我们的意志工作”[41]；基辛格提出“启蒙运动的终结”思考，认为“人工智能系统表现出了一种此前被认为仅有人类才会具有的信息处理能力”。[42]如今，生成式人工智能大模型的持续突破，则使得上述担忧和警告具备了更加丰富的现实技术基础。随着智能系统复杂程度越来越高，越来越具有“人性”，雷·库兹韦尔（Ray Kurzweil）所预言的“奇点”（即人工智能超越人类智能的转折点）是否已经临近，比人类聪明的人工智能是否会人类“友好”，甚至异化为刘易斯·芒福德（Lewis Mumford）所担心的统治人类的“巨机器”，最终危害人类生存，此类的担忧被不断放大。如2023年11月，OpenAI发布了模型Q–Star，并被证明具备自主推理、自主学习、自主纠错，甚至是自我进化等能力，人们愈发担心人工智能是否已具备自主意识。2023年3月，在ChatGPT发布不久，图灵奖得主约书亚·本吉奥（Yoshua Bengio）等千余名人工智能界的企业家和学者，联名发布题为《暂停大型人工智能实验》的公开信，呼吁暂停训练GPT–4的后续人工智能模型至少6个月，以避免人工智能对人类生存与发展造成不可逆的风险。

五、生成式人工智能的政策治理及实践

面对生成式人工智能引发的伦理风险和社会争议，主要国家特别是人工智能强国都在加紧开展人工智能治理体系建构：在治理架构上增设专门机构，在治理手段上综合运用国家立法、行业规范和政府监管等，同时坚持以技治技、加快推进生成式大模型安全研发，在治理合作上强调全球共建共治。

（一）设立专门机构，形成专业推进力量

当前多国政府已形成发展人工智能的组织管理体系，但生成式人工智能的快速发展依然引起各国政府高度关注，需要专门的建制化力量来“管控”生成式人工智能发展。2023年10月，拜登签署《关于安全、可靠、可信地开发和使用生成式人工智能的行政令》，以安全和负责任地管理人工智能为目标，强调由不同机构在职责范围内进行人工智能监管具体政策制定。随后，联邦政府设立了多个专门负责生成式人工智能相关事务的政府机构。白宫总统科技顾问委员会成立生成式人工智能工作组，以帮助评估关键机遇和风险，就如何确保公平、负责任和安全地技术开发部署提供意见。国防部成立“利马”生成式人工智能特别工作组，负责分析大语言模型和生成式人工智能的潜在任务领域、工作流程和应用案例。2024年4月“利马”工作组计划启用“虚拟沙盒”中心，允许军事人员在其中使用经审核的生成式人工智能工具开展实验，并向国防部提交了230个生成式人工智能用例用于进一步探索验证。商务部建立国家人工智能安全研究所，以审查前沿人工智能模型，制定验证人工智能生成内容的标准，加强模型安全测试和风险评估。此外，法国于2023年9月成立生成式人工智能委员会，专门围绕生成式人工智能发展系统研究，并为国家人工智能战略更新完善提供咨询。欧盟于2024年5月成立由科技专家、律师和经济学家等组成的人工智能办公室，负责评估和测试通用人工智能。

