生成式人工智能的数据安全风险及回应型治理策略

夏志成　施华钧*
深信服科技股份有限公司杭州分公司　浙江杭州　310000

摘　要：随着生成式人工智能技术的快速发展，其在各行业中的广泛应用为社会带来了巨大的便利，但同时也伴随着数据安全风险的隐患。生成式人工智能的核心依赖于海量数据的收集、处理与生成，这在提升生产效率的同时，可能导致个人隐私泄露、数据滥用及恶意生成等问题。本论文通过分析生成式人工智能在数据安全中的潜在风险，探讨了数据隐私保护、算法透明性和平台责任等关键问题，并提出了回应型治理策略。回应型治理强调通过法律法规、技术手段和社会监督等多维度手段，建立多层次的安全防护体系。具体而言，文章提出加强数据加密和去标识化处理、完善风险评估与审计机制、强化AI开发者和平台的伦理责任等措施，以确保生成式人工智能技术的安全可控发展。

关键词：生成式人工智能；数据安全风险；回应型治理策略

序言

生成式人工智能作为一种创新技术，正在迅速渗透到各个行业，包括艺术设计、文本生成、医疗、金融和教育等领域。其强大的自动化内容创作能力为企业和个人提供了前所未有的便利。然而，随着生成式AI模型的广泛应用，其带来的数据安全和隐私保护问题也日益严重，成为技术发展中的重大挑战。生成式AI的应用往往依赖于大量数据的采集、传输、处理和存储，这些数据涉及个人信息、医疗记录、金融数据等敏感内容。因此，数据隐私泄露的风险尤为突出。如果这些数据在未经充分保护的情况下被滥用或泄露，可能会导致个人隐私的严重侵犯，甚至引发身份盗窃、金融诈骗等社会问题。此外，生成式AI在进行数据处理和分析时，可能会受到不公平偏差的影响，导致模型产生歧视性或不公正的决策，进而加剧社会不平等。生成式AI的技术创新速度非常快，这一领域的监管和治理策略仍然滞后，由于缺乏统一的数据合规性框架和规范，很多企业在数据处理和使用过程中可能忽视了隐私保护和数据安全的要求。如何在推动技术创新的同时，有效应对数据安全和隐私保护问题，已经成为学术界和企业界的重点关注方向，通过建立完善的数据安全治理策略，包括加强数据合规性、透明的数据使用政策、提升数据隐私保护技术以及对AI模型的监督和审查，将是确保生成式AI可持续发展和应用的关键。

一、生成式人工智能的数据安全风险

（一）数据隐私泄露风险

生成式人工智能依赖于大量的用户数据进行训练，这些数据往往包含了个人隐私信息，如身份信息、财务数据、健康记录等。在AI系统处理和生成内容的过程中，未经授权的数据访问、存储、或错误的模型训练，可能导致敏感数据的泄露[1]。特别是在跨境数据流动的情况下，不同国家和地区的数据隐私保护规定差异化，使得生成式AI面临更高的隐私保护压力。数据隐私泄露不仅威胁到个人的隐私安全，还可能引发更广泛的法律责任和信任危机[2]。

（二）数据滥用风险

生成式人工智能技术具备强大的数据生成与再加工能力，但其生成的数据也可能被滥用，例如，生成虚假信息、深度伪造内容（Deepfake）或误导性数据等，可能对社会造成严重影响[3]。滥用的风险不仅包括AI生成的数据用于非法目的（如诈骗、造假、操控舆论等），还可能出现AI在没有明确授权或透明度的情况下，通过不当利用个人数据进行数据收集与处理，损害个人或集体

作者简介：
夏志成（1996.12-），男，汉族，江苏兴化人，本科，研究方向为网络安全和数据安全方向。
施华钧（1988.4-），男，汉族，江苏省无锡人，硕士研究生，研究方向为网络安全和数据安全方向。

的利益[4]。

（三）数据偏差和不公平

生成式人工智能的训练过程依赖于大量的数据集，然而这些数据集往往存在偏差。这些偏差来源于历史数据中隐含的社会偏见、性别歧视、种族偏见等，AI模型可能会无意识地复制和放大这些偏见，导致生成内容的不公平性[5]。对于像招聘、金融风控等敏感领域，数据偏差和不公平的产生可能导致不公正的决策，进一步加剧社会不平等和歧视现象。如何消除AI模型中的数据偏差，并确保生成内容的公平性，是当前生成式AI面临的一大挑战。

