编者按　当前，全球人工智能正在加速向新发展阶段迈进，以大模型和生成式人工智能为代表的通用人工智能取得突破性进展。作为引领新一轮科技革命和产业变革的重要驱动力，人工智能对经济社会发展、国际政治格局等方面均已产生重大而深远的影响。党的二十届三中全会审议通过的《中共中央关于进一步全面深化改革 推进中国式现代化的决定》提出，“完善生成式人工智能发展和管理机制”“建立人工智能安全监管制度”。统筹人工智能发展和安全，全面提升人工智能安全治理水平，既是规范引导新兴技术应用和产业健康有序发展的内在需要，也是筑牢国家安全屏障，推进国家治理体系和治理能力现代化的重要议题。为深入研讨如何应对人工智能风险、把握发展战略主动、有效维护和保障公共安全，加强人工智能发展的潜在风险研判与防范，《中国科学院院刊》特策划出版“人工智能与公共安全”专题，邀请相关权威专家学者系统阐述推动人工智能发展安全、可靠、可控的理论与实践，为提升人工智能安全治理注入更多思想力量，共同促进全球人工智能有序安全发展。专题由中国工程院院士、中国科学院计算技术研究所研究员、《中国科学院院刊》副主编李国杰指导推进。

引用格式：曹娟, 盛强, 李国杰. 智能时代公共安全体系面临的技术挑战——兼序《中国科学院院刊》“人工智能与公共安全”专题. 中国科学院院刊, 2025, 40(3): 399-407, doi: 10.16418/j.issn.1000-3045.20241027003.
Cao J, Sheng Q, Li G J. Challenges on public security system in AI era—Preface for special column “Artificial Intelligence and Public Security” in *Bulletin of Chinese Academy of Sciences*. Bulletin of Chinese Academy of Sciences, 2025, 40(3): 399-407, doi: 10.16418/j. issn. 1000-3045.20241027003. (in Chinese)

智能时代公共安全体系面临的技术挑战

——兼序《中国科学院院刊》“人工智能与公共安全”专题

曹 娟[1,2*]　盛 强[1]　李国杰[1]

1 中国科学院计算技术研究所　北京　100190
2 中国科学院大学　计算机科学与技术学院　北京　100049

摘要　人工智能生成内容（AIGC）技术快速发展，引发新的公共安全隐患，严重威胁国家安全和社会稳定。文章梳理了AIGC技术和检测技术的发展现状，指出了检测技术在实战场景中面临的挑战，提出发展面向公共安全的AIGC检测技术，形成从模型端到平台端的全流程检测技术体系，实现“生成时可赋标、传播中可鉴别、案发后可溯源”。

关键词　人工智能生成内容，生成内容检测，公共安全，全流程治理

DOI　10.16418/j.issn.1000-3045.20241027003
CSTR　32128.14.CASbulletin.20241027003

*通信作者
修改稿收到日期：2025年2月22日

1 AIGC技术对公共安全带来新挑战

人工智能生成内容（AIGC）技术是指基于生成式人工智能算法和模型创作文本、图像、声音、视频、代码等内容的技术。作为近年来人工智能领域的最大突破之一，AIGC方向不断涌现里程碑式突破，人工智能模型由理解判别走向生成创造。以GPT-3.5为代表的语言模型、以Stable Diffusion为代表的文生图模型和以Sora为代表的文生视频模型分别突破通用化文本、图像和视频生成的难关，内容效果愈发逼真，制作成本逐渐降低，可用范围不断扩展。据预测，我国AIGC市场在2030年将达到万亿元规模[1]，人工智能合成数据将成为新增人工智能训练数据的主要来源①。

然而，AIGC技术与应用的蓬勃发展也带来了新的公共安全隐患，严重威胁国家安全和社会稳定。最新AIGC技术的特点可以大致总结为“逼真度高、创作效率高、通用性高”，这种质量、数量、适用范围的全面提升，导致人类和传统技术很难立刻分辨真实来源内容和AIGC。近年来，基于AIGC的违法犯罪行为越来越多，AIGC技术在不断降低传统违法犯罪成本的同时，也催生了新型违法犯罪活动的快速涌现，不断撼动现有社会信任体系，公共安全治理面临更严峻的挑战：① 利用AIGC技术生成虚假信息、操作舆论，是世界各国面临的国家安全难题；② 利用AIGC技术进行身份伪造、学术造假、黑产牟利，是各行各业面临的安全发展难题；③ 利用AIGC技术进行电信诈骗、隐私侵犯，是困恼每个公民的个人安全难题。

