人工智能生成内容技术在内容安全治理领域的风险和对策

乔喆

（中国移动通信集团有限公司信息安全管理与运行中心，北京 100053）

摘　要：近年来，人工智能生成内容（artificial intelligence generated content，AIGC）技术取得了颠覆性成果，成为 AI 领域研究和应用的新趋势，推动着人工智能进入新时代。首先，分析了 AIGC 技术的发展现状，重点介绍了生成对抗网络、扩散模型等生成模型和多模态技术，并对现有的文本、语音、图像和视频生成的技术能力进行调查阐述；然后，对 AIGC 技术在内容安全治理领域带来的风险进行重点分析，包括虚假信息、内容侵权、网络与软件供应链安全、数据泄露等方面；最后，针对上述安全风险，分别从技术、应用和监管层面，提出应对策略。

关键词：人工智能生成内容；生成模型；多模态技术；内容安全治理

中图分类号：TP399

文献标志码：A

doi: 10.11959/j.issn.1000−0801.2023190

Risks and countermeasures of artificial intelligence generated content technology in content security governance

QIAO Zhe

Information Security Management and Operation Center, China Mobile Communications Group Co., Ltd., Beijing 100053, China

Abstract: Recently, artificial intelligence generated content (AIGC) technology has achieved various disruptive results and has become a new trend in AI research and application, driving AI into a new era. Firstly, the development status of AIGC technology was analyzed, focusing on generative models such as generative adversarial networks and diffusion models, as well as multimodal technologies, and surveying and elaborating on the existing technological capabilities for text, speech, image and video generation. Then, the risks brought by AIGC technology in the field of content security governance were focused and analyzed, including fake information, content infringement, network and software supply chain security, data leakage and other aspects. Finally, in view of the above security risks, counter strategies were proposed from the technical, application and regulatory levels, respectively.

Key words: AIGC, generative model, multimodal technology, content security governance

0　引言

随着人工智能（artificial intelligence，AI）的快速发展和普及，AIGC 已经成为当今社会中一个备受瞩目的领域。AIGC[1]是指利用深度学习等人工智能技术生成的内容，包括但不限于文本、音频、图片和视频。严格意义上说，1957 年莱杰伦·希勒和伦纳德·艾萨克森完成的人类历史上第一支

收稿日期：2023−05−24；修回日期：2023−10−11

由计算机创作的音乐作品——《伊利亚克组曲》，就可被看作 AIGC 的开端，至今已有 66 年。

在人工智能发展初期，受限于数据、算力、算法等各方面因素，AIGC 技术大多基于预先定义的规则或模板[2]，与真正的智能创作相去甚远。近年来，一方面，大数据技术不断成熟，基础算力大幅提升，基于生成对抗网络和扩散模型等生成式人工智能的 AIGC 技术快速迭代[3]，彻底打破了模版化、公式化生成方法的局限，可灵活生成丰富的多模态内容。例如，对于文本生成而言，早期大多采用基于模版的方法，如早期的对话系统[4]基于符号规则和模版填充相关信息，而如今的大型语言模型[5]能够进行可控文本生成，内容长度不断突破。另一方面，随着数实融合的不断深入，人类对数字内容的质量和丰富度需求空前高涨，大量 AI 生成作品的出现降低了高质量数字内容生产的专业门槛[6]。直到 2022 年，才算是 AIGC 的爆发之年，从 AI 绘画到 ChatGPT，人们看到了 AIGC 无限的创造潜力和未来应用的可能性。

然而，AIGC 技术的广泛应用带来了一系列内容安全治理方面的挑战。在这个数字信息爆炸的时代，信息的真实性、准确性和合规性成为亟待解决的问题[7]。滥用 AIGC 技术产生的大量虚假信息、不良信息、违法信息可能会给人们带来严重的负面影响，甚至威胁人们的生命财产安全。例如，AIGC 技术可能被用于制造虚假信息进行敲诈勒索、生成恶意代码进行网络攻击等，破坏网络生态和社会安全。AIGC 技术的普及使内容安全治理面临着前所未有的风险。同时，与风险并存的是 AIGC 技术所带来的巨大机遇。例如，利用 AIGC 技术，可以自动化地统计分析数字内容，从而提高内容审核的效率和准确性，降低人力成本，为内容平台提供更好的服务。

因此，本文将对 AIGC 技术在内容安全治理领域面临的风险和机遇进行深入探讨。首先，概述 AIGC 技术的发展现状，以及其在文本、语音、图像和视频等内容创作上的技术能力和应用案例；接着，探讨 AIGC 技术在内容安全治理领域所面临的主要风险，包括虚假信息、网络安全和隐私保护等方面；最后，针对这些风险挑战，分别从技术、应用和法律层面总结分析了在当前 AIGC 技术发展和应用背景下的工作建议与应对策略。本文为相关领域的学术研究和实践工作提供了有益的参考，致力于推动 AIGC 技术的健康发展和内容安全治理的不断进步。

