DOI:10. 3876/j. issn. 1671 - 4970. 2024. 04. 009

AIGC 赋能新质生产力的数据风险及其敏捷治理

郑煌杰

(中南大学法学院,湖南 长沙 410083)

摘要:党的二十届三中全会进一步强调了加快培育新质生产力,其重要内容之一是完善生成式人工智能(AIGC)的发展和管理机制。随着"人工智能+"行动的开展,以 ChatGPT、Sora、文心一言等为代表的 AIGC 将逐步实现商业化落地,成为新质生产力的重要质态。然而,AIGC 在提高人们生产力与创造力的同时,亦引发了诸多数据风险。基于其技术原理,可将这些风险阶段化划分为数据非法获取风险、数据安全风险、不良内容生成风险。作为一种新型治理模式,敏捷治理是对彼时探索式、回应式、集中式等传统治理模式的省思与超越,将其嵌入 AIGC 治理,AIGC 在治理理念、治理主体、治理对象、治理工具等方面都将有新的内涵特性。AIGC 适用敏捷治理,既是顺应智能革命的趋势、贯彻科技向善的要求,亦能摆脱传统治理困境,而国家政策方针的指引亦为其适用营造了良好的法治环境。在敏捷治理的引导下,应贯彻风险预防原则,加强数据风险应对机制;强化算法解释权,构建数据安全保障体系;完善应用生成机制,推进技术融通型法治,以有效纾解 AIGC 数据风险,赋能新质生产力发展的同时参与引领全球治理体系改革和建设。

关键词:新质生产力;ChatGPT;生成式人工智能;敏捷治理;数据风险

中图分类号:D923 **文献标志码**:A **文章编号**:1671 - 4970(2024)04 - 0089 - 14

历史进程的每一次跃进,皆根植于生产力的持续演进。新质生产力标志着对现行生产力范式的变革与重塑,其核心动力源自科技领域的根本性跃迁,亦伴随着生产要素配置策略的根本性调整以及产业结构深层次的重构与优化升级,彰显了生产力演进的前沿趋势和高级阶段特征。同时,随着"人工智能+"战略的实施,诸如 ChatGPT、文心一言、Sora 等生成式人工智能应用(下称"AIGC")不断涌现并持续升级迭代,预示着我国数字技术即将步入快速发展阶段,进而为加快形成新质生产力提供强大驱动力。然而,当前 AIGC 在进行大规模数据收集、分析等活动时,易引发诸多数据风险:2023 年,北京互联网法院作出国内第一例 AIGC 生成作品侵权纠纷的判决;2024 年,广州互联网法院作出全球首例 AIGC 服务侵权生效判决;同年,北京互联网法院一审开庭宣判了全国首例 AIGC 生成声音人格权侵权案。这一系列案例表明,AIGC 引发的数据风险不仅引起了人们的关注与担忧,也可能随之发展而持续发酵。鉴于此,网信办联合国家发改委、科技部、教育部等七部门出台了《生成式人工智能服务管理暂行办法》,其中明确了 AIGC 的适用范围、责任主体认定、服务规范等方面内容。2024 年 7 月世界人工智能大会于上海召开,正式发布了《人工智能全球治理上海宣言》,提出要促进 AI 发展,维护 AI 安全,构建 AI 的治理体系等内容。

引用本文:郑煌杰. AIGC 赋能新质生产力的数据风险及其敏捷治理[J]. 河海大学学报(哲学社会科学版),2024,26(4):89-102.

基金项目:国家社会科学基金项目(22VRC175);中南大学研究生科研创新项目(1053320231259、1053320230720)

作者简介:郑煌杰(1995—),男,博士研究生,主要从事数字法学研究。E-mail:Yuppiejay@ 163. com

与此同时,学界对 AIGC 风险治理问题也展开了一系列研究:有学者提出应当建立 AIGC 风险治理的元规则[1],或是从法律与技术层面共同推进 AIGC 的可信治理[2],或是基于算法安全审查视角展开对 AIGC 风险的治理[3]。亦有个别学者针对 AIGC 生成内容的信息风险展开研究,但忽视了信息风险本质还是源于数据风险[4]。简言之,当前大多数研究仅仅是在分析 AIGC 时附带提到数据风险治理问题,并未对其进行系统化研究,由此导致不能深入阐释 AIGC 数据风险治理的特殊性,难以提供具有针对性、精准性的治理方案。因此,基于 AIGC 的技术原理,本研究阶段化地剖析其引发的数据风险,并以国内外现行数据治理立法与实践为依据,对 AIGC 治理模式的选择展开学理论证,并提出契合本土实情的治理框架,以期为加快形成新质生产力推动高质量发展贡献智慧借鉴。

一、AIGC 的技术原理与数据风险:以 ChatGPT 为例

自 Vaswani 等于 2017 年发表开创性论文以来,ChatGPT 所依托的技术基石——Transformer 架构,便在 AI 应用领域中广泛传播,并展现出深远的影响力。然而,引发学界与业界的疑问是:ChatGPT 如何能利用自然语言处理技术(NLP)与人类进行流畅且精确的交互对话?因为对这一问题的探讨,不仅是理解 AIGC 的关键,亦是剖析其数据风险成因的基础。尽管 OpenAI 尚未直接发布关于 ChatGPT 的具体技术细节,但通过其官网资料可知,ChatGPT 采用了与 InstructGPT 相同的技术机制,主要区别在于数据收集方法上有细微调整[5]。因此,可基于 InstructGPT 的技术特性为参照,探析 AIGC 的运作机制。

1. ChatGPT 的技术透视

(1)自回归生成模型:流畅生成的关键

自回归生成模型与 BERT(bidirectional encoder representations from transformers)模型所采纳的遮蔽语言模型(masked language model)截然不同,后者主要通过遮蔽句子中的某些元素,并依靠上下文来推断这些被遮蔽部分,实现了一种双向的学习过程,从而提升对词义和语法结构的理解能力。尽管 BERT 在文本理解方面展现出卓越的性能,但其结构并不适合直接生成自然语言,不能无缝地与用户进行交互对话。不同于 BERT,ChatGPT 采用的自回归生成模型(autoregressive model)是一类单向语言模型[6],主要利用一种连贯的文字预测过程,即根据当前的上文和概率分布,从多个可能的语素中选择并生成下一语素,如此循环,直至形成一个完整的文本或达到设定的最大长度。与 BERT 相比,尽管 ChatGPT 在训练过程中对数据的需求量较大,但自回归生成模型的运用使得其能够流畅地生成自然语言,从而实现与用户进行交互式对话。

(2)预训练:准确生成的基础

预训练指的是通过使用大规模的数据集和通用性质的任务来进行的第一阶段训练,旨在使机器学习模型掌握具有广泛适用性的参数,从而提升模型的泛化能力[7]。以 OpenAI 的 GPT 模型为例,其数据集规模从 2018 年的 5GB 到 2020 年达到了 45TB,之间经历了数量级增长,这为 ChatGPT 展现出强大能力奠定了坚实基础。此外,OpenAI 还采取半监督学习(semi-supervised)方式,即将预训练过程分为两个步骤:第一步是无监督预训练阶段,第二步是微调阶段。第一步的目的是为模型寻找一个优良的初始化状态。通过对大量未标注数据的学习,使得模型初步掌握文本聚类、话题模型构建、机器翻译等方面能力[8]。然而,由于缺乏明确的指导,这一阶段可能会导致模型难以区分有用与无用信息,所产生结果的稳定性和可靠性较差。对此,OpenAI 在第二步中采用了微调技术,通过利用标注过的数据集对模型进行优化,向模型提供标注的虚假或有害信息数据集,增强其识别假信息和有害信息的能力,进而降低生成不良或非法内容的可能性。

