

人工智能生成内容的潜在风险及治理对策

李亚玲^{1,2}, 覃缘琪^{1,2}, 魏阙³

(1. 之江实验室智能社会治理研究中心, 浙江 杭州 311121;

2. 浙江省哲学社会科学试点实验室——之江实验室智能社会治理实验室, 浙江 杭州 311121;

3. 之江实验室发展研究中心, 浙江 杭州 311121)

摘要: 人工智能生成内容技术正重塑数字内容的生产方式和消费模式。基于人工智能生成内容技术的基本概念和技术发展历程, 重点研究了人工智能生成内容引发的数据泄露、算法霸权、数字鸿沟、知识产权保护、就业冲击等潜在风险和治理挑战。在梳理了国外发达国家的典型治理路径和治理经验后, 概括提出了我国面对人工智能生成内容技术引发的潜在风险的应对措施和发展建议。具体包括坚持安全与发展并重, 做好统筹布局; 完善人工智能技术治理规则体系; 加强人工智能技术研发全要素投入; 加快人工智能治理技术研发; 构建全方位多层次的技术治理体系。

关键词: 人工智能生成内容; 风险; 技术治理; ChatGPT

中图分类号: T-01

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202332

Potential risks and governance strategies of artificial intelligence generated content technology

LI Yaling^{1,2}, QIN Yuanqi^{1,2}, WEI Que³

1. Research Center for Intelligent Society and Governance, Zhejiang Laboratory, Hangzhou 311121, China

2. Zhejiang Labs for Philosophy and Social Sciences: Laboratory for Intelligent Society and Governance in Zhejiang Laboratory, Hangzhou 311121, China

3. Center for Science and Technology Strategy, Zhejiang Laboratory, Hangzhou 311121, China

Abstract: Artificial intelligence generated content technology is reshaping the production and consumption patterns of digital content. The research was based on the basic concepts and technological development process of artificial intelligence generated content technology, focusing on exploring ethical risks and governance challenges such as privacy leakage, algorithm hegemony, digital divide, intellectual property protection and employment impact caused by artificial intelligence generated content. After reviewing the typical governance paths and experiences of developed countries abroad, this paper summarized and proposed measures and development suggestions for China to address the potential risks of AI generated content. One was to prioritize development and coordinate layout; Second was to strengthen the full factor investment in the research and development of artificial intelligence technology; Third was to improve the governance rule system of artificial intelligence technology; Fourth was to accelerate the research and development of artificial intelligence technology.

收稿日期: 2023-04-28; 修回日期: 2023-07-25

通信作者: 魏阙, weiq@zhejianglab.com

基金项目: 国家重点研发计划 (No.2022YFC3303103); 浙江省科技厅软科学计划研究项目 (No.2023C25074); 中共浙江省委政法委员会 浙江省法学会 2023 年度法学研究课题重点项目 (No.2023NA02); 浙江省科学技术情报学会科研项目 (No.2023qbxh006); 杭州市余杭区社会科学研究常规资助课题 (No.Yhsk23C01)

Foundation Items: The National Key Research and Development Program of China (No.2022YFC3303103), Zhejiang Provincial Department of Science and Technology Soft Science Research Program (No.2023C25074), Research Project of Zhejiang Law Society (No.2023NA02), Research Project of Zhejiang Society for Scientific and Technical Information (No.2023qbxh006), Social Science Research Project of Yuhang District (No.Yhsk23C01)

gence governance technology.

Key words: artificial intelligence generated content, risk, technology governance, ChatGPT

0 引言

近年来,随着核心算法的突破、计算能力的提升和海量数据的支撑,人工智能技术得以迅速发展,并推动着社会朝着更加智能化的方向加速跃升。我国高度重视人工智能的布局和发展。习近平总书记强调,人工智能是引领新一轮科技革命和产业变革的战略性技术,具有溢出带动性很强的“头雁”效应,要深刻认识加快发展新一代人工智能的重大意义,加快发展新一代人工智能。随着 ChatGPT 的爆红,人工智能生成内容技术作为人工智能中极具发展前景的细分领域,再次引发了各界的广泛关注。在当前数字世界和物理世界加速融合的大背景下,人工智能生成内容技术是未来全面迈向数字文明新时代不可或缺的支撑力量,在论文写作、医疗、传媒、金融、工业机器人以及互联网服务等众多领域和场景中被广泛应用。这种应用提高了效率并降低了成本,但同时也带来了一些挑战。人工智能内容生成系统的自主性逐渐替代了人类决策,有时这种变化不仅未能解决现有问题,甚至引发了新的问题。这些问题成为限制人工智能生成内容技术推广的重要因素,论文写作领域是其中最典型的例子。由于社会对论文写作的巨大需求以及技术的不断成熟,该领域已成为全球研究和发展的热点。然而,论文写作技术带来的伦理问题也受到了人们的关注。现有的法律和政策难以适用于论文写作技术,新的政策又缺乏明确性,这给技术的推广带来了困难。此外,由于技术具有的算法黑箱性、数据依赖性等特性,引发了数据泄露、算法霸权、数字鸿沟、知识产权保护、就业冲击等潜在风险和治理挑战。面对人工智能生成内容技术所带来的潜在风险,人们迫切需要进行广泛而普遍的探讨,并在此基础上找到路径和制定规范,以确保人工智能生成内容的良性发展。

