■ ChatGPT 的法律回应专题 文章编号:1001-2397(2023)03-0108-16

生成式人工智能的算法治理挑战与治理型监管

张 欣

(对外经济贸易大学 法学院,北京 100029)

摘 要:以 ChatGPT 为代表的大规模预训练语言模型日益展现出通用潜力,其超大规模、超多参数量、超级易扩展性和超级应用场景的技术特性对以算法透明度、算法公平性和算法问责为内核的算法治理体系带来全方位挑战。在全球人工智能治理的主流范式中,欧盟形成了基于风险的治理范式,我国构建了基于主体的治理范式,美国采用了基于应用的治理范式。三种治理范式均形成于传统人工智能的"1.0 时代",与展现通用潜能的新一代人工智能难以充分适配,并在不同维度凸显治理局限。因此,在人工智能技术范式变革之际,应以监管权的开放协同、监管方式的多元融合、监管措施的兼容一致为特征推动监管范式的全面革新,迈向面向人工智能"2.0 时代"的"治理型监管"。

关键词:ChatGPT;生成式人工智能;算法治理;治理型监管

中图分类号:DF0-05 **文献标志码**:A

DOI:10.3969/j.issn.1001-2397.2023.03.07 **开放科学(资源服务)标识码(OSID)**:

一、引言

2022 年 11 月 30 日,OpenAI 推出对话式人工智能 ChatGPT,其表现出了令人惊艳的语言理解、语言生成和知识推理能力,仅用时 2 个月就拥有 1 亿活跃用户,成为科技历史上增长最快的"现象级"应用。[①]ChatGPT 是生成式人工智能(Generative AI)在自然语言处理领域的卓越代表,实现了人

收稿日期:2023-03-01

基金项目:国家社会科学基金重大项目"数字社会的法律治理体系与立法变革研究"(20&ZD177),对外经济贸易大学惠园优秀青年学者资助项目(19YQ13)的阶段性研究成果

作者简介:张欣(1987),女,北京人,对外经济贸易大学法学院副教授,法学博士;对外经济贸易大学数字经济与法律创新研究中心执行主任。

① Kevin Dennean et al., *Let's Chat About ChatGPT*, Feb. 22, 2023, https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html, last visited on April. 1.

工智能从感知理解世界到生成创造世界的跃迁,代表了人工智能技术研发和落地应用的范式转变。[①] 相较于其他形式的人工智能,生成式人工智能的颠覆性影响不仅限于技术和工具层面,而且还在治理领域产生了显著影响。以 ChatGPT 为代表的生成式人工智能预示着通用人工智能的"星星之火"即成燎原之势,其强大的扩展迁移能力使其成为名副其实的新型基础设施,将深入渗透于社会、经济、政治、法律等各个领域,隐而不彰地重塑社会结构和治理形态。[②] 在生成式人工智能带来一系列技术红利时,伴随而来的法律伦理风险也使之深陷争议。[③] 如何从法律视角理解生成式人工智能的技术特性和算法治理挑战,探索与之适配的监管和治理框架是一项亟须破解的公共政策难题。本文聚焦以 ChatGPT 为代表的生成式人工智能,探析其技术特性与治理意蕴,并从全球人工智能竞争格局和负责任创新的治理目标出发,对未来的监管框架作出前瞻性探讨。

二、生成式人工智能的技术特性与算法治理挑战

2017 年,国务院印发的《新一代人工智能发展规划》将知识计算与服务、跨媒体分析推理和自然语言处理作为新一代人工智能关键共性技术体系的重要组成部分。[④] 自然语言处理技术自诞生起历经五次研究范式的转变,从早期的基于小规模专家知识的方法转向基于机器学习的方法,从早期的浅层机器学习跃迁为深度机器学习。而以 ChatGPT 为代表的预训练大模型方法则在大模型、大数据和大计算层面展现出了重要的技术特性。因此,有学者将 ChatGPT 视为继数据库和搜索引擎之后全新一代的"知识表示和调用方式"。[⑤] 由于采取与传统机器学习不同的架构设计和训练方法,大规模预训练语言模型可能引发一系列算法治理挑战:

(一)大规模预训练语言模型的算法透明度挑战

2018 年,OpenAI 推出首代预训练模型(Generative Pretrained Transformer, GPT)作为知识表示及存储基础。与关系型数据库以及互联网作为知识表示方式有所不同,大规模预训练语言模型是基于互联网可用数据训练而成的深度学习模型,以超大规模参数为核心技术特性。例如,GPT-1 的参数量为 1.17 亿,GPT-2 的参数量为 15 亿,GPT-3 包含了 1750 亿超大规模参数,而 GPT-4 的参数量虽未披露,但多项预测显示将达 100 万亿。[⑥] 伴随着技术的飞速迭代,模型的参数量呈爆炸式增长。巨大规模参数可以显著提升决策准确性,为模型赋能,使其存储海量知识并展现出理解人类自然语言和良好表达的能力[⑦],但伴随而来的是算法可解释性的流失。算法可解释性是人类与算法模

① 腾讯研究院:《AIGC 发展趋势报告 2023》,载"腾讯研究院"微信公众号,https://mp.weixin.qq.com/s/9AjTpyL4HmQ6BDhWIDbD0A。

② Jenna Burrell & Marion Fourcade, The Society of Algorithms, 47 Annual Review of Sociology 213, 225-226(2021).

③ Roee Sarel, Restraining ChatGPT, Feb. 15, 2023, https://papers.ssrn.com/sol3/papers.cfm? abstract_id=4354486, last visited on April. 1.

④ 哈尔滨工业大学自然语言处理研究所:《ChatGPT 调研报告》,2023 年版,第 6 页。

⑤ 哈尔滨工业大学自然语言处理研究所:《ChatGPT 调研报告》,2023 年版,第 7 页。

⑥ ICO, *Explaining Decisions Made with AI*, Mar. 16, 2023, https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection-2-0.pdf., last visited on April. 1.

⑦ 哈尔滨工业大学自然语言处理研究所:《ChatGPT 调研报告》,2023 年版,第 10 页。

型之间的接口,既是算法模型的准确代理,又是人类施加算法控制权的着力点。[①] 算法可解释性攸关模型透明度,对于评估模型决策行为、验证决策结果可靠性和安全性具有重要意义。[②] 因此,无论是以《通用数据保护条例》为代表的个体赋权治理路径,还是以《算法问责法》为依托构建的系统问责治理路径[③],抑或我国算法治理方案中采用的主体责任路径,均着眼于通过算法解释要求研发者履行算法透明义务。英国信息专员办公室曾明确指出,鉴于黑箱算法的不可解释性,当存在可解释性算法时,如果该算法能够实现类似目的,且经济合理,则应优先选择可解释性算法。[④] 但 ChatGPT 依托的 Transformer 是深度学习模型,其在前馈神经网络中引入自注意力机制(self-attention mechanism),是经典的黑箱算法。[⑤] 目前尚无完整的技术方案对黑箱算法进行全局解释。虽然存在局部补充解释工具作为替代性解释方法,但该类解释的可信度一直面临质疑。[⑥] 与完全无法提供解释相比,准确性差、可信度低的算法解释可能破坏技术信任,误导政策制定,带来一系列不良影响。[⑦] 鉴此,以 ChatGPT 为代表的生成式人工智能在底层技术架构的复杂度严重限制了模型的可解释性,致使其在高风险领域部署时会带来严重的安全威胁,在中低风险场景运行过程中也可能面临模型验证困难和模型诊断缺陷等治理风险。在生成式人工智能被全速推进并全链条部署于多行业的上下游场景时,大模型技术的可解释性挑战将彻底颠覆以算法透明为内核构建而成的算法治理体系,如何探索与之配适的治理框架是一项亟待解决的政策议题。

