DOI:10.12154/j.qbzlgz.2024.02.004

# 生成式人工智能数据安全风险及其应对*

赵梓羽 （浙江大学光华法学院 杭州 310088）

**摘 要：[目的/意义]**生成式人工智能的广泛应用催生了多重数据安全风险，传统的回应型治理和集中式治理模式显然难以应对新挑战，而敏捷治理则以其灵活的姿态彰显了独特的优越性。**[方法/过程]**首先，梳理生成式人工智能在数据安全领域引发的多重风险，进而对三类治理模式在风险管理上的成效进行对比分析，在此基础上提出我国应当尽快实现治理模式转变。然后以此为指导，构建具体的治理制度。**[结果/结论]**敏捷治理模式凭借其适应性、柔韧性和包容性特质，为应对生成式人工智能等新兴技术的数据安全风险提供了有效的方案，能够在实践中不断优化生成式人工智能数据安全治理体系。在敏捷治理模式下，应当树立“预防与应对并重”的适应性治理理念，构建“多元参与，合作互动”的韧性治理机制，运用“技术叠加法律”的包容性治理工具，从而形成综合的生成式人工智能数据安全治理体系。

**关键词：**生成式人工智能 ChatGPT 数据安全 敏捷治理

**Data Security Risks and Countermeasures of Generative Artificial Intelligence**

Zhao Ziyu (Guanghua Law School of Zhejiang University, Hangzhou, 310088)

**Abstract: [Purpose/significance]** The widespread use of generative artificial intelligence has given rise to multiple data security risks. Traditional reactive and centralized governance models seem inadequate to address these new challenges. In contrast, agile governance, with its flexible approach, demonstrates a unique superiority. **[Method/process]** Firstly, a meticulous examination of the multifaceted risks in data security posed by generative artificial intelligence is conducted. Subsequently, a comparative analysis of the efficacy of three governance models in risk management is carried out. It is suggested that China should promptly transition its governance model. Guided by this, a specific governance framework is constructed. **[Result/conclusion]** The agile governance approach, characterized by its adaptability, flexibility, and inclusiveness, offers a highly effective strategy for addressing data security risks associated with emerging technologies like generative artificial intelligence. It allows for the continuous optimization of the governance system for generative artificial intelligence data security. Within the agile governance framework, it is essential to embrace the adaptive governance concept of "equal emphasis on prevention and response", construct a resilient governance mechanism that promotes "diverse participation and collaborative interactions", and employ inclusive governance approaches that combine technology with legal measures. This way, a comprehensive framework for ensuring the security of generative artificial intelligence data is constructed.

**Keywords:** generative artificial intelligence ChatGPT data security agile governance

*本文系国家社会科学基金一般项目“数字经济时代个人信息侵权损害赔偿责任研究”（批准号：22BFX079）的研究成果之一。

## 1 引言

2022年11月，美国人工智能研究实验室OpenAI正式推出了人工智能聊天机器人程序ChatGPT（Chat Generative Pre-trained Transformer）。作为生成式人工智能的最新成果，ChatGPT依赖于其强大的自然语言模型，实现了与用户进行深度交流的突破，在完成艺术创作、文本校对和编辑、代码修改等多样任务中展现出惊人表现，真正颠覆了以往人机交互体验中的机械性和被动性局面，昭示着生成式人工智能领域的新巅峰[1]。然而，卓越成就带来喜悦与机遇之余，我们亦需审慎面对新的挑战与风险。随着ChatGPT的广泛应用，技术带来的数据安全、隐私保护、知识产权、算法偏见等问题成为备受关注的焦点，尤其是数据安全问题几乎触及数字时代的核心，构成不可忽视的国家安全隐患。

ChatGPT采用深度学习和Transformer架构，借助自注意力机制高效地处理序列数据[2]。它通过深入学习大规模文本数据来积累语言知识，然后将这一模型用于执行特定任务，例如生成式对话。在这一过程中，数据始终担任着关键的基础角色，为模型的性能和应用质量奠定坚实基础。但正因如此，数据非法获取、数据泄漏和数据偏见等潜在的数据安全风险随之产生。这些风险威胁着个人隐私等合法权益，甚至可能危及国家安全。目前，许多科技巨头已投入生成式人工智能模型的研发，百度的文心一言、谷歌的Bard、阿里的通义千问等多款类ChatGPT生成式人工智能模型产品正竞相待发。可以预见，随着生成式人工智能技术应用范围的不断扩张，潜在的数据安全风险也将不断加剧，从而给国家安全带来更大的威胁和隐患。

