元宇宙内容生成风险与人工智能素养的教育实践

赵小康
南京警察学院，江苏 南京 210023
DOI: 10.61369/SSSD.2025060014

摘　　要： 生成式人工智能（GenAI）作为当前最具颠覆性的技术之一，为元宇宙发展提供了强大动力。两者的结合不仅改变了创建和体验虚拟世界的方式，更将重塑未来数字经济、社交互动和文化创作模式。然而随着 GenAI 组件集成度的不断提高，一系列新风险、新挑战也随之产生。这既包括与传统数字系统处境相同的网络安全风险，也包括由 GenAI 组件和资产引发的特定于人工智能的漏洞。

关 键 词： 元宇宙；人工智能；教育实践

Risks of Metaverse Content Generation and Educational Practice of Artificial Intelligence Literacy

Zhao Xiaokang
Nanjing Police College, Nanjing, Jiangsu 210023

Abstract: As one of the most disruptive technologies currently, Generative Artificial Intelligence (GenAI) provides a strong driving force for the development of the metaverse. The combination of the two not only changes the way of creating and experiencing virtual worlds, but also will reshape the future models of digital economy, social interaction and cultural creation. However, with the continuous improvement of the integration degree of GenAI components, a series of new risks and challenges have emerged. These include not only the network security risks that are the same as those faced by traditional digital systems, but also the artificial intelligence-specific vulnerabilities caused by GenAI components and assets.

Keywords: metaverse; artificial intelligence; educational practice

一、生成式人工智能对元宇宙的基础支撑作用

元宇宙本质上是数字孪生、区块链和 AR、5G、GenAI等先进技术的融合体[1]。GenAI在元宇宙中扮演着基础性的技术支撑角色，解决了元宇宙大规模多样化内容高效生成的问题，极大地提升了元宇宙构件生产效率。它可以生成元宇宙中的建筑、玩家、NPC、环境要素、故事情节，根据用户数据和偏好生成定制化内容与体验，实现多语言实时互译，与数字孪生技术协同连接现实与虚拟世界。

GenAI能够提供比人工设计更出色的真实感和多样性，从而增强沉浸感与体验感[2]。GenAI赋予了元宇宙个性化定制的能力，可以根据用户需求生成多种方案，以及完成基于模糊需求的定制，使得元宇宙中的每个用户体验都可以是唯一的、特定的。多语种实时交互功能使得分布在世界各地的用户无障碍协作成为可能，对构建全球化的元宇宙社会至关重要。GenAI与数字孪生技术具有显著的融合创新效应。前者从大规模复杂环境中提取关键数据，优化并预测知识逻辑，进而创建自治系统应用于后者；而后者可以创建基于现实世界的数字模型。二者结合为元宇宙提供了虚实融合的基础架构[3]。

GenAI与元宇宙的协同创新突出体现在多模态算法与数字孪生技术两个维度。多模态大模型构成了GenAI支持元宇宙的核心技术基础，通过跨模态注意力机制提取语义表示，最终在虚拟场景中重新生成多模态数据，支持高质量图像、文本和音视频内容创作，满足元宇宙对多样化媒介的需求[4]。GenAI通过学习数据集上的模式和规律实现高度泛化的内容生成[5]。基于数字孪生技术，元宇宙将现实世界映射至虚拟空间，在这一过程中，GenAI能够模拟内容的自动化生成与运营，进而创建自治系统。影谱科技 ADT Meta版引擎就是 GenAI与数字孪生融合的典型案例。

二、元宇宙内容生成的网络安全风险

GenAI面临的网络安全风险涵盖整个 AI系统，既包括智能交互、数据驱动、功能增强等 AI组件，也包括建立在多个组件和工具之上的完整系统，还包括其训练数据。

（一）供应链攻击

人工智能系统与传统软件面临同样的困境，未能摆脱供应链中的漏洞危害。这些漏洞可能破坏训练数据集、模型和平台，导致输出偏差及其他安全问题。随着 LoRA （Low-Rank

基金：由2022LL69项目资助
作者简介：赵小康（1982-），男，河北宣化人，副教授，教务处副处长，研究方向为情报分析、公安情报。

Adaptation）等 PEFT （Parameter-Efficient Fine-Tuning）框架技术在大模型优化微调中的应用，以及设备端模型的出现，进一步加剧了人工智能供应链的复杂性[6]。供应链攻击对于 GenAI 的影响主要在于数据篡改和模型中毒攻击两方面。

GenAI基础模型的成功训练高度依赖训练数据的规模和多样性。用于预训练 GenAI模型的海量数据集主要来源于互联网，其内容往往未经核实，甚至具有潜在危害性。这些缺乏监管的数据集因此成为巨大的潜在攻击对象。攻击者可能在其中嵌入对抗样本，引入漏洞、后门、偏见，进而损害模型性能，最终用于传播虚假信息、提供不安全代码等。