（二）加快推进立法，提供法律保障

法律法规为生成式人工智能监管提供了最根本的制度性保障。ChatGPT、GPT-4等生成式人工智能工具的出现加速了人工智能的立法进程，部分国家和经济体以安全为优先导向，对生成式人工智能等新兴技术实施具有强制性、约束性的立法监管。根据斯坦福《2023年人工智能指数报告》统计，2023年有49个国家在立法程序中提及人工智能，共计2175次，几乎是2022年（1247次）的两倍。[40]2024年4月，美国众议员亚当·希夫（Adam Schiff）向众议院提交了一份新提案《2024年生成式人工智能版权披露法案》，要求生成式人工智能平台披露其在训练人工智能模型时对受版权保护的作品的使用情况，并具有追溯效力。2024年8月，欧盟《人工智能法案》正式生效，成为全球最具有法律绑定效力的人工智能管制法律，采用分级分类的风险判定和监管框架，规定生成式人工智能必须遵守透明度义务和欧盟版权法，发布训练数据的详细摘要，禁止智能系统操纵人类行为等，大部分规则将于2026年8月2日开始生效，但涉及ChatGPT等人工智能模型的规则，将于2025年8月便开始适用。[43]加拿大于2023年推出了《人工智能和数据法案》（草案），并宣布将在该法案框架下优先监管生成式人工智能系统。但同时需要认识到，目前人工智能法律规则制定仍处于初期的探索阶段，为人工智能设计、研发、制造和使用等行为设计一套完整的法律规则，并应用到执法、司法等活动中，是一项复杂的任务。目前多数法律仍以部分环节、具体场景为重点，而且重点解决的是当下生成式人工智能应用涉及到的法律问题。

（三）制定标准规范，探索风险评估

生成式人工智能技术在快速进步，而人类自身的认知能力在特定历史条件和阶段具有局限性，对生成式产品的风险认知往往也具有滞后性。对新技术进行及时全面有效的风险评估和监测，以制度标准中的原则规范约束技术的向善发展，是规避风险的必要手段。[44]部分国

家积极推动风险评估工作的标准化进程，以识别规避风险，实施敏捷治理。2023年1月，美国国家标准与技术研究院（NIST）发布《人工智能风险管理框架》，为用户设计、使用和部署生成式人工智能系统提供了标准依据。日本经济产业省发布《生成式人工智能在内容制作中的应用指南》，重点围绕游戏、动漫、广告等行业，明确了使用生成式人工智能需要注意的法律问题和应对措施等。2024年6月，法国竞争管理局发布《关于生成式人工智能领域的竞争运作》，分析了OpenAI公开发布ChatGPT以来生成式人工智能领域市场竞争的运作情况，提出在价值链上游存在的潜在竞争风险，并提出维护公平竞争、促进创新发展的建议。2023年4月，欧洲成立算法透明度中心（ECAT），监督和调查大型网络平台算法的合规性。2023年7月，美国联邦贸易委员会正式发起对生成式人工智能聊天机器人风险的审查，就OpenAI是否违反消费者保护法开展调查。英国人工智能安全研究所初步建立了可操作的评估方法体系，并推出名为“Inspect”的人工智能安全测试平台，向全球人工智能社区开放。2024年4月，美国和英国宣布在人工智能安全和测试方面建立伙伴关系，两国计划在可公开访问的模型上对前沿人工智能模型开展联合测试，同时还将共同推进人工智能风险和安全技术研究。

（四）加大支持力度，推动技术研发

从技术源头进行安全治理是防范技术风险产生的有效手段。部分国家通过资金补贴、政府采购等干预性财政工具，提高生成式人工智能的可信性、准确性、安全性，推动生成式人工智能技术安全研发和应用。2023年，英国政府划拨1亿英镑投资成立基础模型工作组，联合产学研领域专家，聚焦可信赖的基础模型研发攻关，并通过公共服务采购，促进安全可靠的基础模型的广泛应用。英国人工智能安全研究所正在多个领域开展基础性人工智能安全的研究工作，包括人工智能模型的可解释性、对模型行为的干预以及人工智能对齐的新方法等。美国国防部计划在2024财年增加与人工智能相关的网络安全投资，总额高达2457亿美元，其中674亿美元用于网络IT和电子战能力。DARPA自2019年以来持续投资语义取证（SemaFor）项目，以开发能够自动检测、溯源和表征虚假多媒体信息（如文本、图像、音频和视频等）的技术，用于抵御大规模生成式虚假信息攻击。美国国家标准与技术研究院先后启动了GenAI计划、“人工智能风险和影响评估”（ARIA）项目，前者专注于评估各种生成式人工智能技术（包括文本和图像生成），后者主要聚焦大模型相关风险与影响开展评估方法研究，旨在产出指南、工具、方法论和指标等体系化评估方法。美国能源部在2025财年预算中拨款4.55亿美元，用于支持人工智能基础模型研发和人工智能系统安全性、可靠性、稳定性研究。2024年7月，美国技术现代化基金向商务部拨款1000万美元，支持其下属人工智能安全研究所研发面向生成式人工智能等各类人工智能系统的标准化评估流程。2024年3月，欧盟发布《在科研领域负责任地应用生成式人工智能指南》，分别为科研人员、研究机构和研究资助机构提供指导，确保生成式人工智能能够在科研领域中发挥积极作用，同时避免其潜在风险和滥用。