出于数据安全保护，意大利个人数据保护局宣布从3月31日起禁止使用ChatGPT，限制OpenAI处理意大利用户信息数据，并启动了针对其隐私安全问题的调查。德国、法国、爱尔兰等国家也开始准备效仿意大利的做法，加强对ChatGPT的监管。在我国，国家网信办联合国家有关部门制定的《生成式人工智能服务管理暂行办法》（以下简称《办法》）已经于2023年8月15日正式实施。全国信息安全标准化技术委员会也已制定技术文件《生成式人工智能服务安全基本要求》，并于2023年10月11日起公开征求意见。这些法规文件规定了语料库数据来源、语料库数据标注、生成内容安全评估等内容，对于及时应对和治理生成式人工智能带来的数据安全风险和冲击迈出了关键一步。学界也针对生成式人工智能应用中的数据安全问题展开了深入研究，并提出了构建以数据解释机制为核心的生成式人工智能数据治理框架[3]，构建语料库数据类型化基础上的数据安全分类治理机制[4]，融合制度和技术推进AIGC数据算法可信治理[5]等一系列应对措施。然而，大多研究未能结合生成式人工智能技术特征以及数据安全治理的时代背景进行治理模式的反思与优化，仅仅停留在提出数据安全问题并进行修补的层面，这未免限制了现有治理方案在实际效益方面的发挥。

鉴于此，本文将深入剖析生成式人工智能运行全过程中的主要数据安全风险，涵盖从预训练、运算到生成的所有环节。结合生成式人工智能技术特征、数据安全治理实践以及整体政策环境，综合评估现有的三种数据安全治理模式。在此基础上，提出我国应当尽快实现从回应型和集中式到敏捷治理的模式转变，并构建具体的治理路径，为生成式人工智能数据安全问题的研究与分析提供更为深入、更全面的指导方案。

## 2 生成式人工智能数据安全风险

生成式人工智能之所以具备卓越的生成能力，得益于其出色的自然语言模型和基于大规模数据的训练，然而这种优势也隐含着一系列数据安全风险，诸如数据非法获取、数据泄漏以及数据偏见等。这些风险纵贯生成式人工智能应用的全过程，直接或潜在地威胁着国家利益。

### 2.1 语料库数据非法获取风险

生成式人工智能依赖于足够多且多样化的数据进行预训练，从而建立更全面、准确的语言模型。为了使模型能从大量的文本数据中学习到更广泛的语言知识和进行上下文理解，预训练数据的规模通常可以达到数十亿甚至上百亿个参数量[6]。例如，GPT-3中的语言处理引擎使用了超过1750亿个参数和变量，最新版本GPT-4的参数量甚至达到了GPT-3的500倍，约为100万亿[7]。大规模的训练数据需求量驱使ChatGPT不间断的进行数据获取，难免会导致“未经授权”和“超越授权”获取数据的情形。

首先，在ChatGPT的语料库数据中，若包含信息主体个人数据的汇集，必须遵守《个人信息保护法》第13条规定，确保在事先获得信息主体同意的前提下进行数据收集。然而，由于ChatGPT的语料库数据规模庞

大，逐一获得信息主体的同意变得极为繁琐，甚至不现实，这为数据获取的合法性带来了挑战。其次，即使ChatGPT语料库中获取的个人信息来自已公开数据，技术服务提供者可援引《个人信息保护法》第27条“在合理的范围内处理个人自行公开或者其他已经合法公开的个人信息”来证实数据获取行为的合法性，但对于ChatGPT获取个人信息作为数据源是否属于“合理的范围内”仍存在不确定性。如果技术服务提供者的数据获取行为超出最小必要范围或者对个人权益产生重大影响，仍然可能被认定为非法[8]。再次，根据OpenAI官方网站“我们的方法”页面，ChatGPT的大量训练数据还源于图书数字化和学术文献的收集，而这样的数据收集行为受到著作权限制。由于文本挖掘的主要目的是用于商业用途，通常难以满足合理使用的条件，在未获得作者明确授权的情况下，此类数据收集行为容易被认定为侵犯著作权，从而构成数据的非法获取。最后，在所有数据获取方式中，OpenAI使用最为广泛的数据获取方式是数据爬虫[9]。根据我国《数据安全法》第32条的规定，ChatGPT等生成式人工智能系统爬取我国境内数据时，必须严格遵守我国数据保护相关规定。爬取我国未公开数据、出于非法目的爬取我国数据或者采取其他恶意手段等行为都明显超出正当性边界，构成对语料库数据的非法获取。