针对上述风险内容的治理,2023年4月,国家互联网信息办公室(以下简称国家网信办)发布《生成式人工智能服务管理办法(征求意见稿)》,明确了对人工智能生成内容的监管细则。如何促进人工智能生成内容技术的发展与应用,同时防范其风险,推动技术沿着有利于人类福祉的方向和谐发

展,将成为亟待思考和解决的重要命题。

已有部分研究者针对人工智能生成内容引发的风险和治理对策展开了研究。郁建兴等人^[1]从生成式大模型带来的巨大变革出发,提出大模型治理急需关注的问题,强调有效整合行政和技术手段以确保生成式大模型风险可控,推动其造福人类。张凌寒等人^[2]探讨了生成式人工智能引发的治理基础变化,提出了针对生成式人工智能的敏捷治理范式,强调治理的全面性、适应性与灵活性。宋恺等人^[3]梳理了生成式人工智能引发的个人信息泄露、虚假信息泛滥、仿制服务涌现、治理挑战加剧等问题,提出了以统筹发展与规范为核心的治理策略。于水等人^[4]从伦理、法律规制和技术3个层面提出了生成式人工智能的风险治理路径。张欣^[5]针对生成式人工智能的多维属性和产业链特点,提出应以数据解释机制为核心,构建精准多元的数据主体责任矩阵,打造灵活高效的数据治理监管工具体系。总体而言,现有的研究多集中于对生成式人工智能风险的应用层面或产业层面的分析,缺乏从技术视角对生成式人工智能引发风险的归纳;此外,关于国外生成式人工智能治理模式的比较研究较为缺乏。本文从人工智能生成内容的技术逻辑出发,梳理归纳出其引发的潜在风险;通过对国外人工智能治理模式的比较研究,填补上述空白,并最终提出我国对人工智能生成内容技术治理的对策建议。

1 人工智能生成内容的技术逻辑与治理起点

1.1 人工智能生成内容的概念与分类

人工智能生成内容(artificial intelligence generated content, AIGC)技术是指从人类提供的指令中提取和理解意图信息,并根据提取的意图信息来生成内容,包括文本、图像、音频、视频等^[6]。人工智能生成内容使得内容创作过程更加高效,能够在短时间内迅速生产大量高质量内容。根据接收的指令和生成内容的模态是否一致,生成模型可以分为两类:单模态模型和多模态模型。单模态模型接收与生成内容模态相同的模态指令,而多模态模型接收跨模态指令并产生不同模态的结果。典型的人工智能生成模型梳理如表1所示。

表1

常见的人工智能生成模型

模型类别	模型名称	发布机构	发布时间	应用领域
单模态	LaMDA	Google	2021年5月	文本生成、对话
	ChatGPT	Open AI	2022年11月	文本生成、对话
	LLaMA	Meta	2023年2月	文本生成、对话
	Alpaca	斯坦福大学	2023年3月	文本生成、对话
	文心一言	百度	2023年3月	文本生成、对话
	星火	科大讯飞	2023年5月	文本生成、对话
	StyleGAN	NVIDIA	2018年12月	图像生成
多模态	DVD-GAN	DeepMind	2019年7月	视频生成
	DALL·E 2	Open AI	2022年7月	从文本生成图像
	Stable Diffusion	Stability.AI	2022年8月	从文本生成图像
	GPT-4	Open AI	2023年3月	支持图片和文本的多模态输入
	DeepFloyd IF	Stability.AI	2023年4月	从文本生成图像
	ImageBind	Meta	2023年5月	打通视觉、红外图像、文本、音频、深度信息、动作捕捉6种模态

1.2 人工智能生成内容的技术发展历程

人工智能生成模型有着悠久的历史，起源可追溯到20世纪50年代的隐马尔可夫模型^[7]和高斯混合模型^[8]，这些模型被用于生成简单的语音序列和时间序列数据，但性能一般。

生成模型的第一次飞跃发展源于深度学习算法的应用。2006年，深度学习算法取得重大突破，生成模型的性能有了显著提高。早期的深度生成模型只能处理较短的序列。例如，在自然语言处理中，传统的句子生成方法是使用N-gram语言模型^[9]学习单词分布，然后搜索最佳序列。在引入递归神经网络^[10]后，算法能够捕捉到相对较长的依赖性，因而可处理更长的句子。随着长短期记忆^[11]和门控循环单元^[12]模型的发展，利用门控机制在训练过程中控制记忆，能够单次处理长约200个词的样本。在计算机视觉领域，传统的图像生成算法主要使用纹理合成^[13]和纹理映射^[14]等技术，难以生成复杂和多样化的图像。随着深度学习算法的应用，图像生成领域迅速发展。2014年，生成对抗网络技术^[15]被首次提出，在图像生成应用中取得了令人惊叹的效果，是视觉生成领域的重要里程碑。变分自编码器、扩散生成模型^[16]等技术的发展和应用，实现了对图像生成过程更加细粒度的控制，进一步提高了生成图像的质量。