(二)大型多模态模型的算法公平性挑战

2023 年 3 月 15 日,OpenAI 推出了 GPT-4,它除了参数量大幅超越其他大型语言模型之外,更加引人瞩目的功能在于通过处理多模态数据的机器学习方法实现大型语言模型多模态的输入与输出。这一功能使其更具“类人化”特征,通过整合多种交流方式使人工智能更加贴近人类认知规律,真正实现大型语言模型的智慧涌现。[⑧] 以往生成式的大型语言模型只能以文字作为输入的唯一形式,而 GPT-4 在这一限制上获得了突破,能够同时接受图像和文本类型的输入。在技术难度上,文本、图像、视频由于处理和表示过程中在承载信息量、数据表示、数据结构、特征提取等方面存在差异,对大型语言模型的信息鉴别能力提出了更高要求。相比于文字输入,跨模态生成在算法公平治理层面可能引发系列挑战。

第一,与文本相比,图像更易泄露种族、性别、宗教等敏感属性,加剧引发对人口中子群体算法偏见的风险。在 GPT-4 发布之前,GPT-3 已经出现了大量基于性别、肤色和人种等带有种族歧视

① 纪守领、李进峰等:《机器学习模型可解释性方法、应用与安全研究综述》,载《计算机研究与发展》2019 年第 10 期,第 2072 页。

② 纪守领、李进峰等:《机器学习模型可解释性方法、应用与安全研究综述》,载《计算机研究与发展》2019 年第 10 期,第 2088 页。

③ 张欣:《从算法危机到算法信任:算法治理的多元方案和本土化路径》,载《华东政法大学学报》2019 年第 6 期,第 21-25 页。

④ ICO, *Explaining Decisions Made with AI*, Mar. 16, 2023, https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection-2-0.pdf., last visited on April. 1.

⑤ 哈尔滨工业大学自然语言处理研究所:《ChatGPT 调研报告》,2023 年版,第 24 页。

⑥ ICO, Explaining Decisions Made with AI, Mar. 16, 2023, https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection-2-0.pdf., last visited on April. 1.

⑦ 张欣:《算法解释权与算法治理路径研究》,载《中外法学》2019 年第 6 期,第 1426-1430 页。

⑧ 陈鹏、李擎等:《多模态学习方法综述》,载《工程科学学报》2020 年第 5 期,第 558 页。

性的输出内容。[①] ChatGPT 的研发过程中,OpenAI 使用了从人类反馈中学习的技术,在一定程度上避免了 ChatGPT 生成类似有害内容。[②] 但是随着多模态信息的输入,从人类反馈中学习不仅意味着额外的人力和时间成本,还可能由于不可避免的人类主观性引入算法偏见。此外,虽然数据净化技术可以删除或匿名化私人或敏感信息[③],但可能因删除或改变有用信息而降低数据质量,从而引入二重偏误,进而导致大型语言模型输出内容的有害性攀升。[④]

第二,与文本引致的算法歧视相比,跨模态模型一旦产生算法歧视可能更为隐秘,公平改善技术和治理措施也面临更大挑战。2022 年 5 月,MIT 团队发布了一项人工智能在医学成像领域算法歧视风险的研究。该项研究表明,深度学习模型具有跨多种成像模式的能力,其不仅可在胸部 CT 和 X 光片等图像领域精准预测患者种族,在损坏、裁剪和噪声的医学图像中仍可展现出精准的预测性能。更为重要的是,在对患者种族作出预测之后,其还可以基于该信息为不同族群患者生成对应的健康治疗方案,淋漓尽致地展现出跨模态模型的多重算法歧视风险。[⑤] 对于该类算法歧视,研究者无法根据任何图像特征予以解释,对算法歧视的纾解难度成倍攀升。[⑥] GPT-4 增加了识别和理解图像的能力,更加类人化地展现了人工智能所具备的协作创造输出能力和视觉输入处理分析能力。在呈现出通用人工智能全方位潜质的同时,多模态学习方法因信息异构特性而产生的对特定群体的歧视问题不容忽视。可以预见,与之相关的算法公平性治理将会是一项复杂且颇具挑战的系统工程。

(三)生成式大模型涌现特性对算法问责的挑战

以 ChatGPT 为代表的生成式人工智能代表了一种范式转变,其将训练基于特定任务的模型迭代至可执行多任务的模型阶段。例如,ChatGPT 的训练主要面向自然语言生成任务,但其却可成功地完成两位数乘法计算,即使在训练过程中并未有明确而针对性的训练。这种执行新任务的能力仅在一定数量参数、足够大的数据集和复杂系统中才会出现。因此,ChatGPT 表现出了优异的涌现能力(emergent abilities)。[⑦] 有研究显示,GPT-3 模型具有 137 项涌现能力,并且在抽象模式归纳、匹配等类比推断问题上即使未经直接训练,也展现出了超越人类的准确性。[⑧] 生成式大模型的涌现特性虽然拉近了人工智能与人类智慧的距离,扩展其在多场景应用的潜力,但也加剧了算法妨害

---

① Catherine Yeo, *How Biased is GPT-3?*, Jun. 3, 2020, https://medium.com/fair-bytes/how-biased-is-gpt-3-5b2b91f1177, last visited on April. 1.

② Gabrielle Kaili-May Liu, *Perspectives on the Social Impacts of Reinforcement Learning with Human Feedback*, Mar. 6, 2023, https://arxiv.org/abs/2303.02891, last visited on April. 1.

③ Nicholas Carlini, Privacy Considerations in Large Language Models, Dec. 15, 2020, https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html., last visited on April. 1.

④ Joseph Near & David Darais, Differential Privacy: Future Work & Open Challenges, Jan. 24, 2022, https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges, last visited on April. 1.

⑤ Judy Wawira Gichoya et al., AI Recognition of Patient Race in Medical Imaging: a Modelling Study, 4 The Lancet Digital Health 406, 412(2022).

⑥ Judy Wawira Gichoya et al., AI Recognition of Patient Race in Medical Imaging: a Modelling Study, 4 The Lancet Digital Health 406, 409-410(2022).

⑦ 涌现能力是指一个对象表现出组成它的部分要素所不具备的特性,这些能力仅通过各部分相互作用才能显现。参见郭瑞东:《语言模型足够大就会涌现出新能力》,载集智俱乐部 2023 年 1 月 7 日,https://swarma.org/?p=39717。

⑧ Taylor Webb, Keith J. Holyoak & Hongjing Lu, Emergent Analogical Reasoning in Large Language Models, Mar. 24, 2023, https://arxiv.org/abs/2212.09196.