可见，生成式人工智能预训练阶段存在多方面的数据非法获取风险，若不加以控制和规范，这些风险可能导致多种严重后果，包括侵害个人信息权益、个人隐私权、著作权等，甚至涉及我国的数据主权问题。数字化时代，数据被认为是现代社会的核心资产，对国家的安全、经济发展和社会稳定至关重要，尤其是在全球数字化竞争日趋激烈的背景下，数据主权的争夺也变得愈加复杂和紧迫。以棱镜门事件为教训，若任由生成式人工智能技术服务提供者无视数据主权，通过非法手段获取大量敏感数据并越境使用，将可能对我国的数据主权构成威胁，损害国家核心利益。

**2.2** 数据泄漏风险

在生成式人工智能应用的全过程中，从预训练到模型运算再到输出，始终伴随着数据泄漏的潜在风险。自ChatGPT发布以来，科技巨头纷纷告诫员工不要在该模型上上传敏感信息，以防止数据泄露事件的发生。具体而言，生成式人工智能应用中的数据泄漏风险主要可归纳为以下两类。首先，迭代训练导致的数据泄漏风险。尽管OpenAI声称严格遵守隐私和安全政策，不会使用用户提交的数据来训练或改进模型，但从其运行原理和过程来看，始终存在着一种难以克服的、广泛存在的数据泄漏风险，即迭代训练导致的数据泄漏风险。在生成式人工智能模型训练过程中，它需要持续地通过大规模的数据集学习语言知识和上下文理解能力来不断升级和优化自身。这些模型参数所包含的学习成果，可能涵盖了用户与ChatGPT交流过程中输入的敏感信息、个人数据和商业机密等内容。尽管发布模型时可能采取参数裁剪或脱敏处理等措施，但部分数据信息可能仍会残留于模型中，并在模型运行时通过文本输出再次呈现，造成数据泄漏。其次，网络攻击和漏洞导致的数据泄漏。生成式人工智能系统可能会受到网络攻击或自身存在漏洞，黑客或恶意攻击者可能利用这些漏洞获取系统中的数据，导致数据泄漏[10]。近期发生的数据泄漏事件也突显了这一问题。据OpenAI于2023年3月24日发布的公告显示，ChatGPT Plus中1.2%用户的数据可能被泄露，一些用户称已经看到其他人的聊天记录片段，以及其他用户的信用卡的最后四位数字、到期日期、姓名、电子邮件地址和付款地址等信息。该事件被确认是由于内部开源数据库错误所致。可见，尽管随着数字技术的发展，数据安全保密技术逐渐成熟，然而在面对恶意攻击时，仍难以完全抵御，生成式人工智能仍存在着不可忽视的数据安全风险。

ChatGPT作为拥有庞大数据体量的生成式人工智能模型，一旦发生数据泄露事件，将带来难以挽回的灾难性后果。首先，个人隐私和数据安全将受到严重威胁。ChatGPT涵盖用户与其进行的聊天记录，其中包含大量敏感信息，如个人身份、偏好、习惯等。这些数据若落入不法之手，将导致个人隐私遭到侵犯、身份盗用、金融欺诈等风险，对用户造成直接损害。其次，企业和组织的商业机密将会面临泄露风险。ChatGPT在预训练阶段可能涉及商业数据，如企业策略、市场计划、产品规划等信息。这些机密一旦曝光，将引发不正当竞争、知识产权侵权等问题，对企业造成巨大损失，甚至威胁其生存发展。最后，若ChatGPT中包含我国重要基础战略性信息，例如能源、国防等方面的数据，泄漏后将对国家安全造成较大影响。通过数据的相互关联，生成式人工智能技术服务提供者可能推测我国的军事意图、政策动向以及经济计划，从而使我国陷入

国际竞争的被动局面，甚至遭受外部干涉和制裁。尤其是近期OpenAI已宣布恢复ChatGPT的互联网访问功能，ChatGPT语料库将不再仅限于2021年9月之前的数据。这意味着生成式人工智能语料库数据基数将进一步增大，相应地，由数据泄露可能造成的损害范围也将变得更加广。