二、生成式人工智能数据风险的治理挑战

（一）数据来源合法性与合规性的治理挑战

随着生成式人工智能的普及，数据成为其核心驱动力。然而，生成式AI所依赖的数据来源是否合法、合规，成为治理中的一个重要挑战。首先，数据的合法性不仅要求收集过程符合相关法律规定（如《数据安全法》和《个人信息保护法》），还必须确保在使用过程中不侵犯个人隐私或商业机密。例如，一些AI模型使用的训练数据可能来自未经授权的数据集或通过不当方式获取的个人信息，这不仅违反了隐私保护法规，也对数据的准确性和公正性构成威胁。其次，数据的合规性要求生成式AI的数据处理和存储符合不同国家和地区的法律框架。在全球化背景下，跨境数据流动带来了合规管理的复杂性，不同地区对数据保护的要求可能存在差异，企业在运营时面临多重合规审查。此外，如何确保数据源的透明度和可追溯性，也是治理中的关键问题，特别是在AI模型产生不明确结果时，回溯数据源的合法性和合规性对于保障系统公正性至关重要。因此，加强数据来源的审查、确保数据采集和使用的合法性与合规性，成为生成式人工智能领域的重要治理挑战。

（二）信息存储与数据流动的治理挑战

生成式人工智能涉及大量数据的存储与流动，如何在保证数据安全的同时确保数据的高效使用，已成为治理的另一个挑战。首先，数据存储面临的主要风险包括数据泄露和丢失。由于生成式AI在处理海量数据时，其存储往往分布在不同的服务器和云平台上，这使得数据存储的安全性成为一个复杂问题。如果数据存储采取不当的加密手段或没有有效的安全控制措施，可能导致数据被黑客窃取或篡改，进而影响AI系统的可信度和安全性。此外，数据存储中的备份和恢复机制也至关重要，一旦发生故障或攻击，如何有效恢复数据，是确保生成式AI系统连续性和稳定性的必要条件。其次，数据流动的治理挑战主要集中在跨平台的数据传输过程中。尤其是对于涉及多个国家和地区的数据流动，如何保障数据传输过程中的安全，防止中途被截取或篡改，是当前生成式AI治理中亟待解决的问题。此时，数据加密、访问控制以及数据流动的透明度尤为重要。最后，在数据存储和流动过程中，合规性和隐私保护也是挑战之一，如何平衡数据流动的自由性与数据安全之间的矛盾，确保数据传输在满足法律要求的同时保持高效性，依然是需要进一步探索的领域。

（三）算法公平与无偏见的治理挑战

生成式人工智能在处理和生成数据时，其背后的算法设计直接影响着结果的公平性和无偏见性。然而，如何确保算法公正性，避免算法偏见带来的社会不公，是生成式AI面临的重大治理挑战。首先，生成式AI的训练数据可能包含偏见，尤其是历史数据中固有的社会偏见和不公平现象。若AI模型未经充分校正，可能会在生成内容时继承这些偏见，进而影响决策结果。例如，AI在生成招聘广告或司法判决时，可能无意识地加强性别、种族等方面的歧视，导致对特定群体的不公平对待。其次，算法的透明性和可解释性也是公平性治理中的一大难题。由于生成式AI的工作机制往往较为复杂，许多模型（如深度学习）本身是“黑箱”模型，这使得人们很难理解或追溯其决策过程。在缺乏透明度的情况下，如何评估算法是否存在偏见并采取有效措施进行纠正，成为治理的一大难点。

三、生成式人工智能数据风险的回应型治理策略

（一）数据来源风险治理

生成式人工智能（AI）模型的性能往往取决于其输入数据的质量和多样性。然而，数据来源的合法性、可靠性和完整性是潜在的数据安全风险所在。首先，数据输入的非法获取问题可能引发侵犯隐私、知识产权或违反数据保护法律的风险。例如，未经授权收集的个人信息、敏感数据或商业机密可能被用于训练生成式AI，导致侵犯用户隐私权或企业竞争力的法律纠纷。其次，数据输入的偏差也可能对生成式AI模型的输出产生不良影响。如果所使用的数据存在性别、种族或文化等方面的偏差，AI模型可能会产生带有偏见的结果，甚至加剧社会不平等。因此，在数据输入环节，必须建立严格的规制措施，包括确保数据来源的合法性、透明度以及符合