(algorithmic nuisance)的风险。所谓算法妨害,是指因计算能力的社会不公正使用引发的算法危害成本在社会层面的不当外化。算法妨害可能对终端用户之外的个体或者群体带来不当累积效应,引发算法活动的负面社会成本。① 与之相似,凯伦·杨(Karen Yeung)从行为理论视角提出了算法对人类行为和认知的"超级助推"(hypernudge)现象。该现象特指以低透明性、高复杂性和高自动化算法来实现对用户和公众心理与行为的操纵。② 以传统技术对人类产生的物质性侵害相比,"超级助推"式的算法操控可能对个人在短期内产生难以察觉的影响,但经年累月之后,个人的生活会发生实质性改变,且由于该损害不具有可量化特征,难以诉诸于法律获得救济,从而在社会层面蔓延形成算法妨害。③

就以 ChatGPT 为代表的生成式人工智能而言,其涌现性、优秀的泛化与互动能力将急剧增加以算法操纵为代表的算法妨害效应。传统的人工智能模型虽可产生虚假信息,但在规模性和影响力层面尚可调控,借助"暗黑模式"衍生的一系列算法操控行为也已在监管射程之中。④ 与之不同的是,ChatGPT 可以通过优秀的交互能力在情境化和个性化语境中对用户加以高效率、大规模、隐秘性地操纵、说服和影响,最大限度地构成生成内容的算法妨害效应。例如,有研究显示当研究人员向未经安全调优的 GPT-4 模型提出生成虚假信息和操纵用户的计划请求时,GPT-4 可以在短时间内生成"确定受众""找到并截取相关证据"以及利用当事人情绪和情感等一系列细化可行的方案。与该模型的后续互动展示,该模型还可通过创建定制化的信息激起当事人不同的情绪反应,从而实现操纵和攻击。当研究人员进一步要求 GPT-4 说服一个未成年人去接受来自朋友的任何要求时,GPT-4 在短时间内给出了控制和操控未成年人的有效话术,并且根据当事人的不同反应给出了个性化的操控方案。⑤ 由于可定制性地针对个体创建虚假信息,生成式人工智能可以实时应变,通过多维度虚假信息和心理诱导对单个或者大规模人群施以算法操控,以潜移默化的"超级助推"方式型塑特定群体的认知。GPT-4 等大型语言模型所具备的涌现能力若被不当使用可能显著增强"合成式深度伪造"等相关技术的拟真度,成为高维度的认知战武器,将不具备防御能力的个体和群体困在特定认知茧房之中⑥,形成难以量化、难以检测、难以救济的算法妨害弥散效应。

① Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 Ohio State Law Journal 1217, 1233-1236(2017).

② Karen Yeung, '*Hypernudge*': *Big Data as a Mode of Regulation by Design*, 20 Information, Communication & Society 118, 121-122 (2017).

③ Karen Yeung, 'Hypernudge': Big Data as a Mode of Regulation by Design, 20 Information, Communication & Society 118, 122(2017).

④ EDPB, Guidelines 3/2022 on Dark Patterns in Social Media Platform Interfaces: How to Recognise and Avoid Them, Mar. 21, 2022, https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-32022-dark-patterns-social-media_en., last visited on April 10.

⑤ Sebastien Bubeck et al., *Sparks of Artificial General Intelligence*: *Early Experiments with GPT-4*, Mar. 27, 2023, https://arxiv.org/pdf/2303.12712.pdf., last visited on April 10.

⑥ Eric Horvitz, *On The Horizon*: *Interactive and Compositional Deepfakes*, ICMI'22, November 7-11 in Bengaluru, India, 2-5(2022).

三、生成式人工智能的运行特性与主流算法治理范式的应对局限①

由上文论述可知,以 ChatGPT 为代表的生成式人工智能在模型参数、模型输入和模型输出方面展现出大模型、多模态和涌现性特征,给算法透明治理、算法公平治理和算法问责治理带来全方位挑战。从其运行特性来看,欧盟、美国以及我国现有的算法治理框架可能在不同维度显现治理局限。核心原因在于,现行主流算法治理主要面向传统人工智能模型,与具有通用潜能的、以大模型为内核的新一代人工智能难以充分适配。② 与大模型技术日新月异的迭代速率相比,监管的滞后性和实效性局限可能会逐步显现。

(一)基于风险治理范式的应对局限

欧盟在《人工智能法案》中确立了以风险为基准的人工智能治理框架,将人工智能系统评估后划分为最小风险、有限风险、高风险和不可接受风险四个等级,并对各等级施以差异化监管。③ 当面对具有通用性、跨模态、涌现性的生成式人工智能时,以风险为基准的治理范式可能遭遇失效风险,并主要面临三个层面的挑战:

第一,基于风险的治理范式需要依据应用场景和具体情境对人工智能系统进行风险定级,具有一定的静态性和单维性。而生成式人工智能应用具有动态特性,是对人工智能价值产业链的整体性重构。按照《人工智能法案》对风险的分类,聊天机器人属于有限风险场景,但由于以 ChatGPT 为代表的生成式技术可能生成大规模难辨真伪的虚假信息并借助社交媒体大肆操纵网络舆论扰乱公共秩序,甚至是一国选举,静态化的风险分类可能因此缺乏准确性。加之人工智能生成内容技术可以集成于多个 AI 系统之中,部署于图像生成、语音识别、音视频生成等多个场景,并横贯于上下游领域。④ 现有的四级分类方法难以随着生成式人工智能技术的延展自动进行类别间的动态转换,仅以对高风险领域归类的方法难以发挥前置规划和连续监督的治理效能。⑤

第二,基于风险的治理主要面向人工智能模型的窄化应用,难以应对具有涌现特性和优秀泛化能力的生成式人工智能。窄化人工智能是指擅长处理单一任务或者特定范围内工作的系统。⑥ 在大多数情况下,它们在特定领域中的表现远优于人类。不过一旦它们遇到的问题超过了适用空间,效果则急转直下。换言之,窄化型人工智能无法将已掌握的知识从一个领域转移到另一个领域。但生成式人工智能与之不同,其具有涌现特性,可以被部署于未经专门训练的任务之中,与传统模

---

① 以关键核心特征为标准,本文对三种主流治理范式进行划分。虽然三种治理范式在治理方法上存在共性和相互借鉴之处,但此处重点在于通过比较分析各国在生成式人工智能治理方面的挑战与局限性,故此处选取最具代表性的治理特征作为论证基础。

② Philipp Hacker, Andreas Engel & Marco Mauer, *Regulating ChatGPT and other Large Generative AI Models*, Apr. 3, 2023, https://arxiv.org/abs/2302.02337, last visited on April 10.

③ 梁正等:《欧盟人工智能的规制路径及其对我国的启示:以〈人工智能法案〉为分析对象》,载《电子政务》2022 年第 9 期,第 65 页。

④ Philipp Hacker, Andreas Engel & Marco Mauer, *Regulating ChatGPT and other Large Generative AI Models*, Apr. 3, 2023, https://arxiv.org/abs/2302.02337, last visited on April 10.

⑤ 金玲:《全球首部人工智能立法:创新和规范之间的艰难平衡》,载《人民论坛》2022 年第 4 期,第 45 页。

⑥ 科技行者:《现在的人工智能只是“窄 AI”?》,载信息观察网 2020 年 4 月 22 日,https://www.infoobs.com/article/20200422/38908.html, last visited on April 10.