(3)人类反馈强化学习:优化生成的手段

早在 2020 年,OpenAI 就推出了第三代生成式预训练变换器(GPT-3),但其在运行过程中仍存在一些局限性,比如在回答问题时容易偏离主题、生成不准确甚至是虚假的信息[9]。对此,OpenAI 引入了人类反馈强化学习(reinforcement learning from human feedback,

RLHF)技术,即通过更精细化的反馈机制来提高生成模型的输出质量,从而使生成的文本更加贴近人类的预期和需求,这一技术主要分为两个核心阶段:奖励模型(reward model)阶段和近端优化(proximal policy optimization)阶段[10]。在前者阶段,训练人员会对 GPT-3 模型生成内容的质量和相关性进行评估,根据评估结果向模型提供正向或负向的激励信号,以减少个体偏好对评估结果的影响,也就是说,通过引入客观性的评判标准来提高整个过程的公正性和透明度;在后者阶段,将这些排序和评估结果反馈给模型,以便在算法层面调整和优化模型的参数设置,使模型能够更准确地预测和生成用户期望的答案,同时降低生成内容的偏差率。

综上,可将 AIGC 的运作机制大致划分为收集阶段、分析阶段和输出阶段。其中,自回归生成模型是 3 个阶段的前提基础,"预训练"可视为收集阶段的组成部分,涉及大规模数据的收集与初步模型构建,以便从中提取模式和规律;"微调"更贴近分析阶段,AIGC 利用自身算法模型来处理输入数据,进行优化训练;"强化学习"则在分析阶段与输出阶段均有体现,根据算法模型的反馈来改进输出结果,以生成高质量内容。

2. AIGC 的数据风险

目前我国数据治理体系主要是建立在《中华人民共和国数据安全法》(以下简称《数安法》)《中华人民共和国网络安全法》(以下简称《网安法》)《中华人民共和国个人信息保护法》(以下简称《个保法》)体系的基础上,要求数据处理者应以保障网络安全、数据安全、个人信息安全为原则,进行数据的收集、使用与处理等活动。在此前提下,基于对 AIGC 技术原理的剖析,可将其引发的数据风险归纳以下阶段性风险:收集阶段,AIGC 通过"被动输入"与"主动爬取"构建语料库,此过程主要关乎数据的合法获取,也隐含着数据泄露、版权争议等初始风险;分析阶段,算法模型在数据处理与优化中主要存在"数据偏见"与"运行时泄露"风险,亦反映出技术操作背后的价值倾向与安全防护的缺失;输出阶段,生成内容的唯一性虽赋予了 AIGC 独特价值,却也使其成为不良内容生成的风险高发区。进言之,AIGC 技术阶段与风险阶段之间的逻辑关系体现为,风险随技术活动的发展而动态演变:收集阶段的风险聚焦于数据的获取与存储,强调合法性与伦理界限;分析阶段的风险转移至算法处理,数据处理的公平性与安全性成为核心;输出阶段的风险则直接关联生成内容的可控性,要求生成内容不违反社会秩序。

(1)收集阶段:"被动输入"与"主动爬取"引发的数据非法获取风险①

AIGC 需要大量数据训练模型,才能建立相应的语料库,语料库是一种存储大量语言表达形式的数据集合,是 AIGC 训练和学习的基石。构建与更新语料库的方式,通常包括"被动输入"与"主动爬取"两种:就前者而言,用户通过对话框输入信息并由系统保存,这些"被动输入"的数据会被整合到语料库。这种方式不仅能捕捉到用户的真实需求,也可以在对话中根据用户的反馈对语料库进行优化和扩充。从后者来看,AIGC 通过"爬虫技术"自动收集互联网上的大量数据,而数据来源的多样性和广泛性使得其既能获取不同主题领域的数据内容,为各种问题提供多样的答案选项,也能确保模型的训练更为高效与准确。然而,无论是"被动输入"还是"主动爬取",均有可能引发 AIGC 训练数据的非法获取风险。

在"被动输入"过程中,用户往往无法确切掌握其信息的收集与处理情况。虽然 OpenAI 在用户协议中规定了对个人信息的收集不得超出法律和伦理的限制,并负有相应的"删除"义务,但未明确说明删除的方式及具体应如何行使该义务,由此将引发一系列风险:其一,用户的个人信息权益可能受到侵害。根据《个保法》第 45~50 条规定,信息主体享有知情同意权、删除权等权益。但由于当前 AIGC 缺乏明

① 需要强调的是,无论是本研究指称的"被动输入"还是"主动爬取",两者引发的风险包括但不限于数据非法获取、隐私泄露、版权等风险,但每个技术阶段的"风险占比"是有所不同的,换言之,每个技术阶段风险也会有所"重叠"。因此,在不同技术阶段,本研究强调的是此技术阶段的"最大风险"(下文亦同),以此进行详细论述与精准施策。比如,版权风险的来源根本在于数据的非法获取,正是因为数据存在非法获取,所以才引发版权争议,即在收集阶段,一个是表象(版权、数据泄露等风险),一个才是内因(数据非法获取风险)。

确的删除操作流程,个人信息可能会被保留并用于模型训练,从而加大了个人信息侵权风险。其二,可能降低用户对 AIGC 的可信度。如果用户对个人信息的删除方式和行使途径缺乏透明和清晰的了解,将难以相信其信息能得到妥善处理,进而不再积极主动地提供真实的数据,这将可能降低训练数据的可靠性与真实性。其三,版权风险亦不容忽视。AIGC 在训练中整合的海量数据若包含受版权保护的内容,未经权利人许可使用,将可能构成侵权。即便"用户协议"试图限定责任,但模糊的删除机制无法有效排除违法使用的可能性,这无疑加剧了版权侵权风险。

数字时代,数据已经被公认为是一种重要的生产要素。因此,在进行"主动爬取"数据的过程中,各国法律普遍承认"正当爬取"行为的合法性。然而,数据爬取并非不受任何限制,应当确保其在合理的范围内进行。我国已有法律对网络爬虫规制主要集中于《中华人民共和国刑法》有关计算机信息系统犯罪第 285 条、第 286 条等规定。如果 AIGC 出于非法目的或超出合理性边界进行数据爬取,很可能构成非法获取数据行为。例如,意大利个人数据保护局就因为 OpenAI 非法爬取用户信息,展开了立案调查,并临时封禁 ChatGPT。此外,非法爬取数据还可能涉及国家数据安全和主权问题,倘若 AIGC 将非法爬取的数据从境内传输到境外,将会侵犯另一个国家的数据主权,对该国家安全构成威胁。