生成模型的第二次飞跃发展源于Transformer架构的提出和应用。2017年，Vaswani等人^[17]首次将Transformer架构用于自然语言处理任务，随后Transformer被应用于计算机视觉等更多领域。在

计算机视觉领域，研究者将Transformer架构与视觉组件相结合，形成Vision Transformer^[18]和Swin Transformer^[19]等模型，使其能够应用于图像领域。除了应用于单模态的模型外，Transformer架构也被用于将不同领域的模型融合在一起，以执行多模态任务^[20]。总之，基于Transformer架构模型的出现彻底改变了人工智能生成内容技术。与之前相比，人工智能生成内容技术得以在更大的数据集上训练更复杂的生成模型，使用更大的基础模型体系结构，以及获得广泛的计算资源的结果。例如，GPT-3的主框架与GPT-2保持一致，但预训练数据量从38 GB增长到570 GB，基础模型大小从1.5 B增长到175 B。因此，GPT-3在人类意图提取等多项任务上的泛化能力优于GPT-2^[21]。

除了数据量增加和计算能力提高带来的性能提升外，也有研究人员开始探索在这些模型的基础上引入新技术。例如，在自然语言处理任务中，ChatGPT利用来自人类反馈的强化学习^[22-23]来确定给定指令的最合适回答，使得ChatGPT在长对话中更好地理解人类的偏好，从而提高模型的可靠性和准确性。在计算机视觉任务中，稳定扩散模型^[24]于2022年被提出。该模型通过控制探索 and 开发之间的平衡来帮助生成高分辨率图像，从而使生成图像的多样性和与训练数据的相似性和谐结合。未来，随着人工智能生成内容技术在各个领域的不断应用和发展，将会有越来越多的技术被引入，使这一领域不断进步。

从人工智能生成内容技术的发展历程可以发

现, 该技术除了具有人工智能算法的固有特点(如算法黑箱特性)外, 还具有其独特之处, 如数据依赖性、交互性强、算力壁垒高等特点。这些特点相应地会引发与以往人工智能技术不同的风险。例如, 由于生成式模型与用户间交互性较强, 用户频繁输入信息并获得反馈, 因此数据泄露风险比一般模型要高。此外, 由于训练算力和推理算力成本较高, 生成式模型的研发容易被寡头垄断, 从而加剧数字鸿沟。后文将对人工智能生成内容引发的潜在风险进行详细分析和介绍。

2 人工智能生成内容引发的潜在风险

对 ChatGPT 安全的担忧是目前争议的焦点。以 ChatGPT 为代表的生成式人工智能技术在网络安全领域是一把“双刃剑”, 一方面, ChatGPT 可以赋能网络安全, 提升风险检测与防护能力; 另一方面, ChatGPT 技术的滥用将为国家、社会、公众带来负面影响, 同时, AI 新技术的脆弱性也使得 ChatGPT 的应用面临多种潜在风险。

2.1 数据/信息泄露风险

移动互联网的普及导致了个人信息被不断收集和使用, 人工智能生成模型中的数据泄露风险主要集中在训练数据收集和生成模型使用过程中。随着移动互联网的发展普及, 个人信息正在不断地被收集和使用。人工智能生成模型中的信息泄露风险主要包括以下两种情况。

(1) 训练数据收集时的信息泄露风险。由于生成模型是在大规模网络抓取数据上训练的, 需要采集大量、多样的数据, 使用目的以外的数据收集可能导致信息泄露。例如, ChatGPT 训练的数据是在未经适当同意的情况下从帖子、网站、文章、书籍和个人信息中系统地抓取的。如果其中涉及个人信息, 或是在未经同意的情况下使用, 可能涉及侵犯隐私。

(2) 生成模型中的信息泄露风险。首先, 有研究表明, 大型语言模型更容易受到隐私攻击^[25]。攻击者可从训练模型中生成序列, 反过来识别出训练数据集中的训练序列。其次, 部分生成模型存在训练数据记忆问题, 导致信息泄露。例如, 稳定扩散模型 Stable Diffusion、DALL·E 2 模型都被报道称在训练数据中记忆图像, 并在生成图像中复现^[26]。这种图像反流与图像在数据集中被多次复制有关。类似地, ChatGPT 在生成回答中也存在信

息泄露现象^[27]。最后, 生成模型在使用过程中存在过度收集用户信息的行为, 从而导致信息泄露。例如, 根据 ChatGPT 的隐私政策, 除了对话信息外, 它还收集用户与网站的交互数据、浏览器设置及其类型、IP 地址等信息。