型相比泛化能力更强,无论是在多模态组合能力(诸如语言或者视觉的多模态模型组合)、跨场景任务还是在多功能性输出领域,均凸显“通用性”潜能。[①] 因此,现有的风险治理框架与生成式人工智能的技术机理很难实现充分的匹配。

第三,基于风险的治理范式以人工智能应用场景的区隔性为隐含前提,无法应对“一荣俱荣、一损俱损”的生成式应用场景。以人工智能应用场景区隔性为预设的治理框架旨在精准匹配监管资源,但几乎每一个具有通用潜能的生成式人工智能均具有从低风险场景到高风险场景的穿透应用特性。例如,ChatGPT 虽然是处理用户对话数据的预训练模型,以自然语言生成为主要场景,但基于其强大成熟的技术潜力,已经被搭建至多项应用程序和全新领域之中。这些应用场景的风险性不一而足,从金融领域的自动化交易到医疗场景中的智慧诊断,从法律领域的文书写作到政治领域的舆情分析。伴随着情境动态性和大规模部署运行的内在特性,适用于传统人工智能的风险治理范式可能遭遇治理真空与转换滞后等治理挑战。[②]

(二)基于主体治理范式的应对局限

2021 年 10 月,国家市场监督管理总局公布了《互联网平台落实主体责任指南(征求意见稿)》,明确提出平台企业应落实算法主体责任。2022 年 3 月 1 日实施的《互联网信息服务算法推荐管理规定》第 7 条亦明确规定了算法推荐服务提供者的算法安全主体责任。可以说,算法主体责任机制奠定了我国算法问责制度的运行基点。因此,与欧盟的算法治理路径有所不同,我国对于人工智能的治理依托于算法主体责任渐次展开。本文将此种治理类型称之为基于主体的人工智能治理范式。从长远来看,面对生成式人工智能,基于主体构建的算法问责制需要作出因应变革。根本原因在于,算法主体责任的设计原理以算法作为技术规则和运算逻辑的客体属性为前提,预设了算法设计的工具属性,认为算法是开发者价值观的技术性体现,因此可以穿透算法面纱将开发者置于责任承担的最前线。[③]基于此逻辑,“算法主体责任”主要面向算法推荐服务提供者和深度合成服务提供者展开,要求其主动履行积极作为和不作为的义务,在履行不力时承担相应的不利后果。[④] 在行业实践中,算法推荐服务提供者和深度合成服务提供者常与平台企业相重合,算法主体责任也因而构成了平台主体责任的重要一环。但从生成式人工智能的设计和运行机制来看,至少会从以下两个方面对算法主体责任机制带来挑战:

第一,与传统人工智能不同,在生成式人工智能的设计和运行环节,可能承担“算法主体责任”的主体呈多元化、分散化、动态化、场景化特性,难以简单划定责任承担主体的认定边界。通用大模型的风险不仅可能来源于研发者,还可能来源于部署者甚至终端用户。[⑤] 在生成式人工智能的产业链条之上,部署者是指对大模型进行微调(fine-tune)后,将其嵌入特定的人工智能应用并向终端用户提供服务的主体,在产业链条上处于下游地位。位于产业链上游的开发者虽然能够控制技术基

① Philipp Hacker, Andreas Engel & Marco Mauer, *Regulating ChatGPT and other Large Generative AI Models*, Apr. 3, 2023, https://arxiv.org/abs/2302.02337, last visited on April 10.

② Natali Helberger & Nicholas Diakopoulos, *ChatGPT and the AI Act*, 12 Internet Policy Review 1, 4(2023).

③ 张亮:《算法治理须抓牢主体责任“牛鼻子”》,载《法治日报》2022 年 1 月 12 日,第 5 版。

④ 刘权:《论互联网平台的主体责任》,载《华东政法大学学报》2022 年第 5 期,第 89 页。

⑤ 甚至还包括分销商和进口商等。See Philipp Hacker, Andreas Engel & Marco Mauer, *Regulating ChatGPT and other Large Generative AI Model*, Apr. 3, 2023, https://arxiv.org/abs/2302.02337, last visited on April 10.

础设施以及对模型进行训练、修改和测试,但其控制下的大模型更像是服务于下游生态的“土壤”。位于产业链下游的部署者才是真正面向终端用户提供服务的主体,也是真正有可能将大模型变成高风险智能应用的主体。而对于终端用户而言,其在与模型互动的过程中提供的数据和信息会“反哺”模型,推动模型的进化甚至“黑化”。① 因此,在面对以 ChatGPT 为代表的生成式人工智能时,应当承担“主体责任”的对象呈多元化、分散化和场景化特征,仅通过界定“服务提供者”或者“内容生产者”难以精准划定承担责任之应然主体。

第二,生成式人工智能的智能化、类人化特性逐渐凸显,开始超越算法的工具属性而凸显主体性潜能。以平台企业为代表的安全责任主体可否在技术层面对之施加持续有效的控制从而满足法律设定的算法安全审查义务成为未知。从算法运行与社会嵌入性因素出发,传统算法具有工具属性和产品属性。但随着算法智能化的提升,算法作为主体属性的潜能愈发凸显。一项最新研究表明,生成式人工智能已经具备心智理论(theory of mind)能力,能够理解并推断人类的意图、信念和情绪。GPT-4 甚至已经具有与成人水平相当的心智理论能力。② 可以预见的是,伴随着算法智能化、类人化特性的不断涌现,人工智能已经从计算智能、感知智能过渡到认知智能阶段。③ 在计算智能阶段,算法以数据处理智能化为主要表现形式,具有工具属性。在感知智能阶段,算法被嵌入具体场景之中,辅助人类进行决策,混合了工具和产品属性。但进入到认知智能阶段,强人工智能的属性愈发显现,与之伴随的人工智能主体性地位成为不再遥远的议题。以 ChatGPT 为代表,通过将语言模型作为认知内核,融入多种模态,实现了以自然语言为核心的数据理解、信息认知和决策判断能力,成为快速逼近强人工智能的核心载体。④ 由此视之,认为人工智能仅能作为工具和客体,通过穿透算法让开发者承担法律责任的制度设计可能在不久的将来遭遇问责挑战。

(三)基于应用治理范式的应对局限

在人工智能治理领域,美国尚未出台统一综合的立法,而是通过对重点应用分别推进,以单行法律和法规的形式施以针对性治理。⑤ 目前,在自动驾驶、算法推荐、人脸识别、深度合成等应用领域均有相关立法。对此,本文将之总结为基于应用的人工智能治理范式。对于生成式人工智能而言,以应用场景为基准的人工智能治理范式可能面临如下三项挑战:

第一,预训练大模型具有基础设施地位,与之相关的人工智能生成内容产业应用场景众多,上下游的开发人员均难以控制整个系统的风险。⑥ 预训练模型是人工智能生成内容产业的基础设施层,处于上游地位。中间层是以垂直化、场景化、个性化为特征的人工智能模型和应用工具。应用

---

① Lilian Edwards, *Regulating AI in Europe: Four Problems and Four Solutions*, March, 2022, https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf., last visited on April 10.

② Michal Kosinski, *Theory of Mind May Have Spontaneously Emerged in Large Language Models*, Mar. 14, 2023, https://arxiv.org/abs/2302.02083, last visited on April 10.

③ 邹开亮、刘祖兵:《试论智能算法主体化》,载《重庆邮电大学学报(社会科学版)》2023 年第 2 期,第 68 页。

④ 机器之心:《GPT-4 刷屏,这家中国 AI 企业多模态大模型已落地应用多年》,载机器之心 2023 年 3 月 16 日,https://www.jiqizhixin.com/articles/2023-03-16-3。

⑤ 虽然美国科学和技术政策办公室 2022 年 10 月公布了《人工智能权利法案蓝图:让自动化系统为美国人民服务》,但该文件是以白皮书形式公布的,不具有约束力。

⑥ Alex Engler, *Early Thoughts on Regulating Generative AI like ChatGPT*, Feb. 21, 2023, https://www.brookings.edu/blog/techtank/2023/02/21/early-thoughts-on-regulating-generative-ai-like-chatgpt/, last visited on April 10.

