人工智能生成内容（AIGC）的安全风险评估框架

郝玉霞

菏泽市牡丹区卫生健康局　山东菏泽

【摘要】随着人工智能生成内容（AIGC）的迅猛发展，相关的安全风险逐渐引发广泛关注。AIGC 不仅带来了技术创新，同时也带来了数据泄露、恶意内容生成、版权侵犯等一系列潜在的安全隐患。制定科学合理的安全风险评估框架显得尤为重要。本文提出了一种基于多维度评估模型的 AIGC 安全风险评估框架，涵盖技术、法律、伦理等方面的风险识别与应对策略，旨在为政策制定者、技术开发者提供系统性指导，从而有效保障 AIGC 技术的安全应用与健康发展。

【关键词】人工智能生成内容；安全风险；风险评估框架；技术隐患；法律伦理

【收稿日期】2025 年 7 月 17 日　【出刊日期】2025 年 8 月 18 日　【DOI】10.12208/j.sdr.20250158

Security risk assessment framework for Artificial Intelligence Generated Content (AIGC)

Yuxia Hao

Heze Mudan District Health Bureau, Heze, Shandong

【**Abstract**】With the rapid advancement of Artificial Intelligence Generated Content (AIGC), related security risks have increasingly drawn widespread attention. While AIGC drives technological innovation, it also introduces potential vulnerabilities such as data breaches, malicious content generation, and copyright infringements. Establishing a scientifically sound security risk assessment framework has become crucial. This paper proposes an AIGC security risk evaluation framework based on a multi-dimensional assessment model, covering risk identification and response strategies across technical, legal, and ethical dimensions. The framework aims to provide systematic guidance for policymakers and developers, thereby ensuring the secure application and healthy development of AIGC technology.

【**Keywords**】Artificial Intelligence Generated Content; Security risk; Risk assessment framework; Technical hidden danger; Legal ethics

引言

随着人工智能技术的快速进步，AIGC 已成为创新和生产力提升的重要工具。其带来的安全风险问题日益严峻，影响范围从个人隐私到社会安全层面。如何评估和应对这些风险，成为当前技术发展的关键课题。探索适用于 AIGC 的安全风险评估框架，能够为全面识别并管控潜在威胁提供系统化的方案，同时为各领域安全应用的标准制定奠定基础，确保技术的长远可持续发展。

1 人工智能生成内容（AIGC）面临的主要安全风险

人工智能生成内容（AIGC）在技术创新的推动下，已被广泛应用于各个行业。随着其快速发展的同时，也暴露出一系列安全风险问题。这些风险不仅关系到技术本身的稳定性和可靠性，更涉及到广泛的社会层面，包括用户隐私、数据安全、内容真实性等方面[1]。AIGC 的生成机制依赖于大量数据的训练和模型的自主生成，这就意味着其生成内容的质量和安全性很大程度上取决于输入的数据质量、模型设计的严谨程度及其对潜在问题的预判。由于 AIGC 在生成过程中并未完全脱离人工干预，其结果可能会受到不当数据训练或不完备模型算法的影响，导致生成内容中存在错误、偏差，甚至产生有害内容。

除了技术层面的问题，AIGC 带来的隐私泄露风险也是其安全隐患之一。AIGC 系统通常需要大量的数据来进行训练和优化，而这些数据可能包含用户的

敏感信息。一旦数据未经适当保护或出现泄露，可能会导致用户隐私遭到侵犯，甚至引发数据滥用。更为严重的是，恶意攻击者可能通过针对 AIGC 系统的攻击手段，操控生成模型，故意生成带有恶意信息、虚假事实或带有歧视性的内容。这不仅会对公众产生误导，还可能被用作社会工程攻击、虚假宣传等恶意活动的工具，从而威胁到网络安全和社会稳定。

AIGC 技术的滥用可能带来版权保护和法律合规性的问题。AIGC 生成的内容在某些情况下可能侵犯原创作品的版权，尤其是在生成涉及艺术、文学等创意领域的内容时。由于生成模型的开放性，创作者和开发者可能无意中使用了未授权的数据源，从而导致法律纠纷[2]。AIGC 的普及也使得一些不法分子能够利用其生成虚假信息、煽动言论或制造极端内容，这在社会道德和法律层面都构成了巨大的挑战。面对这些复杂多变的安全风险，如何在技术、法律、伦理等多方面进行综合治理，确保 AIGC 的安全可靠运行，已成为全球范围内亟待解决的难题。

