# 专家解读｜推动生成式人工智能精细化治理

当前，生成式人工智能正在加速发展，不断催生新场景、新业态、新模式和新市场，赋能千行百业。与此同时，其暴露出的安全风险也引发了各国监管部门的广泛关注，意大利、德国、法国、西班牙等多个国家的数据保护监管机构对ChatGPT表达担忧或发起调查。近日，国家网信办联合国家有关部门公布了《生成式人工智能服务管理暂行办法》（以下简称《办法》）。《办法》是我国促进生成式人工智能健康发展和规范应用的专门立法，界定了生成式人工智能技术的基本概念，规定了生成式人工智能服务提供者的制度要求，为生成式人工智能的健康发展指明了方向。

一、促进发展，鼓励生成式人工智能研发创新

新一代人工智能是推动科技跨越发展、产业优化升级、生产力整体跃升的驱动力量。习近平总书记高度重视人工智能技术发展，指出要加快新能源、人工智能、生物制造、绿色低碳、量子计算等前沿技术研发和应用推广。中共中央政治局4月28日召开会议强调，“要重视通用人工智能发展，营造创新生态，重视防范风险”。

《办法》立足于促进生成式人工智能健康发展和规范应用，在总则中明确了国家坚持发展和安全并重、促进创新和依法治理相结合的原则，并在技术发展与治理一章中提出了一系列鼓励措施。一是鼓励创新应用。生成式人工智能应用场景丰富，可以与各行业深度结合，助力实现各行业智能化变革。《办法》明确提出鼓励生成式人工智能技术在各行业、各领域的创新应用，有助于生成式人工智能赋能千行百业。二是鼓励多方协作。生成式人工智能在技术创新、数据资源建设、转化应用、风险防范等方面涉及多方主体，为此，《办法》鼓励行业组织、企业以及各个相关机构之间开展协作。三是鼓励自主创新与国际合作。生成式人工智能尚处于发展阶段，而我国在生成式人工智能相关算法、芯片等技术研发方面起步相对较晚。《办法》一方面鼓励生成式人工智能相关基础技术的自主创新，另一方面鼓励平等互利开展国际交流与合作，从而促进我国人工智能技术发展。四是鼓励资源共享。算力是支撑数字经济蓬勃发展的重要“底座”，数据是生成式人工智能的重要资源。为此，《办法》提出推动生成式人工智能基础设施和公共训练数据资源平台建设的措施，促进算力资源协同共享，提升算力资源利用效能。

二、划定底线，推动生成式人工智能向上向善

以ChatGPT为代表的生成式人工智能在社会变革中的具有重要作用，但其可能存在的数据违法收集、知识产权侵权、生成虚假信息等问题同样不容忽视。

《办法》积极回应生成式人工智能带来的社会问题，明确提供和使用生成式人工智能服务的法律底线。一是不得生成违法内容。《办法》落实《网络安全法》等法律的规定，强调提供和使用生成式人工智能服务应当坚持社会主义核心价值观，不得生成法律、行政法规禁止的内容。二是防止歧视。训练数据本身存在歧视内容、算法更倾向于某些特征值、标注人员的主观判断等因素可能导致歧视问题，给社会带来不良后果，如边缘化弱势人群或者煽动仇恨与暴力。为减少歧视现象发生，《办法》要求在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止歧视。三是尊重知识产权和他人合法权益。生成式人工智能在个人信息的处理过程中，可能存在违法收集、使用个人信息等风险；生成式人工智能的训练数据中如包括他人已发表并享有著作权的作品，也可能给知识产权造成侵害。为减少侵权事件的发生，《办法》强调尊重知识产权、商业道德；尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益。四是提高生成内容准确性和可靠性。自生成式人工智能面世之初，其存在的“一本正经地胡说八道”现象便引起了人们的警惕。这种虚假信息的产生很可能会误导用户，加剧社会对共享信息的不信任。如何保障生成内容的真实性，既是产业界为进一步扩大生成式人工智能商用范围需要克服的技术难题，也是监管部门需要重点考量的问题。《办法》要求生成式人工智能服务提供者（以下简称“提供者”）基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。这一要求充分体现了发展与安全并重的原则，符合生成式人工智能的技术原理与产业需要。

三、细化责任，构建生成式人工智能治理体系

生成式人工智能从研发到落地需要经历多个环节，包括训练数据处理、数据标注、提供服务等。清晰界定不同环节的要求更有利于降低生成式人工智能的安全风险，增强制度的可落地性。

《办法》明确提供者的义务与责任，提供者是利用生成式人工智能技术提供生成式人工智能服务的组织、个人，包括通过提供可编程接口等方式提供生成式人工智能服务的组织、个人。在训练数据处理环节，提供者应当使用具有合法来源的数据和基础模型，并采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。在数据标注环节，提供者应当制定标注规则，开展数据标注质量评估，并对标注人员进行培训。在提供服务环节，提供者应依法承担网络信息内容生产者责任，按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识；发现违法内容的，应当及时采取处置措施、进行整改并向有关主管部门报告；应依法承担个人信息处理者责任，对生成式人工智能服务使用者的输入信息和使用记录依法履行保护义务，及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求；应履行用户管理义务，明确并公开服务适用情况，指导使用者科学理性认识和依法使用技术，采取未成年人防沉迷措施，依法处置从事违法活动的使用者，建立健全投诉、举报机制；应提供安全、稳定、持续的服务。

四、强化披露，明确生成式人工智能监管手段

人工智能算法具有“黑箱”特性，表现为行为不可控、决策机制难以解释，给人工智能监管带来了一定困难。为此，《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等规范已经明确了算法备案、安全评估等监管手段，提高算法的透明度与可问责性。

《办法》与现有规范一脉相承，延续了此前监管手段，同时明确了分类分级监管的原则，健全了我国人工智能治理体系。一是明确了安全评估与算法备案要求。为有效应对算法“黑箱”问题，《办法》对提供具有舆论属性或者社会动员能力的生成式人工智能服务的，提出了安全评估与算法备案要求。二是明确了信息披露要求。《办法》明确了提供者应配合主管部门的监督检查，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。

《办法》是在区块链、汽车数据、算法推荐、深度合成等立法之后，我国对新兴领域立法的又一次探索。《办法》充分协调了发展与安全之间的关系，一方面提出了生成式人工智能发展的鼓励措施，另一方面明确了生成式人工智能的法律底线，构建了生成式人工智能精细化治理体系，有利于推动生成式人工智能健康有序发展，为世界贡献人工智能治理的中国智慧和中国方案。（作者：辛勇飞 中国信息通信研究院政策与经济研究所所长）

中国网信网2023-07-13