2.2 算法霸权风险

人工智能生成模型的算法公正性和客观性受依赖数据和算法黑箱等因素影响, 导致算法歧视和算法滥用等算法霸权行为, 如训练数据集中的偏见、虚假信息的传播、剽窃和不当行为及过度依赖等。

生成模型的算法歧视主要与训练数据集中隐含的偏见或歧视信息有关。由于生成模型中使用的训练数据是在现实世界中收集的, 因此它们可能会无意中强化有害的刻板印象, 排斥或边缘化某些群体, 并包含有毒的数据源, 这些数据源可能会煽动仇恨或暴力并冒犯个人。例如, 用于训练扩散模型的 LAION 数据集^[28]因包含了社会刻板印象、色情、种族主义辱骂和暴力等相关的问题内容而受到批评。

生成模型中的算法滥用主要包括以下 3 个方面。(1) 虚假、有害信息的传播。ChatGPT 可能被滥用于生成并传播虚假、有害信息。例如, ChatGPT 可能被用于生成色情、低俗和暴力内容, 或是被黑客用于制作病毒或木马等恶意软件。ChatGPT 还可能被用来生成大量捏造的文章, 这些文章出现在博客、新闻、报纸或互联网上, 看起来与其他文章无法区分, 但实际上是假的。这种虚假信息的广泛传播不仅损害公众利益, 还会破坏网络环境, 将对公共信息安全造成灾难性后果。(2) 剽窃和不当行为。例如, ChatGPT 在学术上的滥用。学生可以使用 ChatGPT 生成的内容通过考试并提交学期论文。研究人员可能会使用 ChatGPT 生成的内容提交论文, 并隐瞒 ChatGPT 的使用。许多学校已经禁止使用 ChatGPT。(3) 过度依赖。ChatGPT 可以像搜索引擎一样根据问题迅速生成建设性的答案。这种毫不费力的指导可能会逐渐削弱使用者的批判性思维和独立思考的能力。

2.3 数字鸿沟风险

生成模型能提高知识传播效率, 但高成本和专业性可能导致数字鸿沟加剧。数字鸿沟指由不同性别、年龄、收入、阶层的人在接近、使用新信息技术的机会与能力上的差异造成不平等进一步扩大状况。随着移动互联网的发展, 地区差异、代际差

异造成的数字鸿沟现象日益凸显^[29]。

尽管生成模型可以降低知识学习的门槛，提高知识传播的效率，但也可能使知识传播的差距被进一步放大，导致数字鸿沟加剧。一方面，大规模模型训练和运行所需要的算力成本和数据成本较高。由于生成模型变得越来越复杂，参数数量和计算复杂度呈指数增长，需要GPU和TPU等高性能芯片来进行大规模模型的训练和推理。以GPT-3为例，GPT-3有1 750亿个参数，其推理运行时需要2 048个CPU和2 048个GPU的超级计算机集群，模型1天的运行成本约4 000美元。高昂的算力成本使得大型模型只能被少数企业掌握，将进一步加剧生成模型领域的垄断情形。这可能导致内容生产工具的核心技术被一小部分人主导，造成权力和影响力的集中，从而加剧世界的贫富分化，扩大现有的不平等。另一方面，低收入国家可能会因为数字基础设施落后而难以公平地触达生成模型等先进人工智能工具和技术，从而进一步拉开与发达国家在知识传播上的现有差距。

2.4 知识产权风险

人工智能生成内容的版权和知识产权问题尚未得到解决，存在统一法律认识、难以识别训练数据、数据规定不明确和利益分配机制尚未建立等难点。同时，生成模型可能引发训练数据侵权和生成内容侵权等知识产权风险。随着人工智能生成内容技术的不断发展，对人工智能生成内容所有权和保护已经引起了大量的关注和讨论。根据现行法律，版权法一般保护由人类作者创作并以有形形式固定的原创作品。人工智能生成的内容是否应被视为有资格受到版权保护的原创作品，目前尚且处于争议中^[30]。厘清人工智能生成内容的版权争议尚且存在以下难点。（1）全球对人工智能生成内容的版权问题缺乏统一的法律认识，所有权争议仍未解决；（2）难以识别用于训练生成模型的所有原创作品。理论上这些模型可以生成无限量的内容，因此难以做到对所有生成内容进行测试和识别；（3）对数据收集、使用、权利确认和数据商业使用的规定不明确；（4）模型和数据提供方的利益分配机制尚未建立。

除了版权争议外，生成模型可能引发的知识产权风险有以下两类。

（1）训练数据的侵权风险。生成模型在创作过程中，学习了大量人类作品作为训练数据，这些训练数据如果未能得到原创作品权利人的授权，就会有

知识产权侵权风险。例如，Stable Diffusion模型是在未经知识产权所有者批准的情况下，对来自互联网的数十亿张图像进行训练的，可能会侵犯原作者的知识产权。尽管Stable Diffusion拒绝对生成的图像拥有任何所有权，只要图像内容合法且无害，就允许用户自由使用它们，但这种自由引发了关于所有权归属的争议。