1　AIGC 技术发展现状

1.1　AIGC 技术爆发标志着人工智能研究热点从分析式向生成式转变

分析式人工智能需要搜集大量标注数据形成训练集，并学习训练数据中数据与数据标记之间的最佳映射规则，进而完成各类机器学习任务。生成式人工智能通常采用无监督方式学习海量未标注数据的模式与统计特征，采用少量人工标注数据训练或无须训练就可以完成各类机器学习任务。生成式人工智能可以充分利用海量的未标注数据进行学习，相比分析式人工智能具有更强的泛化能力[8]。判别式模型和生成式模型的区别如图 1 所示，对比机器学习中的判别式模型和生成式模型，判别式模型估计条件概率分布，根据 $P(Y|X)$求得标记 Y，强调直接从数据中学习决策函数；而生成式模型估计联合概率分布 $P(Y, X)$，强调学习数据生成的规律，从而更好地对数据进行表征建模。

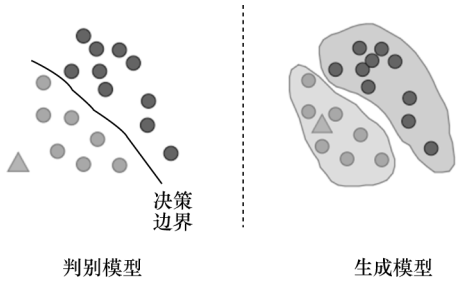

图 1　判别式模型和生成式模型的区别

1.2 生成式人工智能相比分析式人工智能存在三大优势

一是分析式人工智能任务大部分可以直接用生成式人工智能模型来解决，但反之不成立。例如，文本分类是一个典型的分析式人工智能任务，可以引导生成式人工智能模型生成类别标签来实现文本分类。但普通的文本分类模型最终输出的概率分布是针对人类预先定义的类别标签，而不是针对词表的，无法直接生成文本内容。

二是生成式人工智能泛化能力远超分析式人工智能。在生成式人工智能模型构建中，随着模型参数和训练数据不断增加，模型性能稳步提升。目前在自然语言领域，生成式人工智能模型已经全面超越分析式人工智能模型，并不断逼近人类水平甚至超越人类水平。例如，GPT-4 在各类考试中[9]，排名超过大多数人类。

三是生成式人工智能可以完成逻辑场景的自动转换，而分析式人工智能针对不同的应用场景则需要重新设计并训练模型。生成式人工智能一旦训练完成，通常可以处理不同任务（如文本分类、摘要、生成等），在推理阶段能够自动转换逻辑场景。即使有（已知逻辑下的）新的学习任务时，也无须训练人员干预，仅需提供给生成式模型少量样例进行学习（in context learning）[10]，模型就能够完成相应任务，模型本身不必重新训练。生成式人工智能是当前人类迈向通用人工智能最可能的途径。而分析式人工智能一旦训练完成，则只能够处理与训练数据同分布的任务。分析式人工智能只能局限于专有人工智能，在特定领域中可以较好地发挥作用，但通用性受限。

1.3 生成对抗网络、扩散模型、ChatGPT、跨模态对齐等技术催生 AIGC 大爆发

（1）生成对抗网络（generative adversarial network，GAN）、扩散模型（diffusion model）使 AI 生成的内容更加精细逼真。生成对抗网络[11]大幅提升 AI 生成内容的逼真程度。其由一个生成器和一个判别器组成，模型使用合作的零和博弈框架来学习。生成器的训练目标是使生成内容真假难辨，判别器的训练目标是精准辨别内容真假。生成器将内容输出给判别器进行判别，判别器将判别结果反馈给生成器进行改进。得益于双方博弈的学习策略，生成对抗网络生成内容的真实性和清晰度得到极大提升。该技术成功应用于图像、视频、语音和三维物体等多种内容的生成。如英伟达发布的 StyleGAN-XL 模型[12]生成的高分辨率图片，人眼难以分辨真假；又如 DeepMind 发布的 DVD-GAN 模型[13]用来生成连续视频，在草地、广场等明确场景下表现突出。

扩散模型[14]使多媒体内容产业爆发式发展。近年来，扩散模型在图片生成任务中超越了 GAN，并且在诸多应用领域都表现出色。扩散模型原理受热力学概念启发，即通过不断给原始内容增加噪声将内容最终转化为纯噪声，再通过学习如何逆转这种噪声叠加过程从纯噪声恢复原始内容。这一过程与人类进行创作的过程非常近似，人类在创作初期的思路可能是零散而混乱的，接近噪声。在创作过程中人类的思路逐渐清晰，最终经过不断地修改和完善（去噪声）形成最终的作品。相比生成对抗网络，扩散模型无须引入判别器，架构更加简单，训练更加稳定，在各方面任务中都超过了生成对抗网络。例如，著名的 DALL-E、Stable Diffusion、Midjourney 等都是基于扩散模型的产品。2022 年，一位游戏设计师使用 Midjourney 生成的《太空歌剧院》画作获得美国科罗拉多州博览会一项美术竞赛一等奖，这标志着 AI 创作画作首次击败人类艺术家。

（2）ChatGPT 叩响了通用人工智能的大门。通用人工智能[15]是指一种能够像人类一样在各种领域中执行任务的智能系统。ChatGPT 的诞生让人们看到了通用人工智能的曙光。2022 年 12 月，OpenAI 推出的 ChatGPT 不仅具有出色的人类意图