在开发人员依赖开放权重模型的情况下，通过开源平台分发的模型可能携带隐藏风险，比如嵌入在模型源代码中的恶意软件、后门等。这些威胁通常处于休眠状态，在特定条件下才会被激发，比如模型加载时或接受到包含特定单词、短语等“扳机”信息时。这无疑极大地提高了检测难度。一旦攻击者将后门插入预训练模型中，即使在微调或额外的安全训练之后，这些后门也可能持续存在。

（二）提示词直接注入攻击

提示词直接注入攻击是指提示词以意外的方式改变 GenAI模型的行为或输出。这可能导致其违反相关规则、生成有害内容、影响关键决策、实施未经授权的访问等。该类攻击主要包括基于优化的攻击和手动攻击。

基于优化的攻击通过系统地算法优化方法来完善对抗性提示，其目的是将生成的恶意或有害响应的概率最大化。通常会通过优化对抗性后缀来实现这一目的，从而规避 GenAI模型的安全防范措施。这些后缀可以在不同模型间转移。这使得提供白盒访问权限的开放权重模型成为针对仅提供 API 访问权限的封闭系统的可转移攻击的理想载体。

手动攻击的目的是触发 GenAI模型的目标冲突和不匹配泛化。当模型的能力与安全目标发生冲突时，会出现相互冲突的目标。最为突出的情形是利用角色扮演策略，将模型推入与其初始意图相冲突的状态，从而破坏其安全协议。

此外，攻击者还有可能利用辅助语言模型自主生成和优化攻击性提示词。

（三）提示词间接注入攻击

提示词间接注入攻击是指外部输入以意外的方式改变 GenAI模型的行为或输出。此类攻击由恶意第三方实施，无需直接与底层模型交互，主要导致3类问题。一是破坏可用性，包括诱发模型执行耗时操作、阻断模型使用特定 API或工具，以及破坏模型输出等手段。二是破坏完整性，包括诱导模型以攻击者指定的信息作为回应或重定向用户至恶意网站、恶意内容等手段，以此传播误导信息、推荐欺诈产品或服务、压制或隐藏特定信息等。三是破坏隐私性，导致敏感信息泄露，主要的手段是说服用户提供信息，然后将其泄露给攻击者。

三、元宇宙内容生成的信息安全风险

（一）信息析取

GenAI模型在整个生命周期中可能触及大量、多类信息。一是其训练数据可能包含未匿名处理的个人信息、机密训练数据、受版权保护的资料等。二是当系统采用检索增强生成（RAG）流程时，敏感信息可能成为模型输入的一部分。对于系统自身而言，模型的权重、架构、系统提示也属于敏感信息。这些信息可能通过成员推断、模型反转、模型提取等手段被攻击者析取。泄露敏感数据将导致针对 GenAI系统提供商的法律诉讼和处罚等。

成员推断攻击的核心目标是判断某个特定数据样本是否被用于目标模型的训练数据集。此类攻击无需直接访问数据库或文件系统，而是通过分析模型的输出特性间接推断数据归属，具有更强的隐蔽性[7]。由于生成模型通常具有更大的模型容量和更强的记忆力，这使得它们更容易记住训练数据中的独特样本或罕见模式。这种记忆效应虽然有助于模型生成更准确和连贯的内容，但也为成员推断攻击提供了可乘之机。成员推断攻击的实施通常依赖于目标模型、攻击模型、影子模型3个组件。

模型反转攻击是指从模型输出中重建训练数据或推断敏感信息。当模型基于敏感数据（如金融、医疗、执法等信息）进行训练时，模型反转可能会导致隐私泄露。此类攻击需要具备访问模型的权限，通过对模型的逆向工程提取有关原始训练数据的信息。其风险主要来源于生成模型对训练数据的高度再现，在技术机理上同样利用了 GenAI模型的记忆特性。根据攻击者对目标模型的了解程度，可分为白盒、灰盒和黑盒3种场景。

模型提取攻击是从模型中提取参数从而获得模型副本。此类攻击与训练数据提取有关，但目标不同。前者旨在窃取模型参数，而后者则试图提取用于生成这些参数的训练数据。模型提取攻击可以从 GenAI系统的黑盒模型中获得精确信息。

（二）信息操纵

GenAI变革了内容创作模式，它能够帮助元宇宙以极快的速度和规模生成极具说服力的内容。这一能力可用于主导社交媒体讨论、模仿权威新闻机构，进而误导公众，削弱其对媒体的信任，扭曲其认知和决策。这种操纵可能会产生深远的影响。虚假信息的迅速生成往往会超过有效反驳它的能力，因为准确且可验证的回应需要投入大量时间和精力。面对 AI生成的错误 / 虚假信息时，要坚守基于事实的可信交流，维护信息的完整性，往往是一场艰苦的战斗。为 AI生成的内容添加水印，以及通过可信来源核实信息等技术解决方案很有价值，但效果仍然有限，不足以从根本上解决问题。

当 AI被用于生成甚至总结内容时，无意识的偏见已经带来了误导内容的风险。比如使用 GenAI获取信息，相关答案可能会因 AI所基于的训练数据集而产生偏颇。当 AI被有意用于操纵信息时，GenAI的使用就会产生空前的风险挑战，它能够大量、快速生成深度伪造的音视频内容[8]。