伦理标准。具体来说，政府和相关机构应制定清晰的数据采集与使用政策，明确数据使用范围、知情同意要求、隐私保护措施等内容。同时，企业和研发机构应加强对数据来源的审查与验证，确保输入数据的合法性和多样性，从而避免数据输入环节的风险对AI模型的质量和安全性产生不良影响。

（二）数据泄露风险治理

生成式人工智能的训练和应用需要大量的数据存储，因此，数据存储过程中的安全风险不容忽视。随着数据量的增加，数据泄露的风险也日益加大。首先，未经加密或未采取适当保护措施的存储系统容易受到黑客攻击，导致用户的个人数据、公司机密等重要信息泄露。对于生成式AI模型来说，数据泄露不仅涉及敏感信息的直接暴露，还可能影响企业的信誉和市场地位，带来法律责任和经济损失。因此，在数据存储环节，建立健全的数据安全保障措施是应对风险的关键。具体来说，企业和机构应采用先进的加密技术、访问控制机制以及数据备份方案，确保数据在存储过程中的安全性。同时，制定严格的权限管理制度，限制不必要的人员和系统访问敏感数据。政府监管机构应加强对数据存储领域的法律和标准化建设，要求企业遵守数据保护法规，并定期进行安全审计与漏洞检测。通过这些措施，可以有效降低生成式AI在数据存储环节的泄露风险，保护用户隐私和商业秘密。

（三）算法偏见风险治理

生成式人工智能模型的训练依赖于大量的数据和复杂的算法，其过程中的偏见风险可能对模型的公正性和准确性产生严重影响。算法偏见指的是在训练过程中，由于数据的不均衡性或模型设计的局限性，AI模型可能会产生对某些群体或个体的歧视性结果。例如，AI生成的内容可能会强化性别刻板印象、种族歧视或其他社会偏见，导致不公平的决策或输出。在模型训练阶段，规制算法偏见的风险尤为重要。首先，企业和研发机构应加强数据多样性和代表性，确保训练数据能够涵盖不同背景、文化和群体的需求，以避免偏见的发生。其次，算法设计应注重公平性和透明度，确保模型的决策过程能够接受外部审查和监督。政府应出台相关政策，推动算法透明度和公正性的提高，鼓励企业进行算法审计，并采取相应的纠正措施。此外，建立相关法律和伦理标准，要求生成式AI在实际应用中遵循非歧视性原则，避免算法偏见引发的不公平现象。通过这些措施，可以在模型训练过程中有效规制算法偏见风险，提升AI系统的社会责任感和公信力。

结语

综上所述，随着世界新一轮科技革命和产业变革的推进，生成式人工智能技术迅速发展，并与大数据、云计算、区块链等新一代信息技术相结合，推动了社会向智能化、融合化的方向发展。人机物高度融合的智能化社会正在逐步成型，技术的进步在自然语言处理、图像识别、音频处理、视频合成和机器翻译等领域取得了显著突破，极大地改善了生产力和社会运作效率。然而，在数据输入、存储、模型训练和内容输出的过程中，数据质量与真实性的保障成为关键。为应对这些挑战，通过动态治理体系、实时监管和自滤机制是确保数字空间安全可控的必要手段，通过加强监管措施有利于确保数据的安全性和社会公共利益的维护，可以在开放包容的基础上构建一个更加安全、可控的数字空间。

参考文献

[1] 钭晓东.论生成式人工智能的数据安全风险及回应型治理[J].东方法学，2023（5）：106-116.

[2] 刘智锋，吴亚平，王继民.人工智能生成内容技术对知识生产与传播的影响[J].情报杂志，2023，42（7）：123-130.

[3] 曾宇航，史军.政府治理中的生成式人工智能：逻辑理路与风险规制[J].中国行政管理，2023（9）：90-95.

[4] 钭晓东.论生成式人工智能的数据安全风险及回应型治理[J].东方法学，2023（5）：106-116.

[5] 高松.生成式人工智能的数据安全风险与应对措施[J].中国信息安全，2024（6）：32-37.