理解能力和多轮对话能力，还能够帮助人类完成种不可思议的任务，并且通晓 100 多种语言。近期，OpenAI 进一步推出了 GPT-4，其在多种人类考试中排名超过大部分人类，可以结合图片和文本综合理解人类意图，并给出正确反馈。同时，OpenAI 发布了 ChatGPT 的插件系统，可实现与 Web 搜索、科学计算等上千个插件应用对接。OpenAI 将 ChatGPT 定义为一种新型智能计算机架构，它不但可以和人沟通，还可以和任何插件应用沟通，并可以自主协同调用多个应用完成人类的复杂任务需求。显然，ChatGPT 可以与其他多媒体生成产品进行协同，帮助人类更高效地完成各类多媒体内容生成工作。

（3）跨模态对齐技术使能自然语言驱动多媒体生成。没有见过“老虎”的人读再多的书也不知道“老虎”长什么样子，同理，仅学习大量人类的语言限制了 AI 理解现实世界的能力。跨模态数据对齐技术极大地提高了 AI 对现实世界的认知能力。跨模态算法 ViLBERT[16]采用 Bert 提取文本特征，基于预训练的目标检测网络生成图像预选框及其视觉特征，依靠 Co-TRM 完成文本和视觉特征的融合。2021 年，OpenAI 发布的 CLIP[17]采用 Transformer 提取文本和图片特征，能够有效地将文字和图片中的物体、语义、风格等信息进行对齐。此外，多模态模型 UNIMO[18]和 FLAVA[19]将图文特征提取统一起来，即训练一个统一的大模型，使得该模型既能很好地适配单模态数据（文本或图像），又能很好地适配多模态数据。在中文方面，文心 ERNIE-ViLG[20]是目前全球规模最大的中文跨模态生成模型之一，该模型通过自回归算法将图像生成和文本生成统一建模，增强模型的跨模态语义对齐能力，显著提升图文生成效果。

2021 年 3 月 OpenAI 发布 AI 绘画产品 DALL·E，只需要输入一句文字，DALL·E 就能理解并自动生成一幅意思相符的图像，背后的关键技术即 CLIP。类似产品还有 Stable Diffusion、Midjourney 和 Gen-2 等。斩获 7 项奥斯卡大奖的科幻电影《瞬息全宇宙》中的部分视觉效果，使用的是 Runway 公司的视频生成产品，利用自然语言进行视觉场景构建，以大幅提升视觉特效的制作效率。

2　AIGC 技术能力分析

2.1　AI 生成文本：全方位逼近人类水平

- 日常对话问答能力：善解人意，可促膝长谈。ChatGPT 和文心一言都通过学习和理解人类的语言进行对话，不仅能够根据聊天的上下文进行互动，真正像人类一样聊天交流，而且能够主动承认错误和无法回答的问题，大幅提升了对用户意图的理解能力。
- 阅读理解和文本编辑写作能力：考试成绩超越部分普通人类。OpenAI 发布的 GPT-4 具备强大的推理能力，不仅能够理解图表进行数字计算，而且在各大考试（GRE、SAT 等）中，几乎取得了满分成绩[9]。
- 语言翻译能力：支持世界百种语言互译。在文本翻译方面，GPT-3 当时已被证明可以与最先进的机器翻译系统媲美。2022 年 9 月 21 日，OpenAI 发布 Whisper 自动语音识别（automatic speech recognition，ASR）系统，经过 68 万小时的多语言和多任务监督数据训练，支持多种语言的转录，以及将这些语言翻译成英语。
- 图片描述能力：可看懂梗图笑点。VisualGPT[21]是一个由 OpenAI 开发的图像描述模型，能够利用预训练语言模型 GPT 中的知识，其可以在多种领域应用，包括对少见的物体进行描述。大型多模态模型 GPT-4 能接受图像和文本输入，再输出正确的文本回复，不仅能够生成图像的描述，还能理解图像的含义进行问答。
- 代码理解和生成能力：精通多种编程语言

开发与漏洞修复。OpenAI 的大型语言模型训练数据包括 GitHub 上的开源代码库，使得 AIGC 不仅能够理解人类的自然语言，还能读懂计算机语言。利用 AIGC 技术可以给代码生成对应的功能注释，甚至可以生成特定用途的代码、实现代码语言之间的相互转换和漏洞修复等。例如，GitHub Copilot 是一个 GitHub 和 OpenAI 合作产生的 AI 代码生成工具，可根据命名或者正在编辑的代码上下文为开发者提供代码建议。

2.2 AI 生成语音：语音表达连贯自然，声音模仿惟妙惟肖

- 文本生成语音的能力：语音表达连贯自然，字正腔圆。利用 AIGC 技术对给定文本生成语音，已经广泛应用于客服及硬件机器人、有声读物制作、语音播报等任务。例如，喜马拉雅运用文本—语音转换（text to speech，TTS）技术重现单田芳声音版本的《毛氏三兄弟》和历史类作品。
- 语音复制能力：声音模仿惟妙惟肖，真假难辨。AIGC 的语音复制技术可以利用目标人的一小段语音，生成目标人在任意文本上的语音，如 AI 拟声工具 ElevenLabs 提供的嗓音复刻。