（2）生成内容的侵权风险。如生成内容直接复制了部分或全部包含知识产权的内容。例如，2022年11月，Matthew Butterick对微软、GitHub和OpenAI向加利福尼亚州旧金山的美国联邦法院提起集体诉讼，指控GitHub旗下的代码生成服务Copilot非法使用互联网上的代码而未注明出处，违反了版权法。尽管微软承认Copilot是在公共GitHub存储库中接受开源软件培训的，但其声称Copilot的输出只是一系列代码建议，并没有对这些建议赋予任何权利，也不保证所生成代码的正确性、安全性或版权。

2.5 就业冲击风险

人工智能生成内容技术可能导致部分人群失业，但同时也会创造新的产业和就业机会，社会失业率会趋向于稳定，但可能存在两极分化现象。人工智能生成内容技术可以迅速生成高质量内容，因而可能会导致部分人群的失业问题，进而引发人们对社会就业结构变化的担忧。麦肯锡全球研究院（McKinsey Global Institute）预测称，若自动化技术被高速采纳，到2030年，将有30%的工作将会被自动化技术替代，即使在中速采纳下，也将会有15%的工作会被自动化技术所取代，其中中国受影响最大，预计会有1亿人需要转换职业（高速采纳下）。

AIGC也许不会对社会就业结构造成根本上的冲击，但是会削减劳动力在资本谈判中的议价能力，使部分人群成为“无用阶级”。在取代一部分职业和一部分人群的同时，AIGC将大幅提高全社会的生产效率，也会创造新的产业，从而出现新的就业机会。整体上看，社会失业率会趋向于稳定，走向“人-机”协同工作，但是也可能存在两级分化现象，能适应就业结构变化的人群和逐渐失去生产能力的人群。

3 主要治理路径

人工智能技术的治理一直是全球范围内的焦点

议题。全球范围内针对人工智能技术的治理已有较为成熟的规则体系和实践探索,欧盟采取强监管路径,美国采取包容创新监管路径,其他国家也出台了相关政策文件。通过对国外主要发达国家治理模式的比较和研究,总结其先进经验,可以为我国人工智能技术的治理提供参考和借鉴。

3.1 欧盟人工智能技术治理的路径选择

欧盟在人工智能监管领域起步较早。其主要特征是构建完善的法律框架和制度体系,确保人工智能技术朝着有益于人类社会的方向发展。2018年,欧盟发布了《欧盟人工智能》,提出制定人工智能新的伦理准则,以解决公平、安全和透明等问题,捍卫欧洲价值观。随后,欧盟还成立了人工智能领域的高级专家小组,并于2019年发布了《人工智能的伦理准则》,提出构建可信AI框架,重点论述了可信AI的基本原则、关键要求和评估方法。2021年4月,欧盟委员会发布了欧洲议会和理事会《关于制定人工智能统一规则》的立法提案(又称《人工智能法案》),旨在规范AI技术风险并加强全欧洲对AI技术的利用、投资和创新。《人工智能法案》是欧盟第一部监管人工智能系统的横向立法,也是全球第一个人工智能法律监管框架。法案按照风险级别将AI技术及其应用分为4类:(1)具有不可接受风险的技术,包括鼓励未成年人危险行为的语音系统、政府“社会评分系统”等操纵人类行为的AI系统或应用程序,该类技术被禁止;(2)高风险技术,包括用于运输、教育、辅助手术、招聘、移民管理等事项的技术,该类技术在进入市场前需进行风险评估、使用高质量数据集测试、提供详细的记录文件等检验保护措施,其中具有远程生物特征识别的系统使用前需司法授权并受时间和地理位置的限制;(3)聊天机器人等风险有限的系统,需履行透明披露义务;(4)视频游戏等风险最小的应用程序,法规不予以干预。针对生成模型,法案要求其必须遵守额外的透明度要求,例如需披露哪些内容是由人工智能生成的,如果训练数据受版权保护则需要披露数据摘要。2023年6月,该法案经欧洲议会投票通过,但最终生效还需要欧盟成员国逐一审核通过。除了欧盟层面的探索外,已有部分欧盟成员国针对ChatGPT等人工智能生成技术开展监管举措。例如,2023年3月,意大利个人数据保护局表示,已就OpenAI聊天机器人ChatGPT涉嫌违反数据收集规则展开调查,即日起

禁止使用ChatGPT,并暂时限制OpenAI处理意大利用户数据。

总体而言,欧盟对人工智能的监管较为严格,尤其是个人数据的保护方面。欧盟采取较为严格的监管框架与其在人工智能技术领域的落后有关。欧盟希望以较为严格的监管政策体系规范外国技术的应用,在输出符合其价值观的监管体系的同时,促进本地创新企业的发展,从而夺回数字主权。