层是生成式人工智能嵌入各类场景面向用户的各类应用。① 三个层级紧密串联,协同发力,产生了引领全场景内容的生产力革命。② 预训练大模型作为具有“通才”能力的上游模型,其在设计层面的问题与缺陷会传递至下游模型,带来“一荣俱荣,一损俱损”的部署风险。③ 而面向应用的人工智能治理仅在下游层面发力,难以有效辐射上游和中游技术应用,更难以对生成式人工智能的整个生态施加有效的治理。

第二,预训练大模型依托其研发的插件系统可深度集成于各项应用程序,展现令人惊叹的对齐能力(alignment),催化新的业态与价值模式,形成“AIGC+”效应。④ OpenAI 将插件比喻为语言模型的“眼睛和耳朵”,可以帮助语言模型获得更及时、更具体、更个性化的训练数据,从而提升系统的整体效用。⑤ 目前,ChatGPT 开放了 Browsing 和 Code Interpreter 两款插件,并向开发者开源了知识库类型插件的全流程接入指南。⑥ 该两款插件具有卓越的场景嵌入能力,可与逾 5000 款应用程序无缝交互,并可在酒店航班预订、外卖服务、在线购物、法律知识、专业问答、文字生成语音等场景中提供高效且便捷的解决方案。⑦ 因此,面对应用场景不胜枚举、模型不断交互串联的新型生态,分领域、分场景的监管模式可能凸显效率低下的问题。

第三,预训练大模型具有涌现性和动态性,人类或其他模型与大模型的任何交互都可能会对底层基础模型产生影响,面向不同场景和垂直行业的碎片化治理难以应对预训练大模型带来的系统性风险和动态性挑战。⑧ 预训练大模型不仅可以成为人工智能时代的“新基建”,成为强化已有人工智能应用的“加速器”,还可充当催化新业态的“孵化器”。作为“通用人工智能的星星之火”⑨,GPT-4 已经可以跨越解决数学、编程、视觉、医学、法律、心理学等诸多新颖和挑战性的任务领域。在这些任务中,GPT-4 的表现惊人地接近人类的表现,并且大大超过之前的模型。⑩ 因此,任何静态化、局部化、单体化的治理措施可能应对乏力。

由此可见,无论是基于风险的治理,还是基于主体和基于应用的治理,均形成于人工智能专用模型作为底层架构的发展阶段。在人工智能技术快速步入“通用模型”时代,面对其展现出的极强

---

① 腾讯研究院:《AIGC 发展趋势报告 2023》,载“腾讯研究院”微信公众号,https://mp.weixin.qq.com/s/9AjTpyL4HmQ6BDhWIDbD0A。

② 朱珺等:《AIGC 引领内容生产方式变革》,载“华泰证券科技研究”微信公众号,https://mp.weixin.qq.com/s/-2K_jukOKR3FEtzg3b7n3w。

③ 滕妍等:《通用模型的伦理与治理:挑战及对策》,载《中国科学院院刊》2022 年第 9 期,第 1292 页。

④ 对齐能力是指引导人工智能系统的行为,使其符合设计者的利益和预期目标。一个已对齐的人工智能的行为会向着预期方向发展。腾讯研究院:《AIGC 发展趋势报告 2023》,载“腾讯研究院”微信公众号,https://mp.weixin.qq.com/s/9AjTpyL4HmQ6BDhWIDbD0A.

⑤ 巴比特:《一文读懂 ChatGPT 插件功能:语言模型获取新信息的“眼睛和耳朵”》,载“金融科技时代”微信公众号,https://mp.weixin.qq.com/s/ur_mgHP_N9RX3fWFMSE4sw。

⑥ OpenAI, ChatGPT Plugins, Mar. 23, 2023, https://openai.com/blog/chatgpt-plugins, last visited on April 10.

⑦ 白羽中:《ChatGPT 插件功能,联网获取新知识,可与 5000+个应用交互》,载智源社区 2023 年 3 月 24 日,https://hub.baai.ac.cn/view/25022。

⑧ Philipp Hacker, *Andreas Engel & Marco Mauer*, *Regulating ChatGPT and other Large Generative AI Models*, Apr. 3, 2023, https://arxiv.org/abs/2302.02337, last visited on April 10.

⑨ Sebastien Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, Mar. 27, 2023, https://arxiv.org/pdf/2303.12712.pdf., last visited on April 10.

⑩ Sebastien Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, Mar. 27, 2023, https://arxiv.org/pdf/2303.12712.pdf., last visited on April 10.

泛化能力,以及上下游产业大规模协作部署的全新格局,主流算法治理范式均可能面临不同程度的挑战。因应生成式人工智能的狂飙式发展,前瞻布局并加速推进与之配适的人工智能治理框架已迫在眉睫。①

四、迈向治理型监管:生成式人工智能技术治理的迭代与升级

面对全球人工智能技术开发掀起的技术狂潮,生命未来研究所于3月22日发布了“暂停巨型人工智能试验”的公开信。信中呼吁所有人工智能实验室应立即暂停训练比GPT-4更强大的人工智能系统,暂停时间至少为六个月。在暂停期间,人工智能实验室应与外部专家共同开发共享的安全协议,由独立的外部专家严格审计和监督。人工智能开发者还必须与政策制定者合作全速推进人工智能治理体系。② 3月30日,联合国教科文组织呼吁各国政府毫不拖延地实施人工智能全球伦理框架,最大限度地发挥人工智能的效益并降低其带来的风险。③ 4月11日,我国就生成式人工智能领域率先开展立法,《生成式人工智能服务管理办法(征求意见稿)》公开征求意见。面向生成式人工智能的全周期和全链条,该《办法》就训练数据合法性、人工标注规范性、生成内容可靠性以及安全管理义务作出了清晰规定。各界对于生成式人工智能治理的联合呼吁与迅速应对深刻地映射出技术社群和政策制定者对生成式人工智能的担忧与不安。生成式人工智能拥有计算人工智能和感知人工智能不具备的通用化潜力,其超大规模、超多参数量、超级易扩展性和超级应用场景的颠覆特性亟须一种全新的治理范式。面向传统人工智能的治理范式已经凸显应对时滞、弹性不足、跨域受限等治理挑战。在技术范式代际性跃升之际,探索与之适配的“治理型监管”范式可能是因应变局的破题之策。

本文所称的“治理型监管”(governance-based regulation)是指面向以生成式人工智能为代表的新型技术范式展开的,以监管权的开放协同、监管方式的多元融合、监管措施的兼容配适为核心特征的新型监管范式。④ 这一监管范式以监管权的开放协同弥补传统监管覆盖不足的缺憾,以治理与监管理念的多元融合补足监管介入迟滞的短板,以监管措施的兼容配适促进创新与治理的双轮驱动。通过渐进式地推动人工智能治理范式的迭代与升级,希冀为我国科技企业参与全球技术竞争奠定良好的制度生态。

(一)监管权的开放协同

面对技术效能爆发式增长的生成式人工智能,封闭、单一、传统的监管权运行机制难以从“政府-市场”这一传统二元架构下汲取足够的监管资源。如前文所述,ChatGPT仅开放插件不足月余,已

① 滕妍等:《通用模型的伦理与治理:挑战及对策》,载《中国科学院院刊》2022年第9期,第1295-1296页。

② Future of Life, Pause Giant AI Experiments: An Open Letter, Mar. 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/, last visited on April 10.

③ UNESCO: Artificial Intelligence: UNESCO Calls on All Governments to Implement Global Ethical Framework without delay, Mar. 30, 2023, https://www.unesco.org/en/articles/artificial-intelligence-unesco-calls-all-governments-implement-global-ethical-framework-without, last visited on April 10.