2.3 AI 图像生成：水平超越大部分人类画师

- 图像局部生成和更改能力：图片实现移花接木、偷天换日。利用 AIGC 技术可以进行图片去水印和换脸等局部内容修改。AI 去水印的工具，如水印云，能够支持一键去除图片中的文字、标识、人物、瑕疵等内容，消除图片中多余元素。在 AI 换脸方面，基于 Deepfake[22]已经衍生出很多换脸应用，如 Faceswap、FakeApp、DeepFaceLab 等。
- 草图生成高清图像：“神笔马良”成为现实。利用 AIGC 技术以图生图，主要指基于草图生成完整图像，如基于草图生成人脸的 Deep Face Drawing 等。
- 文本生成图像能力：心想“图”成，画出各种奇思妙想。OpenAI 发布的 DALL·E 2[23]能够根据文本描述创建图像，还可以基于文本引导进行图像编辑。从 Midjourney 的 V1 到 V5 版本，文本生成图像的质量提升明显，已经可以生成电影大片级图像。
- 精细化控制图像生成能力：实现图像精准 PS。有时人们希望对生成图片的轮廓、深度、边缘、物体姿态等内容进行精细化控制，而这些信息很难精确地使用自然语言描述。ControlNet[24]开创性地实现精准控图，其可以自动从参考图像中提取上述精控信息，并在这些信息的指导下进行图片的创作。其可以很好地应用于建筑设计、室内装饰风格设计等场景。

2.4 AI 生成视频：轻而易举地合成虚拟视频和特效视频

- 视频属性编辑能力：黑白影片智能上色，模糊视频变高清视频。利用 AIGC 技术，可以进行视频画质修复、局部画面修饰、生成视频特效和自动美颜等。例如，Runway 公司发布的人工智能视频编辑模型 Gen-1[25]，可对视频素材进行转换，并使用文字指令加以剪辑，如更换影片中车辆的颜色。
- 视频自动剪辑能力：智能识别精彩片段创作电影预告片。典型案例包括 Adobe 与斯坦福大学共同研发的 AI 视频剪辑系统、IBM Watson 自动剪辑电影预告片以及 Flow Machine。
- 文本生成视频能力：创意文案速成特效大片。文本生成视频，即根据给定文本生成符合描述的短视频。例如，Meta 公司推出的一款人工智能系统模型 Make-A-Video[26]，可以根据给定的文字提示生成异想天开、独一无二的

视频。Runway 推出的 Gen-2 AI 模型，可直接输入文字生成短视频。

3　AIGC 技术内容安全风险警示

3.1　滥用 AIGC 技术生成虚假信息危害社会安定团结

- 伪造虚假身份实施电信诈骗。利用 AIGC 技术，可以提取音频样本的声纹特征进行语音复制，轻松模仿一个人的声音，从而实施诈骗。同时，可以生成各类话术剧本，还能创建一个拥有图片、语音和视频信息的“虚拟角色”实施电信诈骗。例如，根据《华尔街日报》报道[27]，2019 年 3 月，犯罪分子通过商业化的人工智能语音生成软件，成功模仿并冒充一家德国公司的 CEO，欺骗其多位同事和合作伙伴，一天内多次诈骗并转移资金，使得该公司损失 220 000 欧元（约折合 173 万元）。
- 制造虚假证据用于勒索。勒索者通过 AI 换脸软件将受害人人脸移植到裸模身上或正在从事不法活动的人身上，用于向受害人本人或家属进行敲诈勒索。
- 制造虚假新闻操控舆论走向。利用 AIGC 技术，可以短时间内制造大批谣言，特别是生成关于政治人物的虚假图片、视频、音频，生成带偏见的评论等信息进行舆论走向的操控。
- 生成虚假日志欺骗日志审计。攻击者利用 AIGC 技术根据该系统已有的日志内容生成逼真的虚假操作行为日志，并替换已有的操作日志，从而更好地隐匿攻击行为，同时使日志审计功能失效。
- 使用 ChatGPT 生成诈骗内容。目前可以让 ChatGPT 以举例的方式生成诈骗内容，并让其进一步对内容进行改写和拓展。

AIGC 本身并不具有主动意识和动机，只是一个根据数据和模型生成内容的工具，其生成虚假信息的主要原因有以下几点：一是训练数据偏差，AIGC 的训练数据本身存在偏见、错误或虚假信息，在生成内容时可能会重复这些偏见和错误；二是统计模式自身局限，AIGC 模型是基于统计学习的，通过学习大量数据中的模式和规律生成内容，但有时候这些模式可能并不代表真实的事实，而是简单地反映数据中的频率和共现关系；三是缺乏判断力，AI 缺乏人类的判断力和理解能力，不能像人类一样辨别真实信息和虚假信息，只是简单地根据概率生成内容；四是对抗性样本，对抗性样本是一种特别设计的输入，可以欺骗 AIGC 模型，导致输出错误或虚假结果；五是数据覆盖范围有限，训练数据通常只涵盖特定领域或特定类型的内容，这导致模型在其他领域或内容上可能表现不佳，并容易输出虚假信息。