3.2 美国人工智能技术治理的路径选择

美国在人工智能治理上,以鼓励技术创新为主,强调技术的发展应符合美国价值观。2019年10月,美国发布了全球首份军用人工智能的伦理原则。2020年1月,白宫发布的《人工智能应用的监管指南》中提出了10项监管原则,强调对人工智能的风险进行评估和管理,通过充分地披露保证技术应用的透明性,鼓励公众参与规则制定,以建立公众对AI的信任。2021年1月,白宫成立了国家人工智能计划办公室,作为监督和实施美国国家人工智能战略的专门机构,以确保未来美国在人工智能领域保持领导地位。2022年10月,白宫发布了《人工智能权利法案蓝图》,确定了5项原则以保护人工智能时代美国公众的利益。5项原则分别为:(1)建立安全和有效的系统;(2)避免算法歧视,以公平的方式使用和设计系统;(3)保护数据隐私;(4)系统的通知和解释要清晰、及时和可访问;(5)设计自动系统失败时使用的替代方案、考虑因素和退出机制。2023年1月,美国国家标准与技术研究院(NIST)正式公布《人工智能风险管理框架》,旨在指导机构组织在开发和部署人工智能系统时降低安全风险,避免产生偏见和其他负面后果,提高人工智能可信度,保护公民的公平自由权利。针对生成模型的监管,美国采取了包容审慎的态度,美国商务部在同年4月公开征求意见,讨论新人工智能模型在发布前是否应经过认证程序等内容。

总体而言,美国对人工智能技术的监管较为宽松,以促进技术创新为主。美国在人工智能技术,尤其是生成模型的发展上位居世界前列,因而整体更倾向于通过构建自愿适用的指引、框架或标准等非监管方法对人工智能应用采取软治理。

3.3 其他国家人工智能技术治理的路径选择

其他国家针对人工智能治理领域也出台了相关政策文件。例如,英国在2023年3月发布了《支持创新的人工智能监管方式》白皮书,提出了AI治

理的5项原则，包括安全性和稳健性、透明度和可解释性、公平性、问责制和管理，以及可竞争性；同年4月英国政府宣布投入1亿英镑建立基础模型工作组，以重点推进人工智能安全研究。日本在2019年成立了“以人为本的人工智能社会原则委员会”，强调人工智能技术的发展应遵循以人为本的原则。新加坡在2019年提出了《人工智能监管模式框架》，为企业采用AI技术可能产生的风险提供监管指导。澳大利亚在2019年发布了《人工智能：澳大利亚的伦理框架》讨论稿。巴西在2021年成立国家个人数据和隐私保护委员会，推动制定数据保护规则应用指南。俄罗斯在2021年发布法律规定加强运营商对公民互联网数据的保护，降低隐私泄露风险。智利在2021年对人脑神经技术进行立法规范，成为世界上第一个在该领域立法的国家。

总体而言，在世界范围内，针对人工智能技术的治理已有较为成熟的规则体系和一些实践层面的探索。较为典型是欧盟的强监管路径和美国的包容创新监管路径。上述规则和经验为我国人工智能技术的治理提供了参考和借鉴。

4 对策与建议

我国高度重视人工智能技术的治理，积极构建有利于技术健康有序发展的体制机制。2020年11月，习近平主席在G20峰会上指出，愿同各方探讨并制定全球数字治理规则，推动落实二十国集团人工智能原则，引领全球人工智能健康发展。2022年12月，国家网信办、工信部、公安部联合发布了《互联网信息服务深度合成管理规定》，对深度合成技术做出规范。2023年4月，国家网信办发布了《生成式人工智能服务管理办法（征求意见稿）》，对生成式人工智能服务商的准入、算法设计、训练数据选择、模型、内容，以及用户个人隐私、商业秘密等方面做出了具体规定。

以ChatGPT为代表的、以大模型为基础的人工智能生成内容技术具有内容生成质量高、服务中心化程度高、数据再加工能力强和过程可解释性差等特点。面对这些技术特点所引入的智能技术治理的复杂性，建议我国在完善智能技术治理框架和规则体系的基础上，加强人工智能核心技术与智能治理技术全要素系统性投入，构建多层次全方位的技术治理体系，牢牢掌握国家安全主动权。

4.1 兼顾安全与发展，做好统筹布局

积极倡导安全与发展并重的原则，统筹做好前瞻性布局。抢抓新一轮科技革命和产业变革的重大机遇，催生新发展动能，以技术引领变革、以创新驱动发展、以高端实现跨越。一是要做好系统性前瞻布局，重视人工智能生成内容技术发展过程中出现的趋势性、苗头性、倾向性问题，提前规划和治理，同步推进研发攻关、产品应用和产业培育，大力推动创新链和产业链深度融合，以技术突破推动领域应用和产业升级。二是要遵循集中力量办大事、重点突破的原则，针对我国人工智能生成技术发展的迫切需求和薄弱环节，以重大项目为牵引，明确任务边界和研发重点，形成以重大科技项目为核心、现有研发布局为支撑的人工智能项目群。三是要将安全应用作为发展目标，以与人民生活息息相关的真实场景为导向，有针对性地研发具有重大引领带动作用的相关产品和服务，积极培育人工智能新兴业态。完善多元应用场景与技术的衔接和适配，加快人工智能生成内容关键技术的转化应用。在引入教育、政府、医疗等特殊场景时，需确保数据和隐私安全，并建立反馈、纠偏和问责制度。