**2.3** 训练数据偏见风险

生成式人工智能文本生成的结果受其自然语言模型的影响，这一模型本质上依靠算法选择和用于训练的大规模数据库。然而，这种依赖性也带来了潜在的训练数据偏见风险。

研究发现，最初版本的ChatGPT在面对政治敏感话题时，会毫不避讳地表达自己的立场，表现出明显的左倾自由主义观点，这种现象引发了对生成式人工智能模型偏见和价值观影响的担忧。然而，随着技术的改进，2023年1月版的ChatGPT在处理类似问题时，似乎更加客观中立，试图展现正反两方的观点。这一演进引发了对ChatGPT及类似技术认识和应用的深入思考。与其他语言模型一样，ChatGPT并非完全无偏见，其生成的文本结果受到设计者和训练数据的影响。训练数据的来源、数量和质量，设计者对反馈和立场的处理方式，都可能在一定程度上影响模型的倾向性和偏见[11]。一方面，ChatGPT的数据源于互联网上的大量文本，但这些文本并不一定代表真实观点和经验，从而使得其输出常常偏离全局客观事实。另一方面，ChatGPT采用“人类反馈强化学习”（RLHF）方法，通过将人类反馈引入，模型会逐渐调整自身参数，使得其生成的文本更加符合人类的价值观和意图[5]。然而，这也带来了一个问题，即反馈提供者的个人观点可能会被模型学习和加强，导致最终的文本生成结果缺乏客观性和全面性。这就存在令人担忧的潜在风险：一些生成式人工智能技术服务提供者可能出于特定的政治利益考虑，故意选择偏见化的数据构建语料库和进行模型训练，以传播特定的政治立场或观点[12]。其目的在于试图操纵用户观念，影响公众舆论，从而实现特定的政治目的。

数字时代，各种意识形态观点和价值观在网络空间中流动，人工智能技术将其迅速传播和放大。这种飞速的传播虽然带来了广泛的知识共享和文化交流，但同时也可能造成信息混乱、虚假宣传以及意识形态之间的激烈碰撞。可以说，当今全球多元文化与价值观的交织使我们面对前所未有的意识形态挑战。对于这些新兴技术，如果不加以适当的管理和监督，或会成为操纵和破坏意识形态安全的潜在威胁。

## 3 生成式人工智能数据安全治理模式

治理模式决定了数据安全治理的全局范畴和策略，是进行政策制定和监管执行的先决性要件。纵观各国数据安全立法及实践，历经三种数据安全治理模式的演进，包括回应型治理、集中式治理以及敏捷治理。对这些治理模式逐一分析，再综合治理成效、人工智能技术需求和经济发展等多重因素，对我国当前的生成式人工智能数据安全治理模式进行审慎反思和不断优化，是引导科技向善、化解风险的关键前提。

**3.1** 数据安全治理的传统模式及其缺陷

**3.1.1** 回应型治理：协同薄弱、前瞻性不足

在数据安全治理早期，随着数字经济的蓬勃发展和各类数字技术的迅猛涌现，数据流通成为推动经济增长的重要驱动力。在这一阶段，数据安全治理的首要目标是保障数据的合法流通与共享，为数字产业的良性发展提供坚实保障。为此，数据安全治理规范主要以引导性原则为主，鼓励企业主动承担安全责任，并辅之以适度的政府规范，目的是构建一种包容、互信、开放的数字环境，以推动数据资源的高效利用和数字经济的持续繁荣。

这一时期，各个国家普遍颁布了许多引导性法案和准则，旨在引导企业自觉强化数据安全防护，建立健全数据安全管理体系和自主把控风险，以回应潜在的数据安全风险。这种治理模式主要采取事后回应的方式，总体上呈现出治理范围较窄、覆盖事项较少、措施相对薄弱等特点[13]。与此同时，回应型治理模式还强调多元主体共治。政府在其中发挥引导和监督作用，督促企业强化安全措施，而企业则主动参与制定行业标准和规范，形成了政府与企业积极互动的合作格局，共同推动数据安全的自治和规范发展。这种合作模式为建立全方位的数据安全体系奠定了基础，也增强了整个数字经济生态系统的安全性和稳定性。

可以看出，尽管回应型治理模式在初期能够满足部分需求，但也存在明显的缺陷。一方面，回应型治理模式在处理数据安全问题时往往是零落而分散的，缺乏整体性的规划和统筹，容易导致数据安全措施的不协调和不衔接问题。这种局限性影响了数据安全治理的整体效能，使得数据安全的防范和响应难以形成有

力的防线。另一方面，该模式更多地依赖于数据安全风险的发生，缺乏主动性和前瞻性，可能在事态严重发展之后才采取措施，影响治理效果。特别是对于ChatGPT这样具有巨大数据规模和潜在影响力的大型语言模型，一旦发生数据安全事故将造成难以挽回的损害后果。这种专注于事后回应的治理模式，显然对于保障生成式人工智能数据安全和防范风险来说显得有些力不从心。

**3.1.2** 集中式治理：强监管、制约灵活应对

随着数字经济的不断发展和数字化程度的提高，回应型治理模式在应对新的数据安全挑战时局限不断凸显，风险应对不集中、不及时等问题呼吁新的治理模式。集中式治理模式以治理结果为中心，强调等级划分和权力集中、前瞻性防范及行业整体监管[14]，在一定程度上能够满足政府在加强数据安全保护、降低数据安全风险方面的需求，也体现了从追求技术高速发展到兼顾技术高质量发展的理念转变。