另一种攻击方式是基于 AI的伪装技术来克隆媒体、金融、政府机关等公共机构的真实网站，大规模、系统性的数据污染。

由 GenAI驱动的机器人会助推误导性叙事，操纵信息传播，挑起针对特定群体的争端，引导和放大公众负面情绪，通过“带节奏”“混淆视听”来扰乱元宇宙和现实中的社会秩序。

当然，GenAI也有助于打击信息操纵及其传播。这包括利用生成式人工智能进行清晰且有针对性的沟通、事实核查、识别 AI 生成的内容等。

四、人工智能素养的教育实践

防范元宇宙内容生成风险的有效手段之一在于提升政府、媒体、公众的人工智能素养。通过赋予机构和个人批判性接收 AI生成内容的能力，使其能够辨别虚假或不准确的信息，提供公众对虚拟和现实世界中的内容的质疑分析能力，从而增加决策的正确性，实现对虚假内容的防范抵御。素养提升与技术防范相结合，对应对 GenAI的风险挑战至关重要。

人工智能素养作为一个新兴概念，其定义随着技术发展和社会需求不断丰富拓展。Kandlhofer等人于2016年首次定义了人工智能素养，将其描述为理解不同产品、服务背后 AI基本技术和概念的能力。在最新研究中，赵益民等人进一步丰富了以人工智能意识、人工智能知识、人工智能能力和人工智能伦理为主要维度的人工智能素养结构框架 [9]。

人工智能素养教育关乎个体发展和社会进步双重目标。对个体而言，它帮助人们提高在 AI时代的竞争力。对社会而言，它能够确保更多的人平等享受 AI带来的便利，减少因技术差距导致的社会不平等，以及增强伦理意识和社会责任感。人工智能素养既鼓励积极应用人工智能，同时必须有意识选用正确、积极、有效、创新的生成过程，杜绝错误、消极、无效、陈旧的生成结论 [10]。

从教育实践看，人工智能素养教育经历了“信息素养”“数字素养与技能”“人工智能素养”等演进过程。2018年，教育部颁布《普通高中课程方案和课程标准（2017年版）》，提出了提升学生信息素养的培养目标。自2022年起，中央网信办等四部门每年发布《提升全民数字素养与技能工作要点》。2022年，教育部颁布《义务教育课程方案和课程标准（2022年版）》，提出了提升中小学生数字素养与技能的培养目标。高等教育作为创新型人才培养的主阵地，正在加速人工智能通识教育的布局。2021年，浙江大学、复旦大学、中国科学技术大学、上海交通大学、南京大学、同济大学、华为、百度、商汤在上海成立新一代人工智能科教育人联合体，推出了 AI+X 微专业，专业实现共建共选、学分互认、证书共签和 SPOC授课形式。2024年，浙江大学发布《大学生人工智能素养红皮书（2024版）》，推出了“人工智能基础”系列通识课程，成立了人工智能教育教学研究中心。

五、结语

面对元宇宙内容生成风险，全球视野下的 AI素养教育正呈现出趋同发展与多元创新并存的格局。各国都在根据自身的技术基础、教育传统和社会需求，探索适合本国国情的人工智能素养培养路径。欧盟推出了“XR批判性思维”框架，韩国把数字孪生校园设为实验场。同步兴起的跨国协作平台，正以“开源沙盒”等方式共享风险案例库，形成兼具全球共识与本土韧性的防护网。

参考文献

[1]赵国栋，易欢欢，徐远重．元宇宙 [M]. 中译出版社，2021.

[2]企鹅号．生成式 AI能为元宇宙带来什么？ [EB/OL] [2023-04-04] https://cloud.tencent.com/developer/news/1045383

[3]曹明伟，张迪，彭圣洁，等．元宇宙技术发展与应用综述 [J]. 计算机科学，2025，52 (03)：4-16.

[4]周晨．探索“元宇宙”：空间计算技术、生成式 AI、XR产业、芯片技术，这些领域又有了新动作 [N]. 新闻晨报，2024-10-24.

[5]黎浩田．何以促进：元宇宙支撑技术的治理逻辑与规范路径——以生成式人工智能为例．智慧法制 [C]，2025，1.

[6]赵月，何锦雯，朱申辰，李聪仪，张英杰，陈恺．大语言模型安全现状与挑战 [J]. 计算机科学，2024，51(1)：68-71.

[7]彭铖峰，赵波，刘会，等．针对机器学习的成员推断攻击综述 [J]. 计算机科学，2023，50 (03)：351-359.

[8]许敏，肖书婿．生成技术与视听操纵：AIGC时代深度伪造的内在机理与治理策略 [J]. 北京邮电大学学报（社会科学版），2024，26(4)：27-35.

[9]钟柏昌，刘晓凡，杨明欢．何谓人工智能素养：本质、构成与评价体系 [J]. 华东师范大学学报（教育科学版），2024，42(01)：71-84.DOI：10.16382/j.cnki.1000-5560.2024.01.005.

[10]袁振国．重塑未来——教育数字化之于教育强国建设的突破性意义 [J]. 教育研究，2024，45(12)：4-12.