(2)分析阶段:数据"偏见"与"泄露"导致的数据安全风险①

就"数据偏见"而言,开发者可通过裁剪数据库或操控算法的方式,将其"偏好"植入 AIGC 的训练数据,从而使生成内容呈现特定的价值观。进言之,AIGC 不是一个完全保持"中立"的技术工具,其"认识"取决于算法设计者或数据训练者的决策。尽管 OpenAI 引入了 RLHF,从模型中选取样本并进行人工"评分"标注,不断改进模型生成结果,最终得到符合人类社会价值观的语言生成模型。然而,这种训练机制仍存在以下风险:一方面,出于某些利益考量,设计者可能会使用具有偏见指向的数据样本训练 AIGC,从而在生成内容中"潜移默化"的影响用户的思想或行为。例如,当 AIGC 训练的数据集事先就存在各种偏见和歧视,其可能源自数据采集过程中的人为因素,也可能来自社会中存在的不公平现象,如果这些歧视性因素未能得到有效处理,那么其生成结果将受到影响,并进一步扩大不公平现象[11]。另一方面,体现于 AIGC 对不同国家文化的包容性不足。AIGC 在分析阶段需要大量的高质量语料库支持,但当前其主要还是基于英文语料库进行训练,这就决定了其在输出结果中体现的价值观和文化色彩与西方文化更为契合。

在"数据泄露"层面,已经有大量的现实案例表明,在与 ChatGPT 交互过程中,用户输入的信息容易被视为迭代数据,进而引发数据泄露风险。根据 OpenAI 的隐私协议,如为了优化 GPT 模型,那么其将享有持续处理个人及衍生数据的权利。然而,倘若用户对 AIGC 的数据处理机制认知不足,则会不经意地提交个人敏感信息。比如,根据数据安全服务机构 CyberHeaven 的检测表明,其客户公司的 160 万名员工中存在 4.2%的员工将包含商业秘密的数据输入 ChatGPT 的情况,还有医生则将患者信息输入系统以生成保险公司信函[12]。另外,为满足不同主体的需求而量身定制服务也是 AIGC 的商业拓展方向。在定制化过程中,设计者既要事先收集与定制化需求相关的数据(如用户的个人信息、爱好等),也要将这些数据分为训练集、验证集、测试集,以实时对模型进行优化与性能评估。比如,当想训练 ChatGPT 写一首五言律诗时,必然要将"每句五个字,共八句,音韵上遵循平起仄收"的创作原则与正确样例输入模型,而后依靠其自我学习与迭代能力完成指定任务。然而,当用户提供的"定制数据"传输到服务器时,可能会引发数据泄露风险。例如,2023 年 OpenAI 官网公告表示,因开源数据库存在的错误导致了缓存出现问题,

① 为避免歧义,本研究的分析阶段主要限定为模型优化与用户交互过程,强调在模型优调和定制化配置阶段之后,用户互动中数据的动态处理包括隐私处理和个性化服务的即时生成。

一些用户可能看到其他人聊天记录的片段以及他人信用卡号码、姓名等信息。最初该问题仅波及个别用户,但由于 OpenAI 在对服务器进行更改时再度出现失误,导致问题进一步加剧,受影响的用户扩大至 1.2%的比例,这也暴露了 AIGC 的数据安全隐患[13]。倘若在定制化过程中发生类似数据安全问题,那么引发的数据泄露风险将给用户带来不可估量的损失。

(3)输出阶段:被不法分子利用诱发的不良内容生成风险①

AIGC 通过深度学习算法,从庞大的文本数据集中汲取语言的特征、规则和模式,这一过程本质上是对数据进行复杂的统计分析,旨在让模型能够理解并模仿人类语言的细微差异。在模型训练阶段,海量文本数据的输入使模型能够捕捉到语言的内在结构,进而掌握语义的连贯性和语法的规范性。然而,这种依赖于历史数据的学习方式也埋下了隐患——由于生成的内容基于对过往规律的总结而非即时情境下的深度理解和逻辑推理,AIGC 提供的信息可能仅代表一种经过筛选和标准化的视角,缺乏对当下语境的敏感度和适应性。换言之,相较于传统的搜索引擎如百度、谷歌,AIGC 生成的内容具有显著的独特性。搜索引擎通常呈现多源信息,为用户提供广泛的参考和对比空间,而 AIGC 则倾向于生成单一、连贯的响应,缺乏对信息渠道的具体标注,这无疑增加了评估内容准确性和来源可靠性的难度。受此影响,不良内容生成的风险悄然滋生,因为不法分子可能利用 AIGC 的这一特性,刻意诱导模型生成误导性或有害信息,而用户往往难以察觉或验证这些内容的真实性和完整性。简言之,AIGC 在输出阶段诱发的不良内容生成风险,源自其技术特性的固有局限。由于模型的预测性质和内容生成的封闭性,AIGC 可能在无意间成为传播虚假信息、偏见言论乃至非法内容的载体,对于监管机构和平台运营者而言,识别和过滤这些风险已成为一项艰巨的任务。

二、AIGC 敏捷治理的逻辑证成:意涵、必要性与可行性

面对 AIGC 等科技风险时,主要有两个应对原则:“成本—效益分析”和“风险预防”,虽然这两个原则各有利弊,但考虑数字技术的高风险性,应将“风险预防”置于核心位置[14]。进言之,目前 AIGC 所显露的数据风险,如同浮于海面的冰山一角,其下潜藏的隐患深广且复杂,倘若初期治理存在疏漏,那么极易诱发后续更广泛的违规行为,加剧风险的扩散与治理难度。因此,构建前瞻性且适应性强的治理框架显得尤为迫切。而传统治理架构通常基于层级分明的管理模式,侧重于事后追责,难以贴合 AIGC 快速迭代与跨界融合的技术特性。同时,传统线性治理模式也无法有效应对 AIGC 涉及的多元主体之间的权利与义务分配问题,易造成治理滞后与责任归属不清的“步调失衡”,这种滞后性不仅无法根治风险,反而可能会抑制技术创新与产业发展。鉴于此,AIGC 治理模式的转型迫在眉睫,需从被动反应转向主动预防,从单一封闭转向协同开放,而敏捷治理的提出,恰是对此挑战的积极回应。

1. AIGC 敏捷治理的意涵阐释

(1)敏捷治理的理论内涵

敏捷治理(agile governance)理论,可追溯至软件工程领域的“敏捷方法”[15]。该理论最初应用于企业管理、制造业等领域,随后扩展至数字经济、新兴产业乃至国家治理层面,形成跨学科的研究态势。对于敏捷治理的理论内涵,学界有着不同理解。过去,针对政府 IT 项目中瀑布模型的局限性,《敏捷宣言》提出了敏捷软件开发方法,强调高效响应客户需求、快速交付价值,以及灵活应对业务和技术变化[16]。2007 年,敏捷治理概念被正式提出,其融合了“快捷”“灵活”“协调”等层面理念[17]。数字时代,数据(信息)碎片化的加剧带来了颠覆性变化,组织面临加速、互联、不确定的外部环境,促使学者开始关注组织层面的敏捷治理。如,有学者将敏捷与协调性、灵活性、适应性相结合,视为政府结构的新程序[18];也有学者基于达布林理论,构建了敏捷治理框架[19];还有学者提出

① 本研究的输出阶段,主要是基于其生成内容分析风险,而不是纯粹基于技术层面展开,如前所述,在技术层面,输出阶段与分析阶段是有重合地带的,但就风险占比而言,输出阶段风险主要体现为不良内容生成风险。

了包含十大要素的敏捷治理框架,如维持稳定性与灵活性、快速动员合作、跨部门合作等[20]。2018 年世界经济论坛将敏捷治理定义为:“以顾客为中心,具有柔韧性、流动性、灵活性的行动或方法。”[21]可见,尽管不同领域及学者对敏捷治理的定义各异,但都认识到其是对公共价值的回应,强调多元主体的沟通协作,快速感知响应变化,使治理更具包容性,满足多元需求,以有效应对数字时代挑战。基于此,可将敏捷治理界定为:要求政府、企业、社会组织、公民等主体参与治理,强调多元性、适应性、包容性、灵活性,平衡技术风险与创新,以增进人类福祉为宗旨的治理模式。