AIGC安全治理已进入从高层共识到全民共识的“深水区”、从立法到执法的“深水区”、从探讨危害到实际部署能力的“深水区”。《中国科学院院刊》2025年第3期“人工智能与公共安全”专题，邀请科研和实战一线的领军人物论述智能时代公共安全面临的各方面挑战及其应对策略，为智能时代公共安全体系的重塑提供深度思考和解决方案。受限于篇幅，专题文稿主要关注智能时代对公共安全的技术挑战、业务挑战、算法治理挑战、重要应用挑战4个方面。

（1）技术挑战。随着AIGC技术快速发展，生成内容越来越逼真，肉眼很难分辨，需要依赖技术手段进行检测。面对层出不穷的AIGC新技术和应用，如何构建对新模型可扩展、可溯源的检测技术体系，支撑公共安全治理？本文将介绍生成技术和检测技术的重要进展，梳理当前AIGC检测面临的挑战，提出面向实战场景的应对建议。

（2）业务挑战。AIGC技术的颠覆性与快速迭代性，使未来技术发展可能导致的风险具有高度不确定性，极易引发各类新型犯罪。而现有的法律规制与监管执法手段仍存在漏洞，为犯罪打击带来严峻挑战。北京市公安局高建新副局长等将介绍人工智能犯罪的类型、态势、特点，并针对人工智能犯罪治理现状与挑战提出对策建议。

（3）算法治理挑战。在人工智能时代，算法作为一种新的生产工具，在各种系统服务中扮演着比以往更重要的角色，在推荐系统等场景甚至已经成为人类决策的替代。由于生成式人工智能算法普遍不具有可解释性，在应用场景中会给公共安全带来未知的风险和挑战。中国科学院计算技术研究所程学旗研究员等将聚焦智能算法安全的内涵与科学问题，促进智能算法可信、可管、可控，形成智能算法治理的长效机制。

（4）重要应用挑战。人工智能作为新质生产力，应用场景丰富，发展潜力巨大，各个领域已经开始探索智能化系统的落地应用，但其中的潜在风险特别是

① Is synthetic data the future of AI?. (2022-06-22)[2025-02-11]. https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai.

特定应用场景独有的安全风险仍不容忽视。浙江大学徐文渊教授等将从信息域、物理域、社会域视角出发，探讨具身智能的安全内涵与安全体系，提出具身智能的安全防护体系和综合治理措施。

2 人工智能内容生成技术发展迅速，但安全性问题突出

2.1 人工智能内容生成技术概述

AIGC的质量迅速提升，曾经困扰研究者多年的语句不通顺、视频不连贯、语音不自然等生成瑕疵基本已不存在。经过预训练的大语言模型依赖少量的提示语即可完成各类文字任务；视觉模型仅需要1张照片即可完成换脸任务，5—10张不同角度的照片即可微调实现实时人脸替换；若获取10—20秒的含人脸、声音的视频，即可基于音视频生成技术得到该人物的“数字人”。然而，与生成能力不匹配的是人类仍然缺乏自主辨识AIGC的能力。一项4 600人参与的实验显示，人类还无法凭借自身总结的经验分辨人类和人工智能生成文本[2]；类似的结论也在基于视觉[3]和人声[4]内容的独立实验上分别得到验证。这意味着人工智能生成技术一旦被恶意利用，多数人将无法借助自身知识避免受骗。以下将从AIGC技术主要包含的文本生成、视觉生成和音频生成技术3个方面说明。

（1）文本生成。以GPT系列为代表的通用对话式文本生成大模型主要依赖于关键结构（Transformer网络）、大数据（互联网级语料）和大算力（万级图形处理器训练）3个要素。Transformer是一类基于注意力机制的神经网络结构，其根据各个字词的相关性分配不同权重，能够更好地处理长期依赖关系，具有高度可并行性，非常适合大规模训练。在过去的5年内，语言模型的参数量从亿级（GPT-1）猛涨到了千亿甚至万亿级（GPT-3及后续版本），参数量的增长也带来了惊艳的效果。目前，文本生成模型的总体建模思路暂时趋于稳定，研究者已将更多精力放在对当前模型训练与应用模式的改进和扩展上，具体可分为4个方面：① 交互体验方面，北京月之暗面科技有限公司的Kimi等模型注重长上下文扩展，输入输出窗口最长可达百万级词元（token），可以在短时间内从大量资料中定位所需信息；② 智能提升方面，美国人工智能公司OpenAI提出“超级对齐”（Super Alignment），颠覆现有的“强对弱”对齐模式（如人类对语言模型），期望实现“弱对强”的监督，最终目标是实现“超人智能”；③ 安全输出方面，美国人工智能初创公司Anthropic提出基于人工智能反馈的强化学习框架（RLAIF），通过少量的自然语言准则或指令降低模型输出的有害性；④ 高效训练部署方面，杭州深度求索人工智能基础技术研究有限公司的DeepSeek系列模型关注模型架构效率提升，其V3模型（6 710亿参数）训练所需机时仅为美国Meta公司Llama 3模型（4 050亿参数）的9.1%；北京面壁智能科技有限责任公司的MiniCPM和美国微软公司的Phi等模型关注边缘侧应用，推出的十亿级参数模型可在智能终端本地运行。