2 构建 AIGC 安全风险评估框架的必要性与挑战

随着人工智能生成内容（AIGC）技术的广泛应用，其带来的安全风险问题日益严峻，亟需建立有效的安全风险评估框架来应对这一挑战。AIGC 技术的复杂性和多样性使得风险评估工作不再局限于传统的技术层面，更多的是涉及数据安全、隐私保护、法律合规等多个维度[3]。构建一套全面且高效的风险评估框架，不仅能识别和预测潜在的技术问题，还能确保 AIGC 应用在合法性和伦理性上的合规性。这一框架的必要性主要体现在对新兴技术的深刻理解与适应上，只有通过科学的评估体系，才能为 AIGC 技术的健康发展提供保障，并避免技术滥用和隐患爆发。

构建这样一个安全风险评估框架并非易事，其面临的挑战不容忽视。AIGC 技术的不断演进和迭代使得其安全隐患呈现多样性，传统的风险评估方法往往难以涵盖所有新兴的风险点。AIGC 生成内容的质量控制、数据来源的合法性、以及生成模型的可解释性等问题，都是在框架构建过程中必须加以考虑的复杂因素。此外，随着 AIGC 应用场景的不断拓展，不同领域的特定需求也对评估框架提出了更高的要求。这意味着，评估框架不仅需要具备灵活性，能够适应不同的应用场景，还要能够处理复杂的跨领域安全风险，如跨国数据流动、不同法律体系下的合规性问题等。

在实施过程中，另一个重大挑战在于如何平衡技术创新与安全保障之间的关系。AIGC 技术在提高生产力、创造内容的同时，带来了前所未有的安全风险。如何在推动技术发展的同时，确保不发生数据泄露、内容伪造等重大事件，是评估框架必须解决的核心问题[4]。这要求在技术开发、政策监管、伦理监督等多方面加强合作，确保风险评估不仅仅停留在理论层面，而要具备实际操作性。实际应用中，评估框架需要具备动态更新和自我修正的能力，以应对快速变化的技术环境和日益复杂的安全威胁。

3 AIGC 安全风险评估框架的设计与实施策略

构建 AIGC 安全风险评估框架的设计需要从多维度考虑，确保能够覆盖技术、法律、伦理等多个层面的潜在风险。在技术层面，评估框架需要准确识别和分析生成模型可能面临的风险，数据偏差、算法漏洞、以及生成内容的不当性。为了应对这些技术风险，框架需要包括数据质量监控、算法透明度审查、以及对生成内容的有效监管。数据的来源必须合法且可追溯，算法的设计要避免黑箱效应，确保其输出是可控和可验证的[5]。同时，框架应具备内容审核机制，确保生成的内容符合社会伦理标准，不传播有害信息或虚假新闻。这些技术措施的实施，有助于降低 AIGC 应用中的技术风险，保障其安全稳定运行。

在法律和合规性方面，AIGC 的快速发展要求评估框架能够与现有的法律体系相结合，确保其应用不违反相关的版权法、隐私保护法等。框架的设计需要考虑跨域合规问题，特别是在多国或不同地区应用 AIGC 技术时，必须遵循当地的法律法规，避免引发法律纠纷。框架需要评估 AIGC 生成内容的版权归属问题，确保生成内容不侵犯他人的知识产权。此外，对于涉及个人数据的生成内容，框架应涵盖严格的数据保护措施，确保用户隐私不被侵犯。这些法律合规措施的加入，不仅提升了 AIGC 技术的合法性，也为其应用提供了明确的法律边界和操作规范。

为了确保评估框架的有效性和实施性，必须注重其动态调整能力和跨领域的协同合作。随着 AIGC 技术的不断发展，新的安全风险和法律挑战层出不穷，评估框架不能一成不变，需要具备灵活调整的机制[6]。框架设计应关注实时监控和反馈机制，及时发现潜在问题，并根据技术发展或政策变化进行适时调整。此外，为确保框架实施的可操作性，政府、企业、学术界等多个利益相关方需要在框架的设计