(2)AIGC 敏捷治理的核心特性

在上述分析基础上,倘若将敏捷治理的理论框架应用于 AIGC 领域,可从治理理念、治理对象、治理主体及治理工具 4 个维度探析 AIGC 敏捷治理的核心特性:

第一,治理理念的“适应性”。AI 技术的快速发展和应用推动了传统治理向敏捷治理的演进。与传统治理聚焦于“事后”规制和响应不同,敏捷治理提倡创新与规制的并行发展,要求政府在其中发挥调和作用,以消解发展与创新之间的“创新悖论”“生产率悖论”①。换言之,敏捷治理基于预测性(预防)弹性管理,能够使政府部门具备预防 AIGC 风险的能力,以在促进科技发展的同时保障社会公共利益,突破传统治理下的创新瓶颈。

第二,治理对象的“包容性”。AIGC 敏捷治理主张从底层数据、算法到应用平台的综合治理,建构一个全面而分层的治理体系。这种体系既考虑治理对象的多样性,也重视个性化和普遍性的需求。在我国现行 AI 治理框架中,就明显呈现这种“通用底线+特定场景定制”的治理策略。如,针对数据或算法等突出问题进行规制,政府出台《数安法》《互联网信息服务算法推荐管理规定》等法律规范。此外,敏捷治理亦主张超越传统官僚体系下严苛的层级治理模式,通过考量各方利益,特别是边缘群体的需求,以实现更为人性化的治理方案,从彼时“以事为本”转变至“以人为本”的服务理念。

第三,治理主体的“多元性”。敏捷治理倡导构建一个多层次的治理体系,其中政府、企业、社会及公民等主体均能参与治理过程,形成协作共赢的治理局面。如哈佛大学公共政策教授 Nye 在关于“软权力”的研究中指出,非政府主体在全球治理中的作用日益增强,这一观点同样适用于 AIGC 敏捷治理[22]。通过建立开放的平台和通道,使得利益相关者能够共同参与决策过程,增强政府政策的透明度和公众信任度,从而制定更加科学化、民主化的治理政策。

第四,治理工具的“灵活性”。在治理工具的选择上,敏捷治理强调软硬结合、灵活多变的策略,旨在实现有效且有弹性的治理。通过在宏观法律法规、中观伦理标准与微观自律机制之间的灵活调整,以高效应对快速变化的市场环境和技术革新挑战,从而达到一种弱干预、强效果的治理目标。这种模式不仅仅是对传统治理模式的改进,更是对治理思路的根本创新,促使治理逻辑从静态合规转向动态互动,进而为 AIGC 治理注入新的动力。

2. AIGC 敏捷治理的必要性

敏捷治理并非政策制定者或学界专家基于理论推演所产生的,而是由当前 AI 治理体系面临适应性不足的现实问题所倒逼出来的,亦是对数字时代 AI 治理的复杂性、整体性和迫切性挑战的回应。

(1)顺应智能革命的趋势

在历经机械化、电气化及信息化三次重大的科技革命后,当前人类社会正深处一场以 AI 为先导的智能革命之中。以 ChatGPT、Sora 等为代表的 AIGC 应用创新,不仅加速各行各业的智能化转型,也成为推动数字技术深化应用和社会变革的重要推动力。审视当下,各国对于 AI 治理策略映射出不同价值观与政策导向的碰撞,如欧盟倾向于采用伦理先行的集中式治理路径,强调对个人数据权利的尊重与隐私

① 创新治理悖论是指人类社会会通过维护科技创新体系来寻求一种安全稳定的发展环境,但科技创新活动的开展也会对当前的社会环境造成一定程度的冲击乃至破坏。这个悖论主要包含以下两个问题:如何制定科技创新发展规划以及如何监管科技创新活动。生产率悖论则指的是在技术进步和应用的情况下,生产率并没有相应提高的现象。

保护,力求在严格的规范框架内引导技术向善。相比之下,美国彰显出市场驱动的特点,偏好分散式的灵活监管,鼓励创新优先,放手企业自由探索,期望在竞争中激发技术潜力与经济增长。然而,这两种治理模式均存在一定的局限性,集中式治理虽在伦理框架构建上严谨周密,却可能因决策链条长、调整迟缓,难以适应瞬息万变的技术迭代,导致规制滞后的尴尬;分散式治理虽赋予了市场高度自由,利于快速试错与创新,却也可能因监管松散,难以有效应对 AIGC 带来的跨域数据风险与伦理挑战,如“数据偏见”、隐私泄露等,最终损害公共利益[23]。因此,无论是欧盟的伦理至上还是美国的创新优先,单一模式皆难以完美应对 AIGC 的治理挑战,由此呼唤更为灵活、前瞻且兼顾安全与创新的敏捷治理模式,以实现技术进步与社会福祉的和谐共生。

(2)贯彻科技向善的要求

2023 年,习近平主席在第三届“一带一路”国际合作高峰论坛开幕式上发表主旨演讲,指出愿同各国加强交流和对话,共同促进全球人工智能健康有序安全发展[24]。因此,在充分释放 AIGC 技术红利的同时,还应准确识别并积极应对伴随而来的潜在风险,实现技术创新与治理之间的平衡,并为其提供一个规范有序的发展环境。在此前提下,与传统治理相比,敏捷治理具备以下优势:

首先,在于对不确定性风险的快速响应机制。在 AIGC 快速发展背景下,数据风险呈现出前所未有的复杂性和动态性,传统治理往往基于既定规则和流程,难以迅速适应环境变化。相比之下,敏捷治理通过构建实时监测与反馈机制,能够迅速捕捉到市场、技术、政策乃至公众态度的微妙变化,进而动态调整治理策略,确保治理措施始终与外部环境保持同步。其次,在风险认知上更为“成熟与理性”。传统治理通常追求“零风险”的理想状态,但现实世界中,特别是在创新密集型的 AIGC 领域,完全消除风险几乎不可能,且过度规避风险可能导致创新停滞。敏捷治理倡导在风险与收益之间寻找平衡点,通过引入风险评估、压力测试等工具,对不同风险进行分级管理,实施差异化、精细化的规制策略,这种做法既保障了创新的活力,又确保了风险的可控性,具有较强治理的前瞻性和灵活性。再者,在于强调多元主体参与和协同治理的生态系统构建。传统治理依赖单一中心化的决策结构,忽视了多元利益相关者的角色与影响。AIGC 治理涉及技术开发者、数据提供者、用户、监管机构等众多主体,每一方都是治理链条上的关键节点。敏捷治理鼓励这些主体共同参与治理过程,通过协商、合作建立共识,形成一种网络化、多层次的治理结构。如此,既能够充分利用各方资源和专长,增强治理的针对性和有效性,还能促进治理决策的民主化和社会包容性。