总体而言，我国当前对人工智能的治理即处于集中式治理阶段。从法律规范层面来看，我国先后颁布了一系列涉及人工智能和数据安全的法律法规，如《数据安全法》《网络安全法》《个人信息保护法》等。这些法律法规明确了对数据和人工智能安全管理和使用的要求，体现了政府对人工智能和数据安全的集中式管理和规范。从行政监管层面来看，监管部门对于人工智能等新兴技术产业采取了较为严格的监管态度。例如，对于互联网企业滴滴和阿里巴巴违反数据安全和资本垄断的行为，政府采取了强有力的监管措施，进行了数据安全审查和处罚，以确保企业合规运营和数据安全。这种强监管态度体现了政府在集中式治理模式下对于人工智能产业的主导地位和监督作用。

集中式治理模式的优势在于将权力集中，能够强化监管机构对于新兴技术产业的监督和管理，从而实现对数据安全问题强有力的整治，有助于规范行业发展和维护公众利益。但该治理模式缺乏必要的监管弹性，在应对快速变化的科技发展和数据安全挑战时显得有些僵化和不适应。一方面，数据安全威胁和攻击手段不断演进和变化，集中式治理模式下的政府或中心机构的反应速度和决策过程可能相对较慢，导致无法及时适应快速变化的威胁，难以保障数据安全。另一方面，由于集中式治理模式的刚性，政府强制实施过多限制性措施，可能对ChatGPT等人工智能技术的创新带来限制，影响行业的进步和发展。

**3.2** 敏捷治理：生成式人工智能数据安全治理的新思路

在应对生成式人工智能应用中复杂多变的数据安全挑战和治理难题时，传统的回应型和集中式治理模式已显现出一定的局限性和不足。为此，亟需推动数据安全治理模式的转型，迈向更高层次的治理体系。敏捷治理模式是一种基于敏捷方法论的治理模式，其核心理念是在快速变化和不确定的环境中，通过持续适应和灵活性来有效管理和应对复杂问题[15]。“敏捷”一词最早源自20世纪80年代的制造业，强调在快速变化的市场环境下探索更加高效的生产模式，以提升企业的适应能力和生产效率。随着数字经济的蓬勃发展，这一概念已被引入到其他领域，包括软件开发、项目管理和数据安全治理等。在当前生成式人工智能数据安全风险治理方面，相对于传统的回应型治理和集中式治理模式，敏捷治理展现出以下核心特质和优势。

首先，敏捷治理模式具有适应性，强调风险的快速应对。2018年世界经济论坛上，“敏捷治理”被定义为“一套具有流动性、灵活性或适应性的行动或方法，是一种自适应、以人为本以及具有包容性和可持续性的决策过程”[16]。这一特性赋予了敏捷治理模式针对崭新的技术挑战做出灵活反应的能力。以ChatGPT为代表的生成式人工智能技术，其发展需要经历从产生到完善再到广泛应用的过程，其中充满不确定性、风险突发性的特点，并由外部向内部逐渐扩展[17]。因此，在对待这一类新兴技术时，强调“风险预防”要远胜于“惩罚威慑”，并需要快速实现从被动安全到主动安全的思维方式转变[18]。敏捷治理模式以其高度灵活和适应的特质，通过“主动学习”来增强数据安全与风险防控能力，即使没有经验可供参照，也能够根据不断演化的环境作出及时应变，能够有效弥合数字社会快速变迁与监管滞后之间的矛盾[19]。同时，敏捷治理模式下，管理者更注重预判和预防，而非等风险事件发生后再采取应对措施。这一积极预防的态度使得管理者能够更早地辨识潜在的风险点，并采取相应措施予以遏制。

其次，敏捷治理模式具有柔韧性，强调协同合作。生成式人工智能数据安全风险具有复杂性且来源广泛，需要多个领域和利益相关者的协同配合，否则难以实现风险的有效治理[20]。传统的数据安全治理模式通常以政府机构为核心，强调国家机构的主导地位和法

律法规的权威性，限制了个人、企业和行业组织等其他主体的积极参与，难以形成共赢的新秩序，同时也难以适应数据安全风险的迅速变化和不确定性。相比之下，敏捷治理模式具有高度柔韧性，打破了传统“政府主导、社会参与”的机制，建立了“多元协同、合作互动”的治理结构[21]，展现出独特优势。通过政府、企业、学术界和公众等各方共同参与和贡献智慧，能够发挥多元参与主体的能动优势，形成治理的合力，从而提升数据安全治理的整体效能。