（五）开展治理合作，共建国际框架

让人工智能符合人类社会伦理价值，实现全球普惠发展，是国际社会需要共同协作解决的问题。主要国家、双边和多边国际组织等都在积极倡导和推进人工智能治理合作，特别是在全人类共同面临的生成式人工智能发展与安全问题时，这种治理合作显得更为紧迫。2022年12月，美欧贸易和技术委员会（TTC）发布《可信赖人工智能和风险管理评估与测量工具联合路线图》，从术语规范、标准制定以及风险监测三方面指导双方人工智能风险管理和可信赖人工智能发展，并促进相关国际标准的制定。2023年8月，金砖国家成立人工智能研究小组，以共同抵御风险、制定人工智能治理框架和标准。2023年9月，联合国教科文组织发布

《教育与研究领域生成式人工智能指南》，敦促成员国政府机构规范使用生成式人工智能技术，采取数据隐私保护和用户年龄限制等举措。2023年12月，G7数字和技术部长会议发表联合声明，发布人工智能治理国际框架《广岛进程—人工智能综合政策框架》，其中对生成式人工智能的开发者提出相关安全要求。2024年初，新加坡在世界经济论坛年会上提出生成式人工智能国际治理框架，积极寻求国际反馈。3月联合国大会一致通过了由美国牵头提出的决议案《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》，这是在ChatGPT及其他可制作照片和视频的生成式人工智能工具爆发性出现后，各国就人工智能监管采取各不相同的路径之际出炉的，也是联合国大会历史上首次为针对人工智能治理问题确立全球统一规范所通过的专项决议。

中国在人工智能技术和产业发展上已走在世界前列，同时也一直同步推进人工智能发展与治理，早在2017年国家《新一代人工智能发展规划》中便注重人工智能的技术与社会“双重属性”，强调发展与规制并重。在治理理念上，统筹发展与安全，主张全球人工智能发展应“确保有益、确保安全、确保公平”。在治理架构上，2019年成立国家新一代人工智能治理专业委员会，并在国家科技伦理委员会设立了人工智能伦理分委员会。在治理手段上，坚持倡导性原则、限制性规范和约束性法律同步同向发力，2019年便发布《新一代人工智能治理原则——发展负责任的人工智能》；2023年出台首份针对生成式人工智能的专项管理办法《生成式人工智能服务管理办法》，从价值规范、行为规制和过程监管等方面对生成式人工智能加强监管治理，发布《人工智能伦理治理标准化指南》《人工智能安全标准化白皮书（2023版）》，面向生成式人工智能等新一代技术的安全风险，推动研制风险评估和伦理治理的标准清单；同时，中国人工智能法草案已列入国务院2023年度立法工作计划。在治理合作上，2023年10月在第三届“一带一路”国际合作高峰论坛开幕式上，中共中央总书记、国家主席习近平提出《全球人工智能治理倡议》，愿同各国加强沟通交流和务实合作，共同促进全球人工智能健康有序安全发展。