与执行过程中进行协同合作。通过跨领域合作，不仅能确保技术风险得到有效监管，还能够在伦理和法律层面达成共识，为 AIGC 技术的健康发展创造更为广泛的支持和保障。

4 AIGC 安全风险评估框架的实践效果与优化方向

AIGC 安全风险评估框架在实际应用中展现了显著的效果，特别是在降低技术风险和提高内容质量方面。通过对生成模型的深入分析与持续监控，框架能够有效识别模型在数据处理和内容生成中的潜在问题。针对生成内容的准确性与合规性进行严格审查，能够有效避免信息误导、偏见内容和虚假信息的传播。同时，框架通过对数据源和算法的监控，确保了数据合法性和技术透明度，减少了由算法黑箱问题带来的潜在安全隐患[7]。实践证明，在各大企业与技术开发者的应用中，评估框架已成为技术开发过程中不可或缺的一部分，大大提高了 AIGC 系统的安全性与可靠性。

尽管 AIGC 安全风险评估框架在多个领域取得了积极成效，依然面临一些挑战和局限。在多元化的应用场景中，框架的普适性和灵活性仍有待提升。不同应用领域对 AIGC 的安全要求不同，框架难以在所有情况下实现最佳效果，尤其是在一些高度专业化或跨境的应用中，现有的评估体系可能无法完全满足特定需求。此外，随着 AIGC 技术的不断进化，新的安全风险不断涌现，现有的框架可能在识别和应对这些新兴威胁方面存在滞后性。技术更新、法律政策变动及社会需求的变化，使得框架需要动态调整，以应对不断变化的安全挑战。如何快速响应这些新兴风险，保持框架的灵活性和实时性，成为当前框架优化的重要方向。

为了进一步优化 AIGC 安全风险评估框架，应重点加强框架的智能化与自动化能力。通过结合大数据分析和机器学习技术，框架可以在实时数据流中快速识别潜在风险，并自动调整风险评估模型[8]。这样的智能化调整不仅能够在技术发展过程中自动适应新的挑战，还能够有效降低人工干预的成本和错误风险。此外，跨领域的合作与法规的完善也是框架优化的重要方向。AIGC 的安全问题涉及多方面的利益关系，从技术开发者到政策制定者，再到最终用户，各方需要共同参与风险管理的各个环节，通过建立行业标准与全球范围的合作机制，推动评估框架的完善和实施。通过这些综合措施的优化，可以更好地实现AIGC技术在各个领域的安全应用，为其健康发展提供坚实的保障。

5 结语

AIGC技术的快速发展为各行各业带来了巨大的机遇，但随之而来的安全风险也不可忽视。构建一个科学、全面的安全风险评估框架显得尤为重要。这个框架不仅能够识别并应对技术、法律、伦理等多维度的潜在威胁，还能为技术的可持续应用提供保障。尽管框架在实践中已展现出积极效果，但随着技术的不断进步和应用场景的多样化，框架的优化和更新仍是一个长期的挑战。智能化、自动化的评估机制，以及跨领域的合作，将是未来框架优化的关键方向。只有不断完善风险评估体系，才能确保 AIGC 技术的安全性与合规性，实现其在社会中的健康发展。

参考文献

[1] 李文罡,马自飞,万傲霆,等.作物生产智能装备潜在安全风险分析[J/OL].中国农机化学报,1-7[2025-07-11].

[2] 李春晖,王醴弘.政务数据外部利用衍生数据之国家安全风险法律制度应对[J/OL].系统科学学报,1-10[2025-07-11].

[3] 刘晨晖,鲍宇婷,李清宏,等.压电理论视域下人工智能数据安全风险分析及监管问题研究[J/OL].情报杂志,1-8[2025-07-11].

[4] 陈广鹏.生成式人工智能对国家创新体系的赋能、风险及其因应[J/OL].西安财经大学学报,2025,(04):39-51[2025-07-11].

[5] 张曦.生成式人工智能赋能公共数字文化建设：场景、挑战与纾解路径[J/OL].四川轻化工大学学报(社会科学版),1-13[2025-07-11].

[6] 杜斌,郝敬密.建筑电气安装过程中的安全风险管理与防范技术研究[J].建筑工人,2025,46(07):25-27.

[7] 韩佶.数字技术驱动全球粮食安全治理的风险识别与消弭策略[J/OL].北华大学学报(社会科学版),1-11[2025-07-11].

[8] 龙泳汶.人工智能生成合成内容标识与知识产权归属问题研究[J].中国品牌与防伪,2025,(07):85-87.