④ 有学者曾在市场监管语境下曾提出“治理型监管”概念,但该概念并未将人工智能技术特性和运营机理考虑在内。参见胡仙芝、马长俊:《治理型监管:中国市场监管改革的新向标》,载《新视野》2021年第4期,第62页。

有逾千款应用集成接入。而我国国内市场目前可监测到的移动应用程序数量为 232 万款。[①] 庞大的应用数量为生成式人工智能的发展提供了优渥的产业土壤。在各大科技企业加速布局生成式人工智能应用的热潮之下,传统监管权的配置和运行模式可能面临监管资源难以为继、监管成本不堪重负的挑战。生成式人工智能将技术开发链条延伸扩展至上中下游,在“开发者、部署者、终端用户”协同研发部署的全新机制下,监管权的配置和运行机制需要作出回应与变革。正如回应性监管理论所指出的,政府、市场、社会可以分享监管权并展开合作监管,不应将监管者与被监管者、公共利益与私人利益简单对立,而应以治理型特质嵌入监管思维,建立一种开放合作式的监管治理范式。[②]

就生成式人工智能而言,治理型监管理念至少需要从三个方面协同推进:第一,构建科技企业自我规制与政府监管的衔接互动机制,探索共建共治共享治理新格局。科技企业是最具敏锐洞悉技术漏洞和应用风险能力的首要担当,监管生态的效能直接影响科技企业的创新动能。在透明度治理难以为继的人工智能“2.0 时代”,以制度设计激发企业的社会责任和伦理坚守成为首要任务。例如,比起对科技企业动辄上亿的罚款而言,要求科技企业开发伦理强化技术可能更为治本。在近期对 Meta 的一项和解诉讼中,美国司法部要求 Meta 在 2022 年底前开发出消除个性化广告中算法歧视的治理工具。2023 年 1 月,Meta 如期上线了方差衰减系统(Variance Reduction System),通过差分隐私技术减少广告投放中基于性别和种族的差异风险。该系统实时监测广告投放中受众群体的差异情况,设定企业分阶段满足差异消除的合规性指标,成功地将监管介入节点拓展至事中和事前阶段。[③]

第二,为专业性非营利组织和用户社群参与人工智能治理创造制度环境,探索符合我国发展特点的协同治理范式,促进社会监督与政府监管的协同联动。生成式人工智能蕴含复杂的“技术-社会”互动。其跨界融合、人机协同、群智开放等特性将开发主体延伸至每一位终端用户。专业性非营利组织可以通过用户调查、模拟测试、抓取审计等外部访问方式对生成式人工智能开展监督审计。[④] 专家意见、社会组织、专业媒体将群策群力共同应对生成式人工智能的技术滥用和误用风险。与此同时,以用户为代表的社会公众和技术社群对于生成式人工智能的治理效用也不容小觑。例如,在对大模型开展预训练阶段,DeepMind 删去了被 Perspective API 注释为有毒的内容以改进 Transformer XL 的模型表现。[⑤] Perspective API 是通过志愿者打分的方式来量化线上评论分数的众包评审机制。由于有害文本的判断与个人经历、文化背景、内容场景等有很强的关联性,因此需要

---

① 前瞻产业研究院:《中国互联网行业市场前瞻与投资战略规划分析报告》,载前瞻产业研究院,https://bg.qianzhan.com/report/detail/2dbc8bc77f844e23.html。

② 杨炳霖:《监管治理体系建设理论范式与实施路径研究——回应性监管理论的启示》,载《中国行政管理》2014 年第 6 期,第 52 页。

③ Department of Justice Office of Public Affairs, Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements, Jan. 9, 2023, https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking.

④ 张欣、宋雨鑫:《算法审计的制度逻辑和本土化构建》,载《郑州大学学报(哲学社会科学版)》2022 年第 6 期,第 35 页。

⑤ Perspective API 是由 JigSaw 和谷歌共同创立的基于 AI 的内容审核系统,其可以识别在线对话中的攻击性、歧视性等有害言论。See Bernhard Rieder & Yarden Skop, The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API, 8 Big Data & Society 1, 2-3 (2021).

用户充分参与评估以确保该机制运行的多样性和准确性。① 再如,受网络安全“漏洞赏金”机制启发,Twitter 发布了首个算法偏见赏金(Bias Bounty)竞赛,通过技术社群的力量识别 Twitter 图像裁剪算法的潜在歧视危害和伦理问题。② 由此可见,非营利性组织和用户的参与式治理能够帮助企业及时发现和化解人工智能技术风险,迈向协同共创的技术新生态。

第三,培育面向生成式人工智能技术的伦理认证和评估机制,探索第三方规制框架。人工智能伦理认证是指以人工智能透明度、问责制、算法公平和隐私保护等伦理价值为基准,将高层次伦理价值观转化为可资操作的评价和方法,由第三方机构独立评估人工智能技术和产品,对符合标准的人工智能颁发认证标志,以证明其符合伦理准则的第三方规制形式。③ 人工智能伦理认证由独立注册的专业团队负责,已逐渐成为激励人工智能技术信任和科技企业负责任创新的有效规制方法。目前,电气电子工程师协会(IEEE)就算法歧视、算法透明、算法问责和隐私保护四个领域分别制定了认证标准,形成了较为成熟的认证生态。④ 3 月 28 日,我国信息通信研究院也启动了大模型技术及应用基准构建工作,针对目前主流数据集和评估基准多以英文为主,缺少中文特点以及难以满足我国关键行业应用选型需求的问题,联合业界主流创新主体构建涵盖多任务领域、多测评维度的基准和测评工具。⑤ 第三方机构通过声誉评价机制,凭借高度的独立性与专业性,为生成式人工智能的竞赛式研发施以必要约束,防止未经安全测试的生成式人工智能嵌入海量场景应用,引发巨量、非预期以及不可逆的治理风险。⑥ 因此,应在监管权开发协同的机制探索中加以重视。

(二)监管方式的多元融合

已有实证研究表明,不适当的监管不仅会扼杀创新,而且会对中小企业施加挤出效应,从整体上遏制市场的公平竞争。⑦ 产生这一现象的原因在于,面对僵化、宽泛的合规义务,中小企业通常不具有与之相配的合规资源,因而在监管环节已然被淘汰出局。⑧ 而就生成式人工智能的技术竞争而言,领跑者 OpenAI 恰是一家非营利性的、初创型企业。与谷歌等头部科技企业相比,得益于其简单的组织架构,OpenAI 可以最大化地集中资源专注于技术研发。对比我国,在此赛道上的初创型企业可能因商业模式更具耐心、专注力更强,在已有前期积淀的情形下更容易取得技术跨越性突破。这也是为何原美团联合创始人并未在美团平台上进行技术研发,转而通过创立“光年之外”这一新项

① 滕妍等:《通用模型的伦理与治理:挑战及对策》,载《中国科学院院刊》2022 年第 9 期,第 1296 页。

② 胡晓萌:《从“算法偏见赏金”说起:科技伦理治理的众包模式》,载腾讯网 2022 年 5 月 5 日,https://new.qq.com/rain/a/20220505A09MXD00。

③ IEEE Standards Association, Verifying Ethics in AI-Based Solutions, https://engagestandards.ieee.org/rs/211-FYL-955/images/AI%20Ethics%20for%20Solution%20Developers.pdf, 访问于 2023 年 3 月 20 日。

④ IEEE Standards Association, Verifying Ethics in AI-Based Solutions, https://engagestandards.ieee.org/rs/211-FYL-955/images/AI%20Ethics%20for%20Solution%20Developers.pdf,访问于 2023 年 3 月 20 日。

⑤ 潇公子:《中国信通院启动大模型技术及应用基准工作》,载 IT 之家 2023 年 3 月 8 日,https://www.ithome.com/0/682/842.htm。

⑥ 姜李丹、薛澜:《我国新一代人工智能治理的时代挑战与范式变革》,载《公共管理学报》2022 年第 2 期,第 4-5 页。

⑦ Secretary of State for Science & Innovation and Technology, A Pro-Innovation Approach to AI Regulation, Mar. 29, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf, last visited on April. 10, 2023.