3.2　AIGC 生成侵权内容导致法律纠纷

人工智能撰写的文章等 AIGC 作品存在著作权归属不清的现实困境[28]，其主要根源在于训练数据可能包含版权保护的内容。这一问题不仅可能导致使用 AIGC 技术创作的作品无法获得著作权保护，阻碍人工智能技术发挥其创作价值，还有可能因人工智能的海量摹写行为稀释既有作品权利人的独创性，威胁他人的合法权益。2023 年 3 月 16 日，美国版权局发布新规，人工智能自动生成的作品不受版权法保护，坚持只有人类创作的内容才能得到版权保护。

3.3　AIGC 生成问题代码威胁网络与软件供应链安全

AIGC 的训练数据可以包括各类开源代码库，一方面可能包含恶意软件的代码，导致能够生成网络攻击进而攻击相关代码程序；另一方面开源仓库代码质量良莠不齐，导致生成代码可能存在缺陷。

- 协助非专业人员实施网络攻击。利用

ChatGPT 等大型语言模型可以快速编写恶意软件，如生成钓鱼邮件协助网络攻击；生成加密工具远程锁定他人计算机，由此进行勒索；生成攻击脚本，对网络用户进行用户标志模块（subscriber identify module，SIM）交换攻击（身份盗窃攻击）等。欧洲刑警组织曾警告，ChatGPT 可能被滥用于网络钓鱼、虚假信息和网络犯罪。

- 生成缺陷代码难以追查。软件工程师在项目开发过程中，可能会利用 ChatGPT 快速生成特定功能代码。然而，根据 OpenAI 的评估，Codex 只有 37%的概率给出正确代码。除了存在无法运行的 bug，基于 AI 编写的代码还可能会引入漏洞。Pearce 等[29]通过研究 89 个场景中生成的代码，发现 GitHub Copilot 给出的结果中 40%存在漏洞。

3.4 AIGC 生成偏见内容导致个人、企业名誉受损

模型开发者很容易将自身偏好、偏见、价值观带入模型中。企业或个人在使用 AIGC 技术能力对外提供服务时若出现偏见内容，将会导致名誉受损。例如，Replika 最初的产品定位为“关心人类的 AI 朋友”，可提供陪聊服务，但之后多次爆出性骚扰用户事件。此外，当将 AIGC 技术应用于教育行业时，若对内容审查不严，偏见内容会严重影响学生的价值观。

3.5 利用 AIGC 欺骗智能识别系统

合成伪造生物识别信息。AIGC 技术能够读取并模仿生成生物识别信息，伪造身份以欺骗身份验证，大幅增大了智能识别系统的入侵风险。例如，益博睿的一份报告就概述了企业面临的合成身份欺诈威胁，即网络犯罪分子使用深度伪造的面孔欺骗生物识别验证，这已经被确定为增长最快的金融犯罪类型。这将不可避免地给依赖面部识别软件作为其身份和访问管理策略一部分的企业带来重大挑战。

3.6 使用 AIGC 服务导致核心数据泄露

根据 OpenAI 官网公布的隐私政策，其并未提及类似欧盟《通用数据保护条例》（General Data Protection Regulation，GDPR）等数据保护法规，在“使用数据”条款里，OpenAI 承认会收集用户使用服务时输入的数据，但未对数据的用途做进一步说明。企业员工在使用 ChatGPT 服务完成工作任务时很容易将企业核心数据资料发送给 ChatGPT，从而导致企业核心数据泄露，造成无法挽回的损失。例如，微软和亚马逊已禁止员工向 ChatGPT 分享敏感数据，以防后者的输出包含或出现类似公司机密信息。此外，意大利数据保护局表示，ChatGPT 涉嫌违法收集个人数据且没有建立年龄验证机制，即日起暂时禁止使用，成为首个禁用 ChatGPT 的国家。

3.7 AIGC 训练不当容易导致隐私数据泄露

一方面，如果在训练 AIGC 模型的过程中使用了未脱敏的数据，使用 AIGC 服务的恶意用户可以采用特定的交互模式获取这些隐私信息，从而导致个人隐私数据泄露。例如，Facebook 曾因未经个人同意使用公开图片集进行算法训练，违反了《生物识别信息隐私法》（Biometric Information Privacy Act，BIPA），最终赔偿 6.5 亿美元，微软、亚马逊、谷歌同样曾因此陷入 BIPA 诉讼中。

另一方面，AIGC 模型在垂直行业的落地应用离不开行业数据精调训练。这涉及基础大模型服务提供商与垂直行业企业的联合。但行业数据对于企业具有较高的商业价值，一旦对外泄露会导致严重经济损失。这严重阻碍了 AIGC 技术赋能垂直行业。

4 应对策略

4.1 技术层面：加强 AIGC 相关技术研究

（1）利用数字水印技术实现 AIGC 合成内容追踪溯源

针对人工智能撰写的文章、生成的图片等

AIGC 作品存在著作权归属不清和知识产权剽窃等一系列风险，可以跟进研究合成内容的标记算法——数字水印技术。数字水印技术[30]是一项保护版权的信息隐藏技术，可以有效避免数字产品在传播过程中遭到篡改与非法利用。文本、图像、音频、视频等均可作为数字水印的载体。在自然语言方面，已经有研究人员提出一种为大规模语言模型添加水印的方法[31]，该方法在不影响大模型生成文本质量的基础上实现了水印的添加。验证者可使用算法对文本中的数字水印进行验证，从而快速识别机器生成的文本内容。对于图像数据[32]，可以从训练数据和模型层面将水印信息嵌入图像中，使人眼难以辨别载体图像与含水印图像之间的差异，在不影响图像视觉效果的情况下，实现水印信息的可靠传输，并在传输过程中遭遇失真或一定程度的攻击后仍能完整地提取出水印，可被有效地应用于泄密追踪、版权认证、防伪溯源等。