(3)摆脱传统治理困境的需要

AIGC 时代的到来,逐渐使得传统治理陷入适用困境:其一,治理流程的适应性难题。AIGC 的发展必然会对国家治理制度提出调适与转型要求,若治理程序仍保持一成不变,易阻碍其持续创新。例如,基于“命令—执行”的传统治理,在面对 AIGC 数据风险时往往显得滞后甚至过时,从而易引发监管方面的“信息鸿沟”。官僚体制下的管理思维,难以适应智能社会的扁平化、非中心化和平等参与的新特征。进言之,在技术初始阶段管控难度小、影响力有限,而一旦技术广泛应用,调控成本与难度将急剧增加。因此,为防止人类社会被数字技术所“反噬”,AIGC 采取敏捷治理势在必行。

其二,管理模式的单一化局限。传统治理普遍呈现出一种倾向于单一化和纵向集中的管理特征,忽略多元主体共同参与的价值。审视我国 AI 治理模式的发展历程——从早期无明确治理主体的探索式治理(2016 年以前),到以国家部门独立性领导的回应式治理(2017—2019 年),再到多部门合作的集中式治理(2020—2021 年)——虽逐步成形,但广泛动员多元参与者的能力仍显不足[25]。在智能社会下,政府及各部门单向度的监管模式已难以有效管理 AIGC 所带来的风险,亟需政府、企业、社会组织等主体共同参与治理。

其三,传统责任理论的适用性不足。随着通用大模型的广泛应用,传统责任理论是否还

能适用值得深思:一方面,根据《生成式人工智能服务管理暂行办法》第三章相关规定,AIGC 的开发者、使用者和提供者需承担各自不同的责任,但在不同场景下各主体责任具体应如何划分与界定往往复杂多变。另一方面,AIGC 算法的不透明性、难以问责性、不可解释性等问题,均是现阶段技术层面与法律层面所面临的困难,这些因素进一步加剧治理责任的共识难题。对此,敏捷治理的核心价值之一就是在于能够适应 AIGC 的动态发展,通过不断迭代和调整解决治理问题,而非僵化地适用传统责任理论。

3. AIGC 敏捷治理的可行性

在面对 AIGC 的快速发展带来的不确定性,党的二十届三中全会明确提出要“完善生成式人工智能发展和管理机制”,我国正积极致力于更新并完善法律法规、政策规范,以满足 AIGC 产业与智能社会的需要,这亦为 AIGC 敏捷治理营造了良好的法治环境。一方面,针对 AI 特别是在算法、数据处理等关键领域,我国进行了专门立法,发布相关伦理准则、法规等,为 AI 相关活动提供明确的行为指引。例如,我国在 2019、2021 年分别发布了《新一代人工智能治理原则——发展负责任的人工智能》和《新一代人工智能伦理规范》,其宗旨就在于把伦理和道德原则贯穿 AI“设计—研发—应用”的全周期。此外,根据 2022 年由中共中央办公厅发布的《关于加强科技伦理治理的指导意见》,AI 治理需遵循伦理优先、灵活治理、开放合作等伦理原则。另一方面,国家相关监管部门也针对 AI 的应用管理,尤其是在资源访问、实体认证、权利与责任分配等关键领域进行立法工作,确立 AI 相关活动的基本原则和责任归属,以促进敏捷治理的实现。例如,2021 年,国家网信办等联合多部门发布《关于加强互联网信息服务算法综合治理的指导意见》,要求企业全面提升 AI 治理能力;针对 AIGC 所带来的法律风险,《生成式人工智能服务管理暂行办法》亦明确了其概念、服务者提供者义务等内容,并进行专项监管。

三、敏捷治理下 AIGC 数据风险的防范策略

如前所述,针对 AIGC 正逐步呈现的弥散性扩张数据风险,构建全方位、多维度、重实效的数据安全治理体系已刻不容缓。因此,亟需充分发挥敏捷治理的适应性、多元性、灵活性等优势,为 AIGC 数据风险提供一个强有力的治理框架,形成一个系统化、综合性的应对机制,以促进新质生产力的形成与发展。

1. 贯彻风险预防原则,加强数据风险应对机制

敏捷治理的“适应性”侧重“风险预防”,这在如今大规模数据流动的时代背景下尤其关键。2022 年国际标准化组织(ISO)在发布的 ISO 31073:2022《风险管理——术语》中提出,避免、接受、转移和控制等风险管控策略。在面对 AIGC 带来的数据风险时,需要承认在客观上无法完全消除风险这一事实,但能通过不同的治理策略来预防与控制风险,实现安全与发展的平衡。基于风险来源层面,仅对重要数据进行关联和分析仍不够,还需将数据流动和聚合的整个过程有效地控制好,贯彻风险预防原则,事先防范 AIGC 对数据的“有意”或“无意”地整合与分析。

一方面,对于“被动输入”获取的数据,应完善 AIGC 市场准入制度与管理措施,从源头上预防与控制数据风险。首先,对 AIGC 采取“分级审批”的市场准入机制。通过借鉴欧盟《人工智能法案》的做法,将 AIGC 分为不同等级,相应的等级会关联到具体风险程度。对于高风险等级的应用,开发机构需提交符合规定的评估报告,同时也需要接受相关机构的审查和批准;“中低风险”的监管力度则可“依次递减”。如此,就能够根据不同等级的 AIGC 数据风险,制定契合的监管措施,为 AIGC 的市场准入规范提供基础准则。其次,优化 AIGC 的数据获取和使用规则。AIGC 的训练数据会牵扯到复杂的标注、监管、隐私保护和责任承担等问题,亟须通过数据获取和使用规则的规范化来最大保障相关主体权益。鉴于此,可通过标明 AIGC 语料库数据来源的方式,防止非法数据获取行为的发生。同时,开发机构也应主动采取数据清洗和数据去标识化等技术措施,对收集的数据进行处理,以确保数据集的完整性和真实性,防止数据被篡改、删除或损毁。最后,研

发人员应定期对 AIGC 进行合规审查,及时识别和解决潜在的数据风险。一种可行的方法是建立数据风险评估框架,通过对数据源、数据处理和数据存储等环节进行综合评估,识别和量化数据安全风险,并采取相应的措施进行管控。同时,为加强对数据风险的控制,可要求 AIGC 服务提供者完善数据安全管理制度,包括数据安全策略、数据权限控制、数据加密与传输等内容。

另一方面,在处理“主动爬取”获取的数据时,应构建网络攻击监测与管理系统,防止因不良披露与聚合分析,将重要数据非法纳入 AIGC 的语料库。如美国国家安全局建立的网络监测系统 ICREACH,其能利用先进的网络安全技术和人工智能算法实时监测和分析整个互联网的流量,在大规模网络流量中快速准确地识别出可疑活动和攻击行为。而在构建网络攻击监测与管理系统时,应至少包括以下方面内容:一是完善数据分类分级体系。分类分级是数据保护的前置性要件,我国《个保法》第 13~43 条与第 44~50 条的规定也是采用“一般个人信息”与“敏感个人信息”分类保护的方式。因此,应该根据数据的敏感程度和重要性将数据分为不同级别,并明确各级别数据的保护要求和安全控制措施。例如,个人隐私数据和商业机密数据应该被划分为高级别数据,需要进行更加严格的保护措施。二是采用先进的技术手段对网络流量进行监测和分析。目前,AI 和机器学习等技术已经在网络安全领域取得重大进展,因此,通过利用这些技术来构建自动化的网络监测系统能够处理大规模的网络数据,提供更全面、准确的分析结果。相比传统的手动分析方法,自动化系统可以快速扫描和处理大量数据,有效减少人为分析的工作量,提高分析的效率和准确性。三是建立监测与管理系统与数字平台的联动机制。数字平台在数据流动中起到了重要作用,应与监测与管理系统进行紧密的协调和合作,主动配合监测与管理系统的工作,提供必要的数据支持和协助,共同保障数据的安全。四是加强国际合作与信息共享。网络攻击往往具有跨国性和跨边界性的特点,需要国际社会共同努力来应对。各国应加强合作,分享网络攻击情报和安全技术,共同建立起一个有力的网络安全防线。同时,相关国际组织和标准化机构也应在数据流动治理方面发挥积极作用,推动全球数据的安全流动,GDPR 就是一个成功的借鉴范例。