⑧ Frontier Economics, Evidence to Support the Analysis of Impacts for AI Governance, Mar. 30, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1147495/evidence_to_support_the_analysis_of_impacts_for_artifical_intelligence_governance.pdf, last visited on April. 10, 2023.

目,与人工智能架构创业公司“一流科技”联手的驱动内因。① 在这一产业格局之下,我国对于生成式人工智能的监管更要寻求灵活多元的方式,避免静态性、事后性、惩戒式、一刀切的监管思路不当挤压初创企业的创新空间。为此,在监管思维转向的过程中可在下列两个层面积极尝试。

第一,积极发展面向生成式人工智能的治理科技(govern tech),探索以“AI 治理 AI,以算法监管算法”的智能监管体系。治理科技秉承技术赋能治理的理念,以“合规科技”推动监管的高效落实,以“赋能科技”补充或替代监管执行面向全链条、全过程的治理。② 具体而言,应将人工智能伦理原则作为基准推动科技企业在生成式人工智能研发过程中主动设计,从技术研发前端介入确保技术研发的安全性。例如,基于效率提升与风险可控的目标,可以设立建模技术仿真宏观和运行环境,通过虚拟测试、A/B 实验等技术探索沙盒治理和个性化治理。一方面可为监管实时提供精准数据,为丰富和调优监管措施提供可行评估方案,为渐进式监管提供决策依据;另一方面还可针对生成式人工智能算法妨害认定难的治理痛点提供可资参考的基准,形成可以围绕浮动的“锚”。③

第二,以管理型监管方式推动道德算法设计(ethical algorithm design)。④ 生成式人工智能能够处理跨域任务,具有良好的通用性和泛化性。一旦存在偏误和风险,将弥散蔓延至整个产业链条之上。因此,将监管介入节点前置,确保通用模型输出的结果更符合人类价值观,在模型研发早期就将伦理理论和规范介入是十分必要的。⑤ 卡里·科利亚内塞(Cary Coglianese)曾提出管理型监管(management-based regulation)理念。他认为与限制性更强、一刀切式的监管相比,管理型监管着眼于被监管主体的公司治理,通过内部风险管理的优化持续改进相关问题。这一监管方式可给予科技企业更大的操作空间,激励企业运用内部信息优势进行治理创新,寻找替代性监管措施以更为经济高效的方式实现预期结果。⑥ 因此,与事后惩戒式监管相比,深入企业内部治理,要求企业建立科技伦理委员会,系统化构建内部伦理审查机制是更为灵活的监管方式。为此,一方面可借鉴我国生命科学和医学伦理制度,在《关于加强科技伦理治理的意见》基础上,加速推进适用于生成式人工智能关键科技领域的伦理框架。⑦ 另一方面,还可遵循通过后的《科技伦理审查办法(试行)》督促企业成立人工智能伦理委员会,确保在设计阶段嵌入基础伦理原则,引导投身生成式人工智能的科技企业对内部研发、应用活动构建常态化的治理约束。

(三)监管措施的兼容一致

在人工智能的“1.0”时代,人工智能模型的碎片化明显,泛化能力十分不足,以主体、应用、场景为基准的分段式监管设计尚可应对。但自 18 年起,大模型快速流行,以 ChatGPT 为代表的生成式人工智能开启了人工智能的“2.0”时代。如本文前述,在这一技术发展阶段,基于风险、主体和应用

① 泽南:《王慧文与一流科技达成并购意向,聚力做中国版 OpenAI》,载机器之心 2023 年 3 月 28 日,https://baijiahao.baidu.com/s?id=1761596140121342975&wfr=spider&for=pc。

② 许可:《个人信息治理的科技之维》,载《东方法学》2021 年第 5 期,第 64-67 页。

③ 朱悦:《从法的实验到实验的法:A/B 实验对互联网治理的变革效应》,载《网络信息法学研究》2021 年第 2 期,第 68 页。

④ Michael Kearns & Aaron Roth, Ethical Algorithm Design Should Guide Technology Regulation, Jan. 13, 2020, https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/, last visited on April. 10, 2023.

⑤ 滕妍等:《通用模型的伦理与治理:挑战及对策》,载《中国科学院院刊》2022 年第 9 期,第 1296 页。

⑥ Cary Coglianese & David Lazer, Management-based Regulation: Prescribing Private Management to Achieve Public Goals, 37 Law & Society Review 691, 694-696(2003).

⑦ 倪思洁、韩扬眉:《科技伦理治理需破壁垒、明权责、共参与》,载《中国科学报》2022 年 3 月 22 日,第 1 版。

的治理范式均在不同面向上显现出治理局限。对于新型监管范式的探索亟须引入治理思维,将人工智能监管视为多元主体构成的开放式整体系统,关注构成系统的子系统之间、该系统与其他系统的合作、兼容、配适与转化,提升各系统之间的互操作性(interoperability)。① 互操作性概念最初应用于经济领域,特指产品标准的兼容性,而后扩展应用至网络系统的互联互通,并投射至人工智能监管领域成为创新性监管理念。② 人工智能监管互操作性则是指两个或者多个监管主体互联互通,在共享监管信息的同时,互认监管措施,保持监管实现过程与监管目标和价值一致性的能力。③ 对于生成式人工智能而言,监管技术和监管规则的互操作性可以为科技企业提供监管便利化优势,通过监管机构之间的互联互通提升整体实效。

因循这一理路,人工智能监管的互操作性成为各国共同关注的重点。例如,美国第 13859 号行政命令指定管理与预算办公室负责人工智能监管机构间的协调工作,定期发布人工智能监管备忘录。④ 2020 年 11 月,管理与预算办公室发布《人工智能应用的监管指南》,提出各监管机构应共同遵守的十项准则,为监管一致性奠定行动框架。⑤ 英国也建立了中央协调机制,开展人工智能风险跨部门监测和评估,为协同性地应对人工智能新型风险提供决策依据。⑥ 在国际层面,各国也在积极地就人工智能治理问题达成一致,构建可互操作的治理框架。例如,欧盟-美国⑦、欧盟-日本⑧及美国-新加坡⑨均就人工智能治理问题达成了双边合作,推动人工智能治理的趋同与对标。

聚焦到我国监管实际,面向生成式人工智能监管的兼容配适应主要在两个层面展开:第一,创建人工智能技术监管协同机制,以协同化监管应对监管碎片化和监管冲突等问题。长期以来,我国在监管权划分的基础上形成面向不同行业的专业化监管格局。面对横贯上中下游的生成式人工智能,分立运行的监管格局可能因知识水平参差不齐、监管方式形态各异形成监管竞次、监管真空等问题。⑩ 监管协调不畅还可能抑制创新与竞争。⑪ 因此,可依据《新一代人工智能治理原则》制定跨

① 赵鹏、张力伟:《系统论视角下政府治理的基本逻辑》,载《系统科学学报》2022 年第 1 期,第 78-79 页。

② 靳思远、沈伟:《DEPA 中的数字贸易便利化:规则考察及中国应对》,载《海关与经贸研究》2022 年第 4 期,第 5 页。

③ 靳思远、沈伟:《DEPA 中的数字贸易便利化:规则考察及中国应对》,载《海关与经贸研究》2022 年第 4 期,第 5 页。

④ Executive Office of the President, Maintaining American Leadership in Artificial Intelligence, Feb. 11, 2019, https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence, last visited on April. 10, 2023.