（2）使用 AIGC 技术进行诈骗内容识别和解释

对于利用 AIGC 技术生成的诈骗信息，同样可以使用 AIGC 技术进行诈骗内容的识别和解释。例如，可以利用 ChatGPT 进行诈骗短信的判定，以及对判定结果进行解释分析。尽管 ChatGPT 等 AIGC 相关模型拥有各种强大的能力，但受限于参数量大、训练周期长和未开源等因素，无法完成本地化部署使用。针对此问题，可以利用 AIGC 模型产生高质量数据，在轻量级开源模型上做微调。例如，模型 Baize[33]通过利用 ChatGPT 与自身进行对话，自动生成一个高质量的多轮聊天语料库，并在一个开源的大型语言模型 LLaMA 上进行参数微调，在多轮对话上表现出良好的性能，最大限度地减少潜在的风险。

（3）AI 模型加固，防止 AIGC 合成生物信息攻击

为了避免 AIGC 生成的虚假数据攻击 AI 生物识别系统，导致系统的误判、失效和瘫痪，需要采用深度学习的模型加固等技术提高系统识别准确率和对抗攻击的能力。例如，可以研究使用特征凝结、空间平滑、高斯数据增强和对抗训练等方法增强 AI 模型识别对抗样本的能力，以防御 AIGC 生成的对抗样本对 AI 模型的攻击。

（4）利用 AI 技术鉴别虚假合成内容

一方面，由于 Deepfake 等合成图像和语音技术的日益普及和滥用，需要研究人员构建强大的伪造内容检测解决方案。对于图像数据，进行像素级纹理识别，分析图像中的噪声信息和不正常的轮廓和边缘，判断图像是否属于合成伪造；对于音频数据，进行时域、频域、倒频域和其他域的信号分析，利用深度学习技术进行声纹特征提取和对比。近期，哈尔滨工业大学和南洋理工大学提出全球首个多模态 Deepfake 检测定位模型[34]，不仅能够判断输入图像–文本对的真假，也尝试定位篡改内容（如图像篡改区域和文本篡改单词），让 AIGC 伪造无处可藏。

另一方面，AIGC 的广泛应用可能会产生大量假新闻和谣言，这对 AI 检测技术带来了巨大的挑战。当前，利用 AI 鉴别假新闻和谣言已经有了一定的研究基础。例如，麻省理工学院的计算机科学与人工智能实验室（Computer Science and Artificial Intelligence Laboratory，CSAIL）提出一种可以鉴别信息来源准确性和个人政治偏见的 AI 系统；复旦大学团队提出了 LOREN[35]，一种全新的可解释事实检验范式。在多模态方面，SAFE[36]利用多模态之间的对比分析，依据新闻的文本信息和视觉信息的匹配度识别虚假新闻。此外，大量研究工作[37]结合知识图谱等外部知识来辅助识别虚假信息。外部知识含有丰富的语义信息和客观事实，可以帮助模型更好地理解和对比分析新闻内容，从而识别出虚假新闻中的造假之处。

（5）违规内容检测，防止传播 AIGC 生成的违规信息

一是文本合规检测[38]，利用大型语言模型判

断文本内容是否包含违规信息；二是图片/视频合规检测，通过视觉模型提取图片特征，判断图片包含的信息是否合规；三是音频合规检测，利用深度学习算法分析音频中是否包含敏感词汇、受版权保护的内容，以及是否违反国家相关规定。

（6）隐私计算避免 AIGC 训练数据的隐私泄露

针对多方数据合作训练 AIGC 模型时，可能存在的数据泄露风险，可以进一步加强研究联邦机器学习。通过设计高性能的隐私计算方法，能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下，进行数据使用和 AIGC 模型训练。

（7）AI 反偏见技术，助力 AIGC 公平性研究

为了避免 AIGC 在实际应用场景中产生偏见信息，未来需要进一步加强 AI 模型反偏见技术研究。人工智能的“黑箱”特性使得性别或种族等偏见与更多更复杂的参数相勾连，因此很难通过直接删除或屏蔽模型参数来完成偏见的剔除。算法偏见的根源来自数据，不公正的数据集是偏见的土壤。因此，构建更加公正的数据集无疑是算法偏见根本性的解决方法之一。此外，针对训练好的模型，利用技术手段侦测偏见、解除偏见，也是 AI 反偏见研究的重点内容。例如，哥伦比亚大学的研究者开发了一款名为 DeepXplore 的软件，它可以通过“哄骗”系统犯错，暴露算法神经网络中的缺陷；谷歌推出工具 What-If，其是 TensorBoard 中用于检测偏见的工具；IBM 也将其偏见检测工具 AI Fairness 360 工具包开源，其中包括超过 30 个公平性指标和 9 个偏差缓解算法。从目前的成果来看，大多技术突破还仅处于初级阶段，即检测偏见，消除偏见方面的研究仍亟须进一步努力。