2. 强化算法解释权,构建数据安全保障体系

敏捷治理的“多元性”强调协同合作,整合多方力量,形成多元主体共治的合作机制,而不是单纯依赖政府的强制性措施。AIGC 数据安全治理应充分发挥政府、企业、公众等主体的独特优势与资源,并考虑各方主体权益,构建数据安全保障体系。

一方面,对于“数据偏见”风险,应强化算法解释权,以保障 AIGC 的可信度与公正性。在数字时代,算法的决策过程和结果对个人、组织和社会都会产生重要影响,但由于其具有复杂性和非直觉性特点,导致决策透明度受到一定程度的限制。对此,强化算法解释权的目的在于促使内部机制更加透明和可理解,用户和利益相关方能够知晓算法的运作方式,以及对决策结果产生的影响,进而建立起一个可信赖的数据“训练系统”。算法解释权使得相对人能够知晓算法决策的依据与过程,弥合平台服务提供者与相对人之间的地位鸿沟,推动形式化的数字服务协议转向为实质上的平等[26]。欧盟和美国都较早将算法解释权“法定化”:欧盟在 GDPR 第 22 条中规定,数据主体有权对数据控制者进行干预,并对后者作出的决策可提出异议;美国在《算法透明性和可问责性声明》提出了算法运行的七项原则,其中一条便涉及算法解释权。而我国虽在《生成式人工智能服务管理暂行办法》第十九条首次明确规定了算法解释权①,尽管这一条款在处理算法风险方面取得重要进展,但仍然存在一些不足:该规定仅涵盖监管机构等特定主体对算法解释的要求,相关方权益无法得到有效保障;该规定对于

①《生成式人工智能服务管理暂行办法》第十九条规定:有关主管部门依据职责对生成式人工智能服务开展监督检查,提供者应当依法予以配合,按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明,并提供必要的技术、数据等支持和协助。

描述算法的机制和原理过于笼统,缺乏明确的指引,难以应用于实践。

鉴于此,有必要进一步基于“多元性”来强化算法解释权:首先,应扩大算法解释权的主体范围。除监管机构,其他相关方(如 AIGC 开发者、数据提供方、使用者等)也应该涵盖于算法解释权的范围之内,如此,相关方将有权要求算法服务提供者对其进行解释和说明,以更好地理解算法的运行原理和决策依据,提高算法治理的透明度和公正性。其次,需明确算法解释权的具体内容和标准。算法解释权应该对包括算法的输入、输出、决策逻辑以及对敏感数据的处理方式等方面说明[27]。例如,在使用 AIGC 进行贷款信用评估时,用户应有权要求解释算法是如何根据其信用历史、收入水平和其他相关因素来进行评估。对此,可借鉴美国《公平信用报告法》(fair credit act reporting act, FCRA)与《中华人民共和国商业银行法》第四十四条关于“解释”的规定①。最后,需加强算法监管和评估机制。监管机构应承担对算法解释的监督和执行的责任,确保相关规定的贯彻落实,要求算法服务提供商提交解释和说明的文件,对其进行审查和评估。同时,应建立独立的评估机构或专家团队,负责对算法解释的准确性和合理性进行评估和验证,以确保解释的可靠性和真实性。

另一方面,在处理“数据泄露”风险时,应从事前数据保护与事后应对处理两方面着手,即“数据风险控制”与“数据风险应急”,以构建全方位的数据安全管理体系。就“数据风险控制”而言,需建立全面的数据处理风险预防机制,使 AIGC 的语料库数据时刻处于安全可控的状态。首先,数据分级保护是风险控制的重要前提。根据国家互联网信息办公室发布的《网络数据安全管理条例(征求意见稿)》第九条规定,我国已将数据分为国家核心数据、重要数据和一般数据,并对其进行差异化保护。在数据分级保护中,不同等级的数据应拥有不同的保护措施[28]:对于国家核心数据应采取最高级别的保护措施,特殊权限应该授予有资质的管理员;对于核心数据库可采用物理隔离和访问控制等措施;对于重要数据和敏感数据,可采取加密访问等措施;对于一般数据则可以采用较为简单的保护措施。其次,数据风险监测和评估是风险控制的核心要素。根据《数安法》第 29、30 条规定,AIGC 服务提供者应采取有效手段实施数据风险监测,一旦发现数据安全漏洞或其他风险,应立即采取相应的补救措施,以最大限度降低潜在的数据风险。同时,AIGC 服务提供者还应定期进行数据风险评估,评估范围应涵盖数据获取、存储、传输、处理等环节,风险评估结果应以报告形式提交给相关主管部门,使监管机构能够对数据处理活动的合规性和安全性进行审核和监督。最后,设立数据安全负责人和管理机构是风险控制的重要措施。根据欧盟 GDPR 第 37、38 条规定,数据处理活动必须设置“数据保护官”,负责数据保护方面的工作,我国《数安法》第二十七条对此进行了类似规定。因此,AIGC 服务提供者也应当设立数据安全负责人和管理机构,以履行数据安全保护义务,而具体义务内容可参照《网络数据安全管理条例(征求意见稿)》第二十八条规定。

尽管“数据风险控制”能从整体上提高 AIGC 的数据安全保护能力,但仍无法完全避免数据风险事件的发生,应辅之“数据风险应急”,以应对突发性数据风险,保障用户数据安全。首先,AIGC 服务提供者应明确数据安全责任人的职责。责任人不仅须具备兼备技术和管理能力,能够负责应对潜在的数据安全威胁,还需成立应急响应团队,并按照预设的应急预案履行职责,及时采取措施防范和处理数据安全事件。其次,应急预案既应包括对数据安全事件进行分类,也需制定相应的等级划分标准。划分等级可以依据数据的重要性、涉及的个人隐私及机密性等因素进行,以帮助应急响应团队更好地理解事件的紧急程度,并采取相应的措施。再次,AIGC 服务提供者需事先制定应急

① FCRA 第 607 条(b)款要求消费者信用报告机构必须采取合理的程序收集和公开消费者的有关信息,以最大可能保证报告内容的准确性,避免提供不准确的或过失的信息。《中华人民共和国商业银行法》第四十四条:商业银行开展贷款业务,应当按照法律法规和有关规定遵循风险可控、审慎经营的基本原则,诚实信用,合规经营。

响应措施。在制定措施过程中,应考虑 AIGC 的特殊性,确保应对措施与其数据风险相匹配,具体可包括暂停服务、隔离受影响的系统、采取紧急等措施。最后,针对不同类型的数据风险事件应制定相应的应急预案。由于不同类型的数据风险具有不同的特征和紧急程度,因此应制定区别对待的预案。如,针对数据泄露、未经授权访问、不良攻击等不同类型的事件制定相应的预案,并且预案应根据实际情况进行定期修订,以确保其适应性和有效性。