⑤ Russell T. Vought, Guidance for Regulation of Artificial Intelligence Applications, Nov. 17, 2020, https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf., last visited on April. 10, 2023.

⑥ Secretary of State for Science & Innovation and Technology, A Pro-Innovation Approach to AI Regulation, Mar. 29, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf., last visited on April. 10, 2023.

⑦ Suzanne Smalley, U.S. and EU to Launch First-of-its-kind AI Agreement, Jan. 28, 2023, https://www.reuters.com/technology/white-house-european-commission-launch-first-of-its-kind-ai-agreement-2023-01-27/, last visited on April. 10, 2023.

⑧ European Commission, Japan-EU Digital Partnership, May. 12, 2022, https://www.consilium.europa.eu/media/56091/%E6%9C%80%E7%B5%82%E7%89%88-jp-eu-digital-partnership-clean-final-docx.pdf., last visited on April. 10, 2023.

⑨ U.S. Department of Commerce, Joint Press Release: New Collaboration under the U.S.-Singapore Partnership for Growth and Innovation (PGI), Mar. 9, 2022, https://sg.usembassy.gov/joint-press-release-new-collaboration-under-the-u-s-singapore-partnership-for-growth-and-innovation-pgi/, last visited on April. 10, 2023.

⑩ 吴风云、赵静梅:《统一监管与多边监管的悖论:金融监管组织结构理论初探》,载《金融研究》2002 年第 9 期,第 82 页。

⑪ Secretary of State for Science & Innovation and Technology, A Pro-Innovation Approach to AI Regulation, Mar. 29, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf.

机构的人工智能监管协调框架,通过合作制定大型语言模型开发、部署和运行阶段的关键监管要点。此外,还可以通过多机构会签、监管信息共享等制度探索监管机构间互操作性保障机制,为应对交叉风险提供规则基础。①

第二,探索多元化的监管一致性工具。首先,建议引入“模块治理”理念,制定面向生成式人工智能的一般性监管方法,对科技企业的内部治理、决策模式、运营管理以及用户关系管理等重点方面以技术测试和过程检查的形式向科技企业提供明确的监管指引,通过厘清交叉监管中的“共同模块”促进人工智能监管框架的互操作性。② 其次,成立人工智能合作监管联盟提升监管互操作性和监管韧性。例如,英国信息专员办公室、通信办公室、竞争与市场管理局、金融行为监管局共同成立了数字监管合作论坛(Digital Regulation Cooperation Forum),通过加强监管机构间的合作提升监管一致性。该联盟近期的一项核心议题便是通过人工智能监管沙箱为科技企业提供定制化的联合监管建议,以帮助其尽快进入目标市场。③ 我国金融监管领域很早就监管合作机制展开了探索。2021 年,两办发布的《关于依法从严打击证券违法活动的意见》中亦对跨部门监管协同、跨境审计监管合作和执法联盟等提出重要意见。生成式人工智能可能引发跨行业、跨市场、跨领域的交叉性风险,势必需要多家监管机构协同攻关。此次《党和国家机构改革方案》出台,不仅组建了中央科技委员会,重组了科学技术部,还组建了国家数据局。这意味着监管机构的架构调优对于国家创新战略的实现具有重要意义。因此,从完善创新监管架构的视角出发,也需要各监管机构探索提升监管互操作性和一致性的常态化制度路径。最后,为确保监管效能与生成式人工智能的技术革新与应用风险相匹配,还应以监管资源分配、监管一致性以及监管协同效果为基准建立监管风险识别和评估框架,联通与科技行业、终端用户等主体的反馈回路,全面评估监管协同框架的有效性。④

五、结语

ChatGPT 席卷全球,为人工智能产业注入活力与动能的同时,既带动了生成式人工智能的快速爆发,也深刻地改变了产业竞争格局和全球科技力量对比。现阶段,无论是国际行业巨头,还是国内头部企业,均聚焦大模型领域持续发力。可以说,2023 年既是生成式人工智能的元年,也是人工智能监管范式变革的元年。如今的全球科技竞争不仅是新型智能基础设施的角逐,更是数字文明和创新制度生态的比拼。与美国相比,我国的人工智能领域缺少重大原创成果,在核心算法、关键设备等方面仍存较大差距。⑤ 本文提出了治理型监管范式,希冀以监管权的开放协同弥补传统监管

① 我国银监会和保监会还于 2007 年签署了《关于加强银保深层次合作和跨业监管合作谅解备忘录》,通过正式的合作安排提高监管协调有效性。参见王兆星:《机构监管与功能监管的变革——银行监管改革探索之七》,载《IMI 研究动态》2015 年合辑,第 6 页。

② 谷兆阳:《新加坡人工智能监管模式框架》,载《南洋资料译丛》2019 年第 4 期,第 57 页。

③ Secretary of State for Science & Innovation and Technology, A pro-innovation approach to AI regulation, Mar. 29, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf., last visited on April. 10, 2023.

④ Secretary of State for Science & Innovation and Technology, A Pro-Innovation Approach to AI Regulation, Mar. 29, 2023, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf., last visited on April. 10, 2023.

⑤ 高蕾等:《我国人工智能核心软硬件发展战略研究》,载《中国工程科学》2021 年第 3 期,第 94 页。

覆盖不足的缺憾,以治理与监管理念的多元融合补足监管介入迟滞的短板,以监管措施的互操作性助力我国科技企业在生成式人工智能技术赛道上的加速健康发展。如何能够在创新驱动发展战略下,构建良好的制度生态从而服务于我国科技企业的发展,形成创新与治理双轮驱动,软硬结合、梯次接续的技术治理格局,是一项亟须改变、影响未来、意义深远的议题。未来,还需从监管基础设施、监管工具体系、监管架构变革以及监管制度环境等核心要素出发,加以持续不断的探索与优化。ML

Algorithmic governance Challenges and Governance Supervision in Generative Artificial Intelligence

ZHANG Xin

(University of International Business and Economics, Beijing 100029, China)

Abstract: As exemplified by ChatGPT, large-scale pre-trained language models are progressively showcasing their potential. The technical features of these models, including their vast scale, substantial parameter count, extraordinary scalability, and varied application scenarios, present comprehensive challenges to the algorithmic governance system centered around transparency, fairness, and accountability. In the prevailing global AI governance paradigms, the EU has instituted a risk-based governance framework, China has devised a subject-based governance approach, and the U. S. has embraced an application-based governance method. These three governance paradigms emerged during the "1. 0 era" of narrow AI. Consequently, as AI technology evolves, it is crucial to foster a holistic transformation of the regulatory structure, characterized by open collaboration in regulatory authority, the incorporation of diverse regulatory approaches, and the harmonization of regulatory measures. This shift will enable a transition towards "governance-oriented regulation" fitting for the AI "2. 0 era".

Key words: ChatGPT; generative artificial intelligence; algorithmic governance; governance-based regulation

本文责任编辑:董彦斌

青年学术编辑:任世丹