六、思考与建议

2023年以来，生成式人工智能技术的发展和应用进入了前所未有的加速期，其蕴含的巨大潜力已在经济社会民生等领域快速显现。然而，这一发展路径并非没有困境，如前文所述，生成式人工智能带来一系列新增以及被放大、被增强、被恶化的伦理风险和社会发展争议，有的已经远远超越了一国一域的治理范畴。同时，在人工智能技术和社会属性动态发展演变的过程中，能否前瞻性研判预警各类风险，风险评估的技术手段能否跟上风险自身的变化，各类规制手段的可执行性、有效性等如何综合发力，以及全球治理中如何求同存异、平衡不同国家在治理理念上固有的差异性等[45]，仍然是摆在各国面前的共同难题。

中国作为负责任的大国，应从以下几方面作出积极努力、付诸更多行动。第一，在发展理念上，应坚持以人类为中心发展负责任的人工智能。政府和公共部门应加快提升自身对人工智能发展规律的理解认知和前瞻性预判，深度探讨生成式人工智能发展的目的，前瞻性的提出通过什么方式来有效控制人工智能，强化其作为人类辅助性工具的根本属性，坚定“发展与安全并重、创新与规制同行”的治理自信。第二，在推动发展鼓励创新方面，制定发布生成式人工智能应用场景负面和正面清单。尤其是对于负面清单事项，如生成虚假实验数据等，应明确惩戒标准。同时，为生成式人工智能开发与应用验证提供安全可控的“试验”环境，开展多维度风险监测评估，对进入负面清单边界的研发应用及时终止。第三，在加强治理防范风险方面，坚持法治原则，通过法律法规实现

依法治理。政府与企业、学术界等主要主体要协同共治，特别是鼓励头部企业承担更多社会责任，加快探索风险评估方法和防控措施，共同探讨如何设立合理有限的责任豁免规则等。同时，倡导参照国际《化学武器公约》《特定常规武器公约》等形成共识性约束条件。

参考文献：

[1]肖峰. 生成式人工智能与知识生产新形态[J]. 学术研究，2023(10)：50-57.
[2]王策. 生成式人工智能“生成性”的哲学考辨[J]. 学术界，2024(3)：148-157.
[3]蒲清平，向往. 生成式人工智能——ChatGPT的变革影响、风险挑战及应对策略[J]. 重庆大学学报(社会科学版). 2023(3)：102-114.
[4]EPSTEIN R. The quest for the thinking computer [J]. AAAl Magazine, 1992, 13(2): 80-95.
[5]PANDEY S. The evolution of generative ai: A journey from eliza to deep learning[EB/OL]. (2023-04-24)[2024-05-05]. https://www.linkedin.com/pulse/evolution-generative-ai-journey-from-eliza-deep-learning-pandey.
[6]GARCIA C. Harold Cohen and Aaron—A 40-year collaboration[EB/OL]. (2016-08-23) [2024-05-05]. https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration/.
[7]ZEMCIK T. A brief history of Chatbot[EB/OL].(2016-08-23)[2024-05-05]. https://computer-history. org/blog/harold-cohen-and-aaron-a-40-year-collaboration/.
[8]GARCIA C. Algorithmic music—David Cope and Emi[EB/OL]. (2015-04-29)[2024-09-11]. https://computerhistory.org/blog/algorithmic-music-david-cope-and-emi/.
[9]DESK N. The future of filmmaking is in AI [EB/OL]. (2018-06-13)[2024-09-13]. https://www.thejakartapost.com/life/2018/06/13/the-future-of-filmmaking-is-in-ai.html.
[10]谢君兰. 小冰写诗：诗歌创作的反面教材[EB/OL]. (2017-06-30)[2024-10-21]. http://www.china-writer.com.cn/n1/2017/0630/c407521-29375469.html/.
[11]NAIK N. AI copyright & human authorship: The legal battle over theatre D’opéra Spatial[EB/OL]. (2024-10-16)[2024-11-01]. https://naiknaik.com/2024/10/16/ai-copyright-human-authorship-the-legal-battle-over-theatre-dopera-spatial/.
[12]郭锐. 人工智能的伦理和治理[M]. 北京：法律出版社，2020：15-16.
[13]RAY S. Samsung bans ChatGPT among employees after sensitive code leak [EB/OL]. (2023-05-02)[2024-11-01]. https://www.forbes.com/sites/sila-dityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/.
[14]BOHANNON M. Lawyer used ChatGPT in court—and cited fake cases [EB/OL]. (2023-06-08) [2024-10-01]. https://www.forbes.com/sites/molly-bohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/.
[15]CRANE C. 20 generative AI, ChatGPT & deep-fake statistics you should know for 2024[EB/OL]. (2023-12-11)[2024-05-05]. https://www.thesslstore.com/blog/generative-ai-statistics/.
[16]WIERSON A. Around the world, AI-enabled deep fakes are poised to upend elections[EB/OL]. (2024-03-18)[2024-11-01]. https://www.newsweek.com/around-world-ai-enabled-deep-fakes-are-poised-upend-elections-opinion-1879810.
[17]郭如愿. 论人工智能生成内容的信息权保护[J]. 知识产权，2020(2)：48-57.
[18]UniFans Content Team. Can ChatGPT write porn? Learn the capabilities & limitations of AI [EB/OL]. (2023-05-02)[2024-11-03]. https://www.unifans.io/blog/can-chatgpt-write-porn.
[19]Nature. Tools such as ChatGPT threaten transparent science; here are our ground rules for their uses [EB/OL].(2024-01-01)[2024-11-28].https://www.nature.com/articles/d41586-023-00191-1.
[20]SAJID H. Social impact of generative AI: Benefits and threats[EB/OL]. (2024-01-01)[2024-11-24]. https://www.unite.ai/social-impact-of-generative-ai-benefits-and-threats/.
[21]THROP H. ChatGPT is fun, but not an author[J]. Science, 2023, 379(6630): 313.
[22]BALLESTE P L. Open science and software assistance: Commentary on “artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora’ s box has been opened” [J]. Journal of Medical Internet Research, 2023(25): e49323.
[23]ATHALURI S A, MANTHENA S V, KESAPR AGA V S R K M, et al. Exploring the boundaries of reality: Investigating the phenomenon of artificial