最后，敏捷治理模式具有包容性，促进自由和创新。数字时代，人工智能带来了机会与风险并存的复杂局面。尽管生成式人工智能技术伴随一定的数据安全风险，但总体而言，它代表了智能革命的重大进展，给社会带来了极大福祉[22]。因此，需要审慎运用法律规范，避免过度干预而对其发展造成阻碍。敏捷治理模式具有高度包容性，不主张法律的过早介入，而是强调不断发掘和拓展新的治理工具和治理方法，以提升治理的可持续性。这对于平衡机会与风险、为技术发展创造广阔前景至关重要。在这一点上，我国政策语境中的“包容审慎监管”与敏捷治理不谋而合。为了应对人工智能带来的挑战，2019年国家新一代人工智能治理专业委员会发布了《新一代人工智能治理原则——发展负责任的人工智能》，其中将“包容共享、开放协作”视为人工智能技术开发和应用的重要原则。2023年8月15日起施行的《办法》第3条提出的治理原则也表明了我国坚持发展与安全并重，促进创新与依法治理相结合。这充分体现了我国对人工智能科技创新的包容和鼓励态度，强调在发展中进行治理，在治理中实现持续发展，与敏捷治理的理念相一致。

## 4 敏捷治理模式下生成式人工智能数据安全风险的应对路径

敏捷治理模式以其灵活适应、协同韧性、包容持续等特质为应对生成式人工智能数据安全风险提供了一个强有力的框架。根据敏捷治理模式的三个核心特征，可从理念、机制和工具三个维度重塑生成式人工智能数据安全风险治理的全过程，形成一个系统化、综合性的应对机制。具体而言，基于敏捷治理模式的适应性，树立“预防与应对并重”的适应性治理理念；基于敏捷治理模式的韧性，构建“多元参与、合作互动”的韧性治理机制；基于敏捷治理模式的可持续性，运用“技术叠加法律”的可持续性治理工具。

### 4.1 适应性治理理念：预防与应对并重

治理理念决定着治理的总体方向和目标，是应对生成式人工智能数据安全风险之基石。敏捷治理模式以适应性为核心特征和关键优势，为此，生成式人工智能数据安全治理应树立“预防与应对并重、快速响应”的理念，以适应性和灵活性促进风险的及时应对和持续改进。这一理念将成为数据安全治理的指导原则，引领着未来生成式人工智能数据安全的整体建设。

在风险尚未显现时，适应性治理理念要求全面开展风险预防措施，包括建立详尽的风险评估和分析机制，以深入了解生成式人工智能应用可能面临的数据安全威胁和风险。此过程中应当关注数据处理的各个环节，涵盖数据的收集、传输、存储和输出等方面，同时建立严格的安全管控机制，保障数据的机密性、完整性和可用性。此外，还需要明确各方在数据安全中的主体责任和权责分配，并积极推动行业自律和标准化。在具体战略制定过程中，可以从国际经验中吸取有益之处，特别是参考欧盟最新发布的《人工智能法案》。该法案将人工智能系统的风险分为不可接受风险、高风险、有限风险和轻微风险四个级别，并为每个级别规定了相应的法律框架和义务。鉴于生成式人工智能具有强大的数据处理和生成能力，可能在某些应用中被归类为高风险级别，适用最严格的安全规范。实际上，在我国《办法》第3条中，已初步规定了类似的“分类分级监管”的原则，但尚待进一步详细规定。未来可以借鉴欧盟《人工智能法案》的相关规定来完善我国的相关政策法规或国家标准，更有效地管理生成式人工智能等新兴技术的发展，平衡创新与风险之间的关系。

当风险事件发生时，适应性治理理念要求快速、协调和有效的应对措施。此时，建立应急响应机制成为关键，这一系统性的安全措施包括预案与流程、责任与角色、监测与检测、风险评估与等级划分、通知与报告、处理与恢复、信息共享与合作、媒体与公众关系、事后总结与改进、培训与演练、法律合规等多个方面[23]。通过建立完善的应急响应机制，可以有效控制数据安全事件的危害后果，将其影响范围最小化。这些措施共同构建了生成式人工智能数据安全治理的坚实体系，为风险的识别、预防和应对提供了有效工具和指导。

### 4.2 韧性治理机制：多元参与、合作互动

敏捷治理模式具有极强的韧性，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,而不单纯依赖政府的强制性措施。在生成式人工智能数据安全治理中,政府、企业、公众等各方都拥有独特的优势和资源,需要充分发挥这些优势和资源,并协同合作,以实现数据安全的全面保障。