4.2 应用层面：构建中间层，完善 AIGC 应用部署策略

在部署 AIGC 相关应用时，通过在应用和模型之间构建双向中间层，完善部署策略，可以有效缓解 AIGC 技术带来的不良影响，如图 2 所示。一是过滤层，利用审核分类器在监控和执行管道中过滤掉 AIGC 模型输出的有害内容。例如，当 AIGC 模型输出涉恐等敏感数据时，过滤层对内容进行判定，若触发了过滤策略，则不向应用层输出模型原始的输出内容，而采用相关提示进行代替。二是防护层，对操作者输入模型的内容做审核，防止“越狱”。例如，当应用层传来的用户输入包含恶意内容时，触发防护层的检测，若判定结果是恶意的，则此输入被拦截，不会送入模型进行推理，并向应用层反馈判定信息。

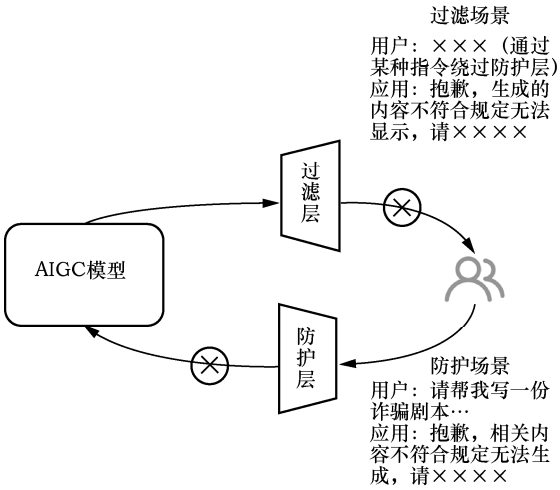

图 2 AIGC 部署策略：过滤层和防护层

4.3 监管层面：不断健全法律法规，为 AIGC 技术健康发展保驾护航

AIGC 技术的健康发展，离不开国家层面及时有效地出台相应的法律法规。自 2019 年 11 月起，我国先后出台《网络音视频信息服务管理规定》《网络信息内容生态治理规定》《互联网信息服务算法推荐管理规定》等文件，对生成合成类内容提出不同程度的监管要求。2022 年 12 月 11 日，国家互联网信息办公室、工业和信息化部、公安部联合发布《互联网信息服务深度合成管理规定》强调不得利用深度合成服务从事法律、行政法规禁止的活动，要求深度合成服务提供者落实信息安全主体责任。此外，国家互联网信息办公室网站 2023 年 4 月 11 日消息，为促进生成式人工智能技术健康发展和规范应用，根据《中华

人民共和国网络安全法》等法律法规，国家互联网信息办公室起草了《生成式人工智能服务管理办法（征求意见稿）》，并向社会公开征求意见。

5　结束语

本文介绍了 AIGC 技术的发展现状，分析了生成式人工智能相比分析式人工智能所具有的优势。针对不同的生成数据格式，分别总结了 AIGC 丰富的技术能力，并提供相应技术案例。但是，任何技术都是一把双刃剑，本文详细列举了 AIGC 技术的飞速发展，以及给内容安全治理带来的诸多风险挑战。最后，为了实现 AIGC 技术健康可持续地发展，分别从技术、应用和法律层面，总结分析了未来的工作建议与应对策略。

参考文献：

[1] 李白杨, 白云, 詹希旎, 等. 人工智能生成内容(AIGC)的技术特征与形态演进[J]. 图书情报知识, 2023(1): 66-74.
LI B Y, BAI Y, ZHAN X N, et al. The technical features and aromorphosis of artificial intelligence generated content (AIGC)[J]. Documentation, Information & Knowledge, 2023(1): 66-74.

[2] CAO Y, LI S, LIU Y, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT[J]. arXiv preprint, 2023, arXiv: 2303.04226.

[3] WU J, GAN W, CHEN Z, et al. AI-generated content (AIGC): a survey[J]. arXiv preprint, 2023, arXiv: 2304.06632.

[4] ADAMOPOULOU E, MOUSSIADES L. An overview of Chatbot technology[C]//IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham: Springer, 2020: 373-383.

[5] TOUSEEF I, QURESHI S. The survey: text generation models in deep learning[J]. Journal of King Saud University - Computer and Information Sciences, 2022, 34(6): 2515-2528.

[6] 詹希旎, 李白杨, 孙建军. 数智融合环境下 AIGC 的场景化应用与发展机遇[J]. 图书情报知识, 2023(1): 75-85, 55.
ZHAN X N, LI B Y, SUN J J. Application scenarios and development opportunities of AIGC in the digital intelligence integration environment[J]. Documentation, Information & Knowledge, 2023(1): 75-85, 55.

[7] ZHOU X Y, ZAFARANI R. A survey of fake news[J]. ACM Computing Surveys, 2021, 53(5): 1-40.