（2）视觉生成。早期的图像和视频生成主要依赖生成对抗网络（GAN），通过生成器和判别器的对抗训练来提高视觉内容质量，但其稳定性一直不高。近年来，基于概率的无监督式生成模型（扩散模型）越来越引人关注，其设计灵感来自于非平衡热力学，模仿扩散过程对图像不断加噪以将其转变为近似噪声的隐编码，然后模型学习逆转加噪的过程，从图像相同尺寸的噪声中不断去噪以还原原始图像。扩散模型的训练相对简单且稳定，比传统的GAN更容易实现。同时，扩散生成模型的表示能力非常强，其加噪去噪过程的设计适合完成图像到图像的转换任务（如图像修复、图像超分辨率、图像风格转换），也适用于表情修改、风格化等编辑任务。更重要的是，扩散模型不容易出现GAN训练中常见的梯度消失和梯度爆炸问题，更适用于作为视觉生成大模型的基础结构，因此成为了近期推出的Flux、Sora等视觉大模型的主要

选择。

（3）音频生成。音频生成主要包括人声生成、环境音合成、音乐生成等任务。与文本生成类似，音频生成大模型也采用了序列建模的框架，音频信号首先通过编码器离散化为音频“字符”，之后输入基于Transformer的模型进行训练。例如：①在音乐生成方面，美国互联网公司Meta推出的AudioCraft工具可以实现输入文本指令，生成指定风格的音乐和音效；英国人工智能初创公司Suno AI推出的Suno V3可以一次性制作带有人声和背景旋律的“广播级”音频；②在人声生成方面，美国人工智能公司OpenAI发布的Voice Engine、阿里巴巴通义实验室发布的CosyVoice等模型仅基于十几秒内的原始音频，即可生成模拟音色、韵律、情感色彩的音频，甚至实现跨语种生成，互联网上广为流传的“AI郭德纲”等视频中的声音合成多采用这类技术。

2.2 人工智能生成技术滥用情况

尽管在生成质量取得了突破，催生了一大批基于AIGC的应用产品，但现有AIGC技术的安全问题仍然十分突出，并且在模型本身安全围栏不牢固与不法分子恶意利用的双重作用下，已开始造成诸多现实危害。

（1）基于AIGC技术批量生成虚假信息，危害国家安全和社会稳定。借助AIGC技术，造假者可以基于热点新闻素材大批量伪造低质假消息，成本进一步降低，传播隐蔽性更强，随时可能引发舆论争议，在政治选举等关键事件中误导民众。2023年9月，一段关于候选人操纵选举的人工智能伪造录音流出，对斯洛伐克议会选举产生了颠覆性的影响[②]；2024年1月，美国新罕布什尔州部分选民接到了“AI拜登”的语音电话，试图阻止他们参与民主党初选[③]。根据调研机构NewsGuard报告，截至2025年2月，全球已出现依赖人工智能生成新闻资讯的低质网站1 254家，涵盖汉语、英语、法语等16种语言[④]；大语言模型仍存在幻觉问题，报告显示DeepSeek-R1的幻觉率高达14.3%[⑤]，在开源模式下其被私有部署用于生成虚假信息的风险可能进一步扩大；世界经济论坛发布的《2025全球风险报告》指出，利用人工智能生成的错误和虚假信息是近2年最大的全球性风险[⑥]。

（2）基于AIGC技术换脸变声的新型诈骗，危害个人安全。随着人工智能换脸、拟声技术的发展，诈骗者只需要获取一张照片、一小段语音，就可以实现低成本的实时换脸变声，实现在线会议、视频通话场景下的长时间稳定身份替换，令普通民众防不胜防。据奇安信监测，基于人工智能的伪造欺诈在2023年暴增3 000%[⑦]；迈克菲一项全球7 000余人参与的调研显示，10%受访者曾经历人工智能语音诈骗[⑧]。不仅如此，基于AIGC的新型诈骗单笔涉案金额越来越大，2024年2月，香港警方披露了一起冒充跨国公司首席

② Slovakia’s election deepfakes show AI is a danger to democracy. (2023-10-03)[2025-02-11]. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/.

③ Fake Biden robocall tells voters to skip New Hampshire primary election. [2025-02-11]. https://www.bbc.com/news/world-us-canada-68064247.