首先,政府担负着法制建设和法规制定的使命,代表公众整体利益,因此应在治理中发挥主导作用,建立全程监管模式。国务院发布的《新一代人工智能发展规划》提出了建立人工智能监管体系的任务,强调双层监管结构,即设计问责和应用监督并重。《办法》延续了这一规定并特别强调了训练数据合规的具体制度。为此,在生成式人工智能数据处理的各个环节,应设立相应的监管措施和责任,确保数据安全得到全程掌控。此外,政府还应制定和推广相关标准,推动整个行业的数据安全水平提高。

其次,企业在生成式人工智能数据安全治理中既是风险的开启者,同时也拥有强大的数字技术能力,能够有效控制风险。政府应当鼓励这些机构投入更多资源用于生成式人工智能数据安全技术的研发与创新。通过技术的不断创新和升级,提升数据安全防护能力,更好地应对日益复杂的安全威胁。同时,政府可以为企业和研究机构提供技术研发的政策支持和激励措施,以鼓励它们投资于数据安全领域。

最后,公众的数据安全意识和保护能力也至关重要。政府和企业可以通过广泛宣传和教育来提高公众对数据安全风险的认识,包括提示个人信息主体避免在生成式人工智能中输入敏感信息,告知公众在生成式人工智能处理其个人信息时享有告知同意和删除权等个人信息权益,增强公众对数据安全的意识和保护能力。此外,政府和企业还应加强数据安全培训和教育,特别是对从事生成式人工智能相关工作的人员进行专业培训,提高他们对数据安全的认识和重视程度。技术人员的培训内容可以涵盖数据加密、隐私保护和安全审计等方面,以增强他们在数据安全方面的专业能力。

**4.3** 包容性治理工具:技术叠加法律

敏捷治理模式的包容性特征为生成式人工智能治理提供了一个全新的视角。在这种模式下,治理策略可以综合运用技术和法律手段,对于能够在基础模型层和专业模型层通过技术自身发展解决的问题,无需过早引入法律约束,从而为生成式人工智能发展留足试错空间。

在技术层面,数据加密是核心技术之一,它为生成式人工智能数据提供了多层次的安全保护。通过在数据的存储、传输和处理过程中引入对称或非对称加密,将明文信息转化为代码并仅向授权者公开解密密钥,可以有效地抵御未经授权的访问和数据泄露风险。此外,数据脱敏技术同样具有重要意义。对敏感数据进行脱敏处理,有效防止数据还原为原始信息,可以在预处理阶段降低敏感信息泄露的风险[24]。隐私保护则专注于个人敏感信息的保护,方法包括差分隐私、同态加密和安全多方计算等,不仅维护了用户隐私,还确保了模型在隐私受限环境中的训练和生成[25]。这些技术手段的综合应用为生成式人工智能数据的安全治理提供了多重防护。鉴于不断演变的安全威胁,持续关注并采用新的安全技术至关重要,同时也需要保持灵活性和适应性,确保生成式人工智能数据在整个生命周期中的安全性和合规性。

在法律层面,制度建设是生成式人工智能数据安全治理的重要支撑。在敏捷治理模式下,制度建设应具备柔性和适应性,以适应快速变化和不确定性的风险。首先,生成式人工智能语料库数据的分类分级保护机制是确保数据安全的重要支撑,应坚持统筹数据安全与数据发展的原则,根据数据的价值和对利益的影响,对语料库数据进行分类和定级,自上而下构建完善的数据分类分级制度,并明确不同类型和不同级别数据的处理规则和安全保护措施,确保数据得到恰当地管理和保护[26]。对于重要数据,应将其识别和标记,设置严格的访问权限和数据加密,以确保其不被未授权的访问所泄露或滥用;对于一般数据,可以采取较为宽松的流通规则,以促进数据共享和利用,推动生成式人工智能应用的发展[27]。其次,为保障生成式人工智能数据的合法性、可靠性和高质量,建立健全的训练数据来源合法性审核机制至关重要。通过对数据来源的合法性进行审查,可以确保生成式人工智能模型的训练数据符合《数据安全法》《个人信息保护法》等相关法律法规的要求,从源头避免数据安全问题的出现,保障模型的可靠性和稳健性。

通过建立技术和法律的融合治理,可以将敏捷治理的包容性理念贯彻到数据安全风险治理的实际操作中,实现审慎而开放的治理。这一体系将兼顾安全保障与发展,形成生成式人工智能数据安全治理的全方位保障和持续优化。