intelligence hallucination in scientific writing through ChatGPT references[J]. Cureus, 2023, 15(4): e437432.
[24]BHATTACHARYYA M, MILLER V M, BHATTACHARYYA D, et al. High rates of fabricated and inaccurate references in ChatGPT-generated medical content[J]. Cureus, 2023, 15(5): e39238.
[25]HANSON I. AI's data fakery is 'scary' say researchers, but the problem is already huge [EB/OL]. (2023-12-05)[2024-05-05]. https://www.medicaldevice-network.com/features/ai-medical-data-fakery-scary-say-researchers-but-problem-already-huge/?cf-view.
[26]COPE. Image manipulation[EB/OL]. (2023-10-06)[2024-05-05]. https://publicationethics.org/publication-integrity-week-2023/image-manipulation.
[27]温德尔·瓦拉, 赫科林·艾伦. 道德机器: 如何让机器人明辨是非[M]. 王小红, 等, 译. 北京: 北京大学出版社, 2017: 223-250.
[28]DILMEGANI C. Generative AI ethics in 2024: Top 6 concerns[EB/OL]. (2024-01-02) [2024-05-05]. https://research.aimultiple.com/generative-ai-ethics/.
[29]HAL Stanford University. Artificial intelligence index report 2024[EB/OL]. (2024-04-26)[2024-08-10]. https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf.
[30]莫宏伟, 徐立芳. 人工智能伦理导论[M]. 西安: 西安电子科技大学出版社, 2022: 212.
[31]卡尔·马克思. 马克思恩格斯文集: 第5卷[M]. 北京: 人民出版社, 2009: 495.
[32]闫坤如. 人工智能的道德风险及其规避路径[J]. 上海师范大学学报(哲学社会科学版), 2018(2): 40-47.
[33]刘益东. 数字反噬、通能塔诅咒与全押归零的人工智能赌局——智能革命重大风险及其治理问题的若干思考[J]. 山东科技大学学报(社会科学版), 2022(6): 1-13.
[34]World Economic Forum. Jobs of tomorrow: Large language models and jobs[EB/OL]. (2023-09-18) [2024-05-05]. https://www3.weforum.org/docs/WEF_Jobs_of_Tomorrow_Generative_AI_2023.pdf.
[35]GMYREK P, BERG J, BESCOND D. Generative AI and jobs: A global analysis of potential effects on job quantity and quality[EB/OL]. (2023-08-01) [2024-05-05]. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@dgreports/@inst/documents/publication/wcms_890761.pdf.
[36]李伦, 孙保学. 给人工智能一颗"良芯(良心)"——人工智能伦理研究的四个维度[J]. 教学与研究, 2018(8): 72-79.
[37]UNESCO. Guidance for generative AI in education and research[EB/OL]. (2023-09-07)[2024-05-05]. https://unesdoc.unesco.org/ark:/48223/pf0000386693.
[38]IEA. World energy outlook 2024[EB/OL]. (2024-10-01)[2024-11-25]. https://iea.blob.core.windows.net/assets/4630e9cd-afda-4722-900e-8c90fc23f71d/WEO2024_Executivesummary_Chinese.pdf.
[39]VRIES A. The growing energy footprint of artificial intelligence[J]. Joule, 2023, 7(10): 2191-2194.
[40]HAL Stanford University. Artificial intelligence index report 2023[EB/OL]. (2023-04-15)[2024-05-05]. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023. pdf.
[41]周翔, 霍金. AI或许能根除疾病和贫穷, 但也可能摧毁人类[EB/OL]. (2017-04-27)[2024-11-25]. https://www.leiphone.com/category/ai/fHCfKVMwc2zNhIed.html.
[42]KISSINGER H A. How the enlightenment ends [EB/OL]. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/, 2024-05-05.
[43]European Parliament. Artificial intelligence act [EB/OL]. (2024-03-18)[2024-05-05]. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN. pdf.
[44]乌尔里希·贝克, 约翰内斯·威尔姆斯. 自由与资本主义[M]. 路国林, 译. 杭州: 浙江人民出版社, 2001: 127.
[45]王彦雨, 李正风, 高芳. 欧美人工智能治理模式比较研究[J]. 科学学研究, 2024(3): 460-468.