[8] 喻国明, 苏健威. 生成式人工智能浪潮下的传播革命与媒介生态[J]. 新疆师范大学学报(哲学社会科学版), 2023, 44(5): 65-73.
YU G M, SU J W. Communication revolution and media ecology under the wave of generative artificial intelligence[J]. Journal of Xinjiang Normal University (Edition of Philosophy and Social Sciences), 2023, 44(5): 65-73.

[9] KATZ D M, BOMMARITO M J, GAO S, et al. GPT-4 passes the bar exam[J]. SSRN Electronic Journal, 2023.

[10] DONG Q, LI L, DAI D, et al. A survey on in-context learning[J]. arXiv preprint, 2022, arXiv: 2301.00234.

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[12] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8107-8116.

[13] CLARK A, DONAHUE J, SIMONYAN K. Adversarial video generation on complex datasets[J]. arXiv preprint, 2019, arXiv: 1907.06571.

[14] CROITORU F A, HONDRU V, IONESCU R T, et al. Diffusion models in vision: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10850-10869.

[15] 马化腾. 从专用人工智能迈向通用人工智能[J]. 中国科技产业, 2019(9): 9.
MA H T. From special artificial intelligence to general artificial intelligence[J]. Science & Technology Industry of China, 2019(9): 9.

[16] LU J, BATRA D, PARIKH D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. arXiv preprint, 2019, arXiv: 1908.02265.

[17] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[J]. arXiv preprint, 2021, arXiv: 2103.00020.

[18] LI W, GAO C, NIU G C, et al. UNIMO-2: end-to-end unified vision-language grounded learning[C]//Proceedings of Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 3187-3201.

[19] SINGH A, HU R H, GOSWAMI V, et al. FLAVA: a foundational language and vision alignment model[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15617-15629.

[20] FENG Z D, ZHANG Z Y, YU X T, et al. ERNIE-ViLG 2.0: improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 10135-10145.

[21] CHEN J, GUO H, YI K, et al. VisualGPT: data-efficient adaptation of pretrained language models for image captioning[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 18009-18019.

[22] ZHANG T. Deepfake generation and detection, a survey[J].

Multimedia Tools and Applications, 2022, 81(5): 6259-6276.
[23] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[J]. arXiv preprint, 2022, arXiv: 2204.06125.
[24] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[J]. arXiv preprint, 2023, arXiv: 2302.05543.
[25] ESSER P, CHIU J, ATIGHEHCHIAN P, et al. Structure and content-guided video synthesis with diffusion models[J]. arXiv preprint, 2023, arXiv: 2302.03011.
[26] SINGER U, POLYAK A, HAYES T, et al. Make-a-video: text-to-video generation without text-video data[J]. arXiv preprint, 2022, arXiv: 2209.14792.
[27] STUPP C. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case[J]. The Wall Street Journal, 2019, 30(8).
[28] 孙山. 人工智能生成内容著作权法保护的困境与出路[J]. 知识产权, 2018, 28(11): 60-65.
SUN S. Predicament and outlet of copyright law protection of artificial intelligence-generated content[J]. Intellectual Property, 2018, 28(11): 60-65.
[29] PEARCE H, AHMAD B, TAN B, et al. Asleep at the keyboard? assessing the security of GitHub copilot's code contributions[C]//Proceedings of 2022 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2022: 754-768.
[30] 向德生, 杨格兰, 熊岳山. 数字水印技术研究[J]. 计算机工程与设计, 2005, 26(2): 326-328, 334.
XIANG D S, YANG G L, XIONG Y S. Survey of digital watermarking[J]. Computer Engineering and Design, 2005, 26(2): 326-328, 334.
[31] KIRCHENBAUER J, GEIPING J, WEN Y, et al. A watermark for large language models[J]. arXiv preprint, 2023, arXiv: 2301.10226.
[32] 吴德阳, 张金羽, 容武艳, 等. 数字图像水印技术综述[J]. 高技术通讯, 2021, 31(2): 148-162.
WU D Y, ZHANG J Y, RONG W Y, et al. Survey of digital image watermarking technology[J]. Chinese High Technology Letters, 2021, 31(2): 148-162.
[33] XU C, GUO D, DUAN N, et al. Baize: an open-source chat model with parameter-efficient tuning on self-chat data[J]. arXiv preprint, 2023, arXiv: 2304.01196.
[34] SHAO R, WU T, LIU Z. Detecting and grounding multi-modal media manipulation[J]. arXiv preprint, 2023, arXiv: 2304.02556.
[35] CHEN J J, BAO Q B, SUN C Z, et al. LOREN: logic-regularized reasoning for interpretable fact verification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10482-10491.
[36] ZHOU X Y, WU J D, ZAFARANI R. SAFE: similarity-aware multi-modal fake news detection[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2020: 354-367.
[37] MRIDHA M F, KEYA A J, HAMID M A, et al. A comprehensive review on fake news detection with deep learning[J]. IEEE Access, 2021(9): 156151-156170.
[38] 蔡鑫. 基于 Bert 模型的互联网不良信息检测[J]. 电信科学, 2020, 36(11): 121-126.
CAI X. Internet bad information detection based on Bert model[J]. Telecommunications Science, 2020, 36(11): 121-126.

[作者简介]

乔喆（1981− ），男，中国移动通信集团有限公司信息安全管理与运行中心策略运营处处长、经济师，主要研究方向为网络信息安全。