3. 完善应用生成机制,推进技术融通型法治

诚如前述,AIGC 生成的不良内容(包括虚假性、歧视性等信息)不仅会侵害个人权益,还会对国家、社会稳定造成巨大威胁。对此,应充分发挥敏捷治理的"灵活性"优势,综合运用技术与法律手段,以全方位预防不良内容生成风险。

(1)技术层面:完善生成机制,规范技术标准

一方面,AIGC 服务提供者应完善应用生成机制,以控制不良内容生成风险。首先,可利用自然语言处理和机器学习技术来检测和屏蔽不良内容。通过构建一个庞大的训练数据集,该数据集包括各种类型的不良内容,例如虚假言论、歧视性言论等,AIGC 在训练过程中可学习辨别这些不良内容的特征和模式,并将其作为生成内容时的参考。其次,可在训练阶段引入对不良内容的约束条件与标签。可以通过人工标注数据集,筛选和标记适合的文本作为监督数据,而后在训练过程中与无监督数据相混合,以提高模型对合适内容的生成概率,进而减少不良内容的生成。再者,采取用户反馈机制来实时监控和改进 AIGC 的生成内容。用户反馈可由人工审核或自动系统来处理,即通过分析用户对生成文本的评价,不断优化模型的生成内容。例如,当用户举报或投诉某一段文本时,可通过人工审核或自动检测系统对问题文本进行分析,并在系统中进行相应的调整与修正。

另一方面,应制定规范化的技术标准,为 AIGC 的发展提供指引,确保其生成内容符合公序良俗和法律要求。AIGC 的不良内容生成风险涉及多个维度,包括伦理、法律和社会等问题,因此,制定全球性的技术标准势在必行,能够为研究人员和从业者提供具体的指导和帮助:第一,国际组织应制定明确的技术标准,标准需涵盖保护用户隐私和个人数据的要求,确保算法的透明度和可追溯性,以防止 AIGC 被滥用或用于恶意目的等内容。例如,在国际标准化组织(ISO)、国际电工委员会(IEC)于 2023 年联合发布的 IEC 42001(《信息技术 - 人工智能 - 管理系统》)中,已将上述技术标准内容纳入其中,并进行了相应的细化①。第二,技术标准应当具备可操作性。可以通过明确的操作指南、示范案例等方式,为 AIGC 的研究与应用提供有效指导:操作指南应详细阐述 AIGC 需要遵循的步骤和原则,以确保技术标准得以落地;示范案例则需清晰地展示技术标准的实施方式,并帮助用户更好地理解与应用。第三,技术标准应定期进行更新和调整。GPT-3.5 在短短 6 个月的时间就迭代到 GPT4.0,AIGC 具有快速迭代升级的特点,而每一次升级都可能会带来新的风险与挑战,Sora 亦是例证之一。鉴于此,技术标准也需要随之进行定期"迭代",使得其始终与应用的实际需求保持一致,从而提高 AIGC 生成内容的质量与准确性。

(2)法律层面:构建发展与规制并行的监管框架

当前,各国对于数据的获取、使用及其监管呈现出显著差异,这些差异直接关系到 AIGC 的创新环境与市场潜力。美国在面临数据保护和隐私权的挑战时,治理策略更多依赖行业标准、企业自律与消费者保护相关法律;欧盟则采取推行 GDPR 的方式,强调数据主体的权利,严格限制数据跨境流动,并要求数据处理需基于合法、透明的原则,这对 AIGC 的数据获取设置了高标准门槛。相比之下,我国在数据治理上采取了更为灵活且具有导向性的策略,如

① ISO/IEC42001 的目标可以总结为:促进开发和使用可信赖、透明和负责任的人工智能系统。在部署人工智能系统时,强调公平、非歧视和尊重隐私等道德原则和价值观,以满足利益相关方的期望。帮助组织识别和缓解与人工智能实施相关的风险,从而降本增效。

《网安法》《数安法》及“数据二十条”的出台①,既保障国家安全与个人隐私,又鼓励数据的合理流动与开发利用,为 AIGC 提供了丰富的数据资源池。但是,我国在数据获取上的相对宽松并未完全转化为 AIGC 产业发展的绝对优势,原因在于数据质量、数据标注的标准化程度以及跨域数据整合能力等方面仍有待提升。此外,国内外对于数据权属、算法透明度的法律界定不一,影响着 AIGC 产业的国际合作与信任构建。如,欧盟对算法决策的可解释性要求较高,而我国虽然倡导算法公开透明,但在实际操作层面仍存在解释权与问责机制的模糊地带。我国虽在《互联网信息服务深度合成管理规定》第六条、第七条等规定中,为强化 AIGC 不良内容生成风险的管理与治理提供了适用规则;在《生成式人工智能服务管理暂行办法》第 4、7、19 条等规定中,对生成内容的准确性、合法性与“违规者责任”作出了清晰的界定。然而,现行法律规范仍存在局限之处:一方面,AIGC 的复杂性和隐蔽性使得监管机构难以适用上述规则,并及时发现和遏制不良内容的生成。由于 AIGC 能够在无需人类干预的情况下生成内容,监管机构往往在不良内容已广泛传播后才能够介入,同时其隐蔽性也使得追踪和确认责任主体变得更加困难。另一方面,现行法律对于 AIGC 生成的不良内容所采取的处罚措施较为单一,主要通过行政罚款、营业整顿等方式进行,对于恶意传播不良信息或侵犯个人隐私等严重违规行为,这些措施并不足以起到应有的惩戒和威慑作用。

有鉴于此,我国亟须推进技术融通型法治,构建发展与规制并行的监管框架②,这要求监管部门深入理解 AIGC 技术的原理、应用及其潜在风险,基于技术中立原则出发,通过制定更为精准和灵活的法律规定和政策措施加强对 AIGC 的监管与执法。具体来说:其一,明确责任主体。立法需明确 AIGC 内容生成的法律责任归属,既考虑技术提供者的责任,也不忽视使用者的责任。例如,可以借鉴欧盟《数字服务法案》对平台责任的规定,对于通过 AIGC 技术传播的内容,要求平台采取适当措施进行管理和监督,同时保障用户权利。其二,加强数据与隐私保护。对于 AIGC 涉及个人隐私的处理,需要进一步细化《个保法》中关于合理必要原则、知情同意规则等要求,并针对 AIGC 的特性进行具体化。其三,完善版权制度。AIGC 的“自主性”对版权制度产生了较大冲击,需对现行版权法进行修订,明确 AIGC 生成内容的版权归属,同时为创新和知识共享提供适当的余地,以避免过度限制技术的发展。其四,增设刑事责任。对于故意利用 AIGC 进行犯罪活动(如散布虚假信息、侵犯隐私等)的行为,应考虑在刑法中增设相关罪名和刑责,使惩处措施与违法行为的严重性相匹配,贯彻“罪责相适应”原则。其五,促进国际合作。随着 AIGC 不断发展,未来相关产品及其数据风险将呈现跨国性特征,因此应加强国际合作,参与国际数据治理规则制定,这是应对 AIGC 数据风险的关键举措。例如,欧盟议会在 2024 年通过的《人工智能法案》,此次法案的中心方法论是根据 AI 对社会造成危害的能力,遵循“基于风险”的方法来监管人工智能——“风险越高,规则越严”,这亦与敏捷治理的意涵相契合,可为我国构建本土化的立法与监管手段提供法治借鉴。