4.2 完善人工智能技术治理规则体系

尽管我国已相继出台一系列政策法规对人工智能技术应用过程加以规范，但是在具体规则和实施落地等方面还需要进一步推进完善。一是尽快研究制定人工智能风险管理的总体框架。对风险分析、可信度考量、风险管理等流程进行细化规定，形成人工智能风险管理规范，为企业、高校、科研院所等机构组织提供风险管理的指南和标准。二是结合人工智能生成内容的技术特性，及其在不同领域的应用情况，对可能引发的法律风险进行预研预判，并做好相应对策。开展针对智能生成内容权属认定等方面的法律法规研究。厘清模仿、抄袭与合理使用的认定标准，探索侵权主体及侵权责任的认定路径，为相关资源投入方提供平衡的保护。遵循谨慎包容的理念，鼓励立法研究、行业监管等领域的多方参与。加强法律法规和监管治理的科学化、精细化与敏捷化。三是针对国外先发的人工智能先进技术，探索技术引进路径与监管方法。规范各类企业应用国外技术平台开展技术服务的方式和路径。

4.3 加强人工智能技术研发全要素投入

不论是面对技术本身特点带来的治理复杂性，

还是面对先进人工智能技术由国外率先掌握带来的国家安全问题,坚持人工智能技术的创新发展,牢牢把握人工智能技术的发展方向,永远是解决人工智能技术治理问题的核心关键。因此,面对目前人工智能技术大模型发展趋势,笔者给出两点建议。一是加强人工智能核心技术研发的组织力度。统筹政企学研等研发资源,在现有优势技术团队的基础上,组织成立国家级的人工智能技术研发机构,持续深入开展人工智能核心技术研究。以核心技术攻关的体制机制来推动人工智能领域关键核心技术的突破。二是加强数据、算力等人工智能研发全要素投入。在数据层面,积极推动构建高质量数据集。建立可信安全的公共数据共享开放机制,推动政府公共数据的深度开发。加快构建我国主导的高质量科研数据集,倡导科研数据的共享开放。加快完善数据标准体系,降低数据流通和使用成本。重视数据隐私和数据安全,确保数据采集过程合法合规,数据使用安全可控。在算力方面,加快算力基础设施建设,保障算力资源的多元供给。强化算力的智能调度,提升算力网络的支撑能力,支持和鼓励国家超级计算中心、区域级计算中心对重大项目的研发,并予以算力资源倾斜。

4.4 加快人工智能治理技术的研发

为系统防范人工智能生成内容技术应用的潜在风险,在完善的机制规范基础上,需要围绕内容审查、应用监管、风险研判、智能对抗等领域建立人工智能治理的技术体系。一是在国家人工智能技术领域重点项目上配套治理技术与技术风险评估的研究,加快形成人工智能治理技术与基础技术协同发展形势。二是围绕人工智能全流程治理,在研发过程管理、信息内容鉴别、数据集与模型伦理与法律评价、生成内容的价值分配、恶意内容对抗等领域,通过国家重点专项牵引、揭榜挂帅地方试点等方式设立一批人工智能治理技术研发项目。三是支持产学研各界在人工智能技术研发过程中主动研究相关伦理与社会治理风险问题。通过优先推广、评估认证等方式,鼓励相关技术产品研发单位提供社会治理风险评估报告与风险监测预防工具。支持可解释、可信赖人工智能技术框架、标准体系、评测机制的研究和推广。

4.5 构建多层次全方位的技术治理体系

人工智能生成内容技术的治理离不开各方的共同参与,应构建由人工智能企业、公众、行业协

会、政府部门等共同组成的治理主体集合,明确权责归属,有效地实现不同治理主体之间的灵活互动和敏捷沟通,更加高效地应对人工智能生成内容技术带来的多重治理挑战。微观层面,强化企业主体责任,在企业技术研发中形成兼顾发展和治理的工作机制。企业应对人工智能生成内容算法的公平性和合理性进行内部审视,加快实施算法备案、算法监督检查、算法风险监测和算法安全评估等举措。中观层面,鼓励由行业协会牵头制定人工智能生成内容的技术标准与行业自律公约,以柔性规制来建立有序开放的行业生态。发挥公众与第三方专业机构的监督作用,提升人工智能生成内容产品的公开度与透明度。宏观层面,政府加强引导作用,明确治理问题及治理目标,充分发挥各方积极性,牵头形成互信、合作和敏捷的治理共同体。

参考文献:

- [1] 郁建兴,刘宇轩,吴超.人工智能大模型的变革与治理[J].中国行政管理,2023,39(4):6-13.
YU J X,LIU Y X,WU C. The social transformation and governance of large generative artificial intelligence models[J]. Chinese Public Administration, 2023, 39(4): 6-13.
- [2] 张凌寒,于琳.从传统治理到敏捷治理:生成式人工智能的治理范式革新[J].电子政务,已录用,2023.
ZHANG L H, YU L. From traditional governance to agile governance: innovation in the governance paradigm of generative artificial intelligence[J]. E-Government, accepted, 2023.
- [3] 宋恺,屈蕾蕾,杨萌科.生成式人工智能的治理策略研究[J].信息技术与政策,2023,49(7):83-88.
SONG K, QU L L, YANG M K. Research on governance strategy of generative artificial intelligence[J]. Information and Communications Technology and Policy, 2023, 49(7): 83-88.
- [4] 于水,范德志.新一代人工智能(ChatGPT)的主要特征、社会风险及其治理路径[J].大连理工大学学报(社会科学版),已录用,2023.
YU S, FAN D Z. The main characteristics, social risks and governance paths of the new generation of artificial intelligence(ChatGPT)[J]. Journal of Dalian University of Technology(Social Sciences), accepted, 2023.
- [5] 张欣.生成式人工智能的数据风险与治理路径[J].法律科学(西北政法学报),2023(5):1-13.
ZHANG X. The data risk and governance path of generative artificial intelligence[J]. Science of Law(Journal of Northwest University of Political Science and Law), 2023(5): 1-13.
- [6] WU J, GAN W, CHEN Z, et al. AI-generated content (AIGC): a survey [EB]. arXiv preprint, 2023, arXiv:2304.06632.
- [7] YOUNG S, BLOOTHOOFT G. Corpus-based methods in language and speech processing[J]. Springer Science & Business Media, 1997.
- [8] REYNOLDS D. Gaussian mixture models[M]/LI S Z, JAIN A. Encyclopedia of Biometrics. Boston, MA: Springer, 2009: 659-663.
- [9] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Advances in Neural Information Processing Systems, 2000.

- [10] MIKOLOV T, KARAFIAT M, BURGET L, et al. Recurrent neural network based language model[J]. In Interspeech, 2010, 2(3): 1045-1048.
- [11] GRAVES A. Long short-term memory[M]// Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Heidelberg: Springer, 2012: 37-45.
- [12] DEY R, SALEM F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//Proceedings of 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Piscataway: IEEE Press, 2017: 1597-1600.
- [13] EFROS A A, LEUNG T K. Texture synthesis by non-parametric sampling[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2002: 1033-1038.
- [14] HECKBERT P S. Survey of texture mapping[J]. IEEE Computer Graphics and Applications, 1986, 6(11): 56-67.
- [15] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.
- [16] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.
- [18] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB]. arXiv preprint, 2020, arXiv:2010.11929.
- [19] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 9992-10002.
- [20] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.
- [21] CAO Y, LI S, LIU Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT [EB]. arXiv preprint, 2023, arXiv:2303.04226.
- [22] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[EB]. arXiv preprint, 2022: arXiv:2203.02155.
- [23] TAN F X, YAN P F, GUAN X P. Deep reinforcement learning: from Q-learning to deep Q-learning[M]//Neural Information Processing. Cham: Springer International Publishing, 2017: 475-483.
- [24] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10674-10685.
- [25] CARLINI N, IPPOLITO D, JAGIELSKI M, et al. Quantifying memorization across neural language models[EB]. arXiv preprint, 2022, arXiv:2202.07646.
- [26] CARLINI N, HAYES J, NASR M, et al. Extracting training data from diffusion models[EB]. arXiv preprint, 2023, arXiv:2301.13188.
- [27] BORJI A. A categorical archive of ChatGPT failures[EB]. arXiv preprint, 2023, arXiv:2302.03494.
- [28] SCHUHMAN C, VENCU R, BEAUMONT R, et al. Laion-400m: open dataset of clip-filtered 400 million image-text pairs[EB]. arXiv preprint, 2021, arXiv:2111.02114.
- [29] 周慧珺, 邹文博. 数字化转型背景下数字鸿沟的现状、影响与应对策略[J]. 当代经济管理, 2023, 45(3): 60-67.
- ZHOU H J, ZOU W B. The present situation, influence and countermeasures of the digital divide under the background of digital transformation[J]. Contemporary Economic Management, 2023, 45(3): 60-67.
- [30] 李春芳, 黄涛. 受众视角下人工智能生成物的版权保护探析[J]. 华南理工大学学报(社会科学版), 2022, 24(1): 24-32.
- LI C F, HUANG T. Copyright protection of artificial intelligence products from the perspective of audience[J]. Journal of South China University of Technology(Social Science Edition), 2022, 24(1): 24-32.

[作者简介]



李亚玲 (1991-), 女, 博士, 之江实验室智能社会治理研究中心助理研究员, 主要研究方向为技术治理和数字经济。



曹缘琪 (1992-), 女, 之江实验室智能社会治理研究中心助理研究员, 主要研究方向为智能交通控制和智能社会治理。



魏阆 (1986-), 男, 博士, 之江实验室发展研究中心助理研究员, 主要研究方向为科技创新生态。