## 5 结语

生成式人工智能的涌现与广泛普及标志着人工智能领域的历史性飞跃，引领了技术创新与应用领域的革命浪潮，为生产生活、经济发展和社会进步带来了极大便利和推动力。然而，这一进步也伴随着层层叠加的数据安全风险。语料库数据的聚集与流动所带来的数据非法获取、数据泄漏、数据偏见等威胁正日益升级，直接威胁个人隐私、企业商业机密，甚至国家安全利益。为此，亟需采取切实有效的措施进行数据安全风险治理。敏捷治理模式以其适应性、柔韧性和包容性等特点构筑了生成式人工智能数据安全治理的坚实防线。通过树立适应性治理理念、构建韧性治理机制、运用包容性治理工具，实施全方位的数据安全保障，能够持续不断地优化治理效能。但从长远来看，在数据安全治理的道路上，不能止步于现状，应积极拥抱新思维、新技术，不断创新与完善治理模式。唯有如此，才能引导科技向善，实现生成式人工智能与数据安全的协同共进，创造更加繁荣、安全和可持续的未来。

### 参考文献


[1] VAN DIS E A M, BOLLEN J, ZUIDEMA W, et al. ChatGPT: five priorities for research[J]. Nature, 2023, 614(7947): 224-226.

[2] 陈永伟.超越ChatGPT：生成式AI的机遇、风险与挑战[J].山东大学学报(哲学社会科学版),2023(3):127-143.

[3] 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.

[4] 刘艳红.生成式人工智能的三大安全风险及法律规制——以ChatGPT为例[J].东方法学,2023(4):29-43.

[5] 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.

[6] 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.

[7] 郭春镇.生成式AI的融贯性法律治理——以生成式预训练模型(GPT)为例[J].现代法学,2023,45(3):88-107.

[8] 刘晓春.已公开个人信息保护和利用的规则建构[J].环球法律评论,2022,44(2):52-68.

[9] 张宇轩.人机对话中个人信息的“设计保护”——以ChatGPT模型为切入点[J].图书馆论坛,2023,43(8):77-87.

[10] SEBASTIAN G. Do ChatGPT and other AI chatbots pose a cybersecurity risk? An exploratory study[J]. International Journal of Security and Privacy in Pervasive Computing (IJSPPC), 2023, 15(1): 1-11.

[11] 王 少.ChatGPT介入思想政治教育的技术线路、安全风险及防范[J].深圳大学学报(人文社会科学版),2023,40(2):153-160.

[12] 王延川,赵 靖.生成式人工智能诱发意识形态风险的逻辑机理及其应对策略[J].河南师范大学学报(哲学社会科学版),2023,50(4):1-7.

[13] 范玉吉,张 潇.数据安全治理的模式变迁、选择与进路[J].电子政务,2022(4):114-124.

[14] 马伊里.集中控制与非集中控制:政府再造的一种制度安排[J].探索与争鸣,2005(5):17-19.

[15] 于文轩,魏 炜.数据的敏捷治理：价值重塑与框架构建[J].广西师范大学学报(哲学社会科学版),2022,58(5):37-49.

[16] 吴 磊,冷 玉,唐书清.数字化时代敏捷治理的学术图景：研究范式与实现路径[J].电子政务,2022(8):77-88.

[17] 赵精武.生成式人工智能应用风险治理的理论误区与路径转向[J].荆楚法学,2023(3):47-58.

[18] 刘金瑞.数据安全范式革新及其立法展开[J].环球法律评论,2021,43(1):5-21.

[19] 于文轩.ChatGPT与敏捷治理[J].学海,2023(2):52-57.

[20] 蔡士林,杨 磊.ChatGPT智能机器人应用的风险与协同治理研究[J].情报理论与实践,2023,46(5):14-22.

[21] 贾 开.数字治理的反思与改革研究：三重分离、计算性争论与治理融合创新[J].电子政务,2020(5): 40-48.

[22] 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),2023,29(3):102-114.

[23] 刘欣然,李柏松,常安琪,等.当前网络安全形势与应急响应[J].中国工程科学,2016,18(6):83-88.

[24] 王毛路,华 跃.数据脱敏在政府数据治理及开放服务中的应用[J].电子政务,2019(5):94-103.

[25] 丁红发,孟秋晴,王 祥,等.面向数据生命周期的政府数据开放的数据安全与隐私保护对策分析[J].情报杂志,2019,38(7):151-159.

[26] 袁 康,鄢浩宇.数据分类分级保护的逻辑厘定与制度构建——以重要数据识别和管控为中心[J].中国科技论坛,2022(7):167-177.

[27] 陈 兵,郭光坤.数据分类分级制度的定位与定则——以《数据安全法》为中心的展开[J].中国特色社会主义研究,2022(3):50-60.


[作者简介]赵梓羽，女，1998年生，浙江大学光华法学院博士研究生。

收稿日期：2023-08-10