【责任编辑 邱佛梅 苏聪文】

(下转第148页)

（上接第127页）

Social Controversies and Governance of Generative Artificial Intelligence

GAO Fang, *WANG Yanyu & WANG Yiying*

Abstract: As a branch of artificial intelligence technology, generative artificial intelligence not only has the ethical risks inherent in traditional artificial intelligence, but also brings new social problems and risks, posing new challenges to the existing governance system. This article systematically sorts out the stage characteristics of generative artificial intelligence based on the development history of artificial intelligence for more than 60 years, analyzes various ethical and social development disputes caused by the current stage, and sorts out the governance practices of the main governance entities. Research results show that the unique attributes of generative artificial intelligence—such as creativity, technical generalization, reality reconstruction, and human-likeness—lead to more concealed ethical issues, including privacy breaches, social distrust, academic fraud, discrimination, and prejudice. They pose both short-term and long-term challenges to human development. In the short term, these challenges affect employment structures, the digital divide, and energy consumption. In the long term, they may lead to shifts in social organization and even the potential "loss of control" over artificial intelligence systems, which could threaten human survival. Major countries and multilateral international organizations strengthen the governance of generative artificial intelligence through diverse practices such as mechanism construction, legislative supervision, standard setting, and governance cooperation, which provide valuable references for enhancing our country's ability to balance development and safety, as well as innovation and regulation, and further improve the artificial intelligence safety supervision system.

Keywords: generative artificial intelligence; large AI models; ethical risk; social controversy; governance; science and technology ethics