四、结　语

习近平总书记提出的新质生产力,既是我国新发展阶段经济增长的引导力与创新推动力,更是塑造经济新增长点及打造国际新竞争优势的战略方向。AIGC 与新质生产力具有深度耦合与相互塑造之关联,前者对后者的赋能效应已成为理论与实践探索的核心议题。技术本身既具有积极的作用,也可能带来负面影响,

①《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》(即“数据二十条”)提出的工作原则:遵循发展规律,创新制度安排。充分认识和把握数据产权、流通、交易、使用、分配、治理、安全等基本规律,探索有利于数据安全保护、有效利用、合规流通的产权制度和市场体系,完善数据要素市场体制机制,在实践中完善,在探索中发展,促进形成与数字生产力相适应的新型生产关系,等等。

② 技术融通型法治是有机融合高新科技并满足复杂多变社会条件下所需快速反应能力与真相鉴证能力、法律适用能力的理念与机制,其有可能带来一种增量效应,带来一种超越于简单算数加法的倍数或指数效应,有可能使我国在法治化治理的某些方面走在前列。

AIGC 虽以算法为引擎,但只有在高质量、大规模的数据中进行学习和优化,大语言模型才能不断迭代升级,展现其卓越性能。然而,传统治理模式在面对 AIGC 数据风险时显得捉襟见肘。敏捷治理作为一种全新理念,正在引领数据安全治理进程,应充分发挥其适应性、多元性、灵活性等优势,构筑 AIGC 数据风险治理的坚实防线。未来应进一步结合《人工智能全球治理上海宣言》所提出的"以高水平数据安全保障高质量数据发展,推动数据的依法有序自由流动,反对歧视性、排他性的数据训练,合作打造高质量数据集,为人工智能发展注入更多养料",在数据治理领域加以持续不断地探索,关注如何将数据伦理与法律治理有机结合,以制定出具有中国特色的数据治理方案,赋能新质生产力发展,顺应乃至引领全球新一轮数字经济的高质量发展。

参考文献:

[1] 商建刚. 生成式人工智能风险治理元规则研究[J]. 东方法学,2023,(3):4-17.

[2] 陈兵. 生成式人工智能可信发展的法治基础[J]. 上海政法学院学报(法治论丛),2023,38(4):13-27.

[3] 邹开亮,刘祖兵. 论类 ChatGPT 通用人工智能治理——基于算法安全审查视角[J]. 河海大学学报(哲学社会科学版),2023,25(6):46-59.

[4] 支振锋. 生成式人工智能大模型的信息内容治理[J]. 政法论坛,2023,41(4):34-48.

[5] BHAVYA B, XIONG J, ZHAI C X. Analogy generation by prompting large language models: a case study of instructgpt[J]. arXiv preprint arXiv: 2210.04186,2022.

[6] WU T, HE S, LIU J, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023,10(5):1122-1136.

[7] ROUMELIOTIS K I, TSELIKAS N D. ChatGPT and open-AI models: a preliminary review[J]. Future Internet, 2023, 15(6):192.

[8] HE W, DAI Y, ZHENG Y, et al. Galaxy: a generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection[C]//Proceedings of the AAAI conference on artificial intelligence, 2022, 36(10):10749-10757.

[9] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30:681-694.

[10] CASPER S, DAVIES X, SHI C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback[J]. arXiv preprint arXiv:2307.15217,2023.

[11] 许中缘,郑煌杰. ChatGPT 类应用风险的治理误区及其修正——从"重构式规制"到"阶段性治理"[J]. 河南社会科学,2023,31(10):50-62.

[12] 张欣. 生成式人工智能的数据风险与治理路径[J]. 法律科学(西北政法大学学报),2023,41(5):42-54.

[13] WEI M, HARZEVILI N S, HUANG Y K, et al. Demystifying and detecting misuses of deep learning APIs[C]//Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, Lisbon, 2024:1-12.

[14] 陈景辉. 捍卫预防原则:科技风险的法律姿态[J]. 华东政法大学学报,2018,21(1):59-71.

[15] MERGEL I, GANAPATI S, WHITFORD A B. Agile: a new way of governing[J]. Public Administration Review, 2021,81(1):161-165.

[16] HOHL P, KLÜNDER J, VAN BENNEKUM A, et al. Back to the future: origins and directions of the "agile manifesto"—views of the originators[J]. Journal of Software Engineering Research and Development, 2018,6:1-27.

[17] QUMER A. Defining an integrated agile governance for large agile software development environments[C]//Agile Processes in Software Engineering and Extreme Programming: 8th International Conference, Springer Berlin Heidelberg, 2007:157-160.

[18] MERGEL I, GONG Y, BERTOT J. Agile government: systematic literature review and future research[J]. Government Information Quarterly, 2018,35(2):291-298.

[19] LUNA A J H O, KRUCHTEN P, de MOURA H P. Agile governance theory: conceptual development[J]. arXiv preprint arXiv:1505.06701,2015.

[20] 于文轩,刘丽红. 合供与敏捷治理[J]. 西北师大学报(社会科学版),2024,61(4):82-90.

[21] 韩兆柱,申帅杰. 敏捷治理:人工智能治理新模式[J]. 华东理工大学学报(社会科学版),2024,39(1):105-119.

[22] NYE Jr J S. The information revolution and American soft power[J]. Asia Pacific Review, 2002, 9(1):60-76.
[23] 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, 42(3):460-468.
[24] 中国新闻网. 习言道|共同促进全球人工智能健康有序安全发展[EB/OL](2024-07-15). http://www.chinanews.com.cn/gn/2023/10-20/10097341.shtml.
[25] 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[J]. 公共管理学报, 2022, 19(2):1-11.
[26] 张凌寒. 商业自动化决策的算法解释权研究[J]. 法律科学(西北政法大学学报), 2018, 36(3):65-74.
[27] 苏宇. 算法解释制度的体系化构建[J]. 东方法学, 2024(1):81-95.
[28] 洪延青. 国家安全视野中的数据分类分级保护[J]. 中国法律评论, 2021(5):71-78.

Data Risks and Agile Governance of AIGC Empowering New Quality Productivity /ZHENG Huangjie(Law School, Central South University, Changsha 410083, China)

Abstract: With the development of "AI+" action, generative AI applications (AIGC) represented by ChatGPT, Sora, ERNIE Bot, etc. will gradually realize commercialization and become an important quality of new quality productivity. However, while AIGC improves people's productivity and creativity, it also raises many data risks. Based on its technical principles, these risks can be divided into stages of illegal data acquisition risk, data security risk, and malicious content generation risk. Agile governance, as a new governance model, is a reflection and transcendence of traditional governance models such as exploratory, responsive, and centralized at that time. By embedding it into AIGC governance, it will have new connotations and characteristics in governance concepts, governance subjects, governance objects, governance tools, and other aspects. AIGC applies agile governance, which not only conforms to the trend of intelligent revolution and implements the requirements of technology for good, but also can overcome the difficulties of traditional governance. The guidance of national policy guidelines also creates a good legal environment for its application. Under the guidance of agile governance, the principle of risk prevention should be implemented and the data risk response mechanism should be strengthened. It is also necessary to strengthen the interpretation power of algorithms and build a data security guarantee system, improve the application generation mechanism, promote technology integration-based rule of law, effectively alleviate AIGC data risks, and empower the development of new quality productivity.

Key words: new quality productivity; ChatGPT; generative AI; agile governance; data risk