④ Tracking AI-enabled misinformation. [2025-02-11]. https://www.newsguardtech.com/special-reports/ai-tracking-center/.

⑤ DeepSeek-R1 hallucinates more than DeepSeek-V3. (2025-01-30)[2025-02-11]. https://www.vectara.com/blog/deepseek-r1-hallucinates-more-than-deepseek-v3.

⑥ Global risks report 2025. (2025-01-15)[2025-02-11]. https://www.weforum.org/publications/global-risks-report-2025/.

⑦ 2024人工智能安全报告. (2024-02-29)[2025-02-11]. https://www.qianxin.com/threat/reportdetail?report_id=311.

⑧ Beware the artificial impostor. [2025-02-11]. https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf.

财务官的AIGC诈骗案件，涉案金额高达2亿港元⑨。

（3）基于AIGC技术生成私人性内容图像，侵犯个人隐私和名誉。随着人工智能算力基础设施日益完善和人工智能应用服务模式不断创新，AIGC能力的获取门槛已显著降低，非专业人士也可以通过个人终端设备轻松生成指定内容，易被抱有不良目的的人利用。2024年8月，韩国爆出AIGC版的“N号房”事件⑩，通信软件Telegram上出现大量聊天群分享和传播人工智能伪造的性内容图像，对象涉及学生、教师、医护等特定职业群体，严重侵害受害者隐私和名誉；受害者遍布500多所学校，规模之大令人震惊。这些内容并非出自少数职业团伙之手，而是由普通民众恶意利用公开AIGC工具制作，已知的加害者中甚至有相当一部分还是在校未成年人。

3 AIGC检测技术是应对AIGC滥用的关键

3.1 AIGC检测技术概述

AIGC检测技术是用于分辨各类AIGC与人类书写、摄录内容的技术的总称，在实际应用中已经取得了一定的成效：① 在互联网流量监管中，检测技术被用于违规内容筛查，支撑公安机关破获多起人工智能伪造相关案件，服务重大任务安保；② 在重大事件舆情监测中，检测技术被用于识别虚假信息，支撑快速形成重大事件虚假内容专题报告；③ 在金融服务中，检测技术被用于防范基于人工智能技术的身份冒充，已成为银行等金融机构交易鉴权环节的必备模块。以下将从AIGC检测技术主要包括的生成文本检测技术、生成图像视频检测技术、生成音频检测技术和生成模型溯源技术4个方面说明。

（1）生成文本检测技术。生成文本检测模型用于区分人工撰写和人工智能模型生成的文本，主要包括基于生成概率和基于风格特征的检测方法。① 基于生成概率的检测方法。此类方法认为大语言模型的预训练和生成采样过程塑造了独特的用词偏好和用词稳定性。例如，人工智能生成的论文审稿意见中“commendable”一词出现的频次明显高于人类审稿意见[5]；人工智能生成文本的写作结构相对于人类而言更加稳定。在ChatGPT问世不久后引发关注的产品GPTZero就利用了这些性质，构建了基于语言模型困惑度（perplexity）和突发性（burstiness）的检测模型。斯坦福大学学者提出的DetectGPT[6]延伸了这一思路，通过扰动生成采样过程，观察当前用词是否遵循了“选择概率最高”的人工智能采样规则作为区分人类和人工智能生成文本的信号。不过由于模型特性仍存在差异，上述模型一般只适用于已知特定模型生成的文本。② 基于风格特征的检测方法。此类方法主要依赖语言学分析和神经网络特征学习，从词汇多样性、连贯性、重复性等文体学特征以及事实要素篇章一致性等文字结构的相关特征区分人类和人工智能生成文本，但这类方法的检测灵敏度正因生成质量的提高和检索增强生成等辅助技术的应用而逐渐降低，其全面性和灵活性明显受限于先验知识。

（2）生成图像视频检测技术。生成图像视频检测的设定与文本类似，一部分检测方法利用自然摄录内容概念的先验性质，另一部分注重挖掘生成与编辑过程的特性。① 基于先验性质的检测方法。此类方法认为生成的视觉内容无法完美复现真实世界中视觉语义概念特性，因此观察概念呈现的合理性更容易发现

⑨ ‘Everyone looked real’: Multinational firm’s Hong Kong office loses HK$200 million after scammers stage deepfake video meeting. [2025-02-11]. https://www.scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage.

⑩ South Korea battles surge of deepfake pornography after thousands found to be spreading images. [2025-02-11]. https://www.theguardian.com/world/article/2024/aug/28/south-korea-deepfake-porn-law-crackdown.

AIGC的细微瑕疵。例如，早年人工智能换脸视频经常出现眨眼频次不合理、不生成说话人牙齿、肤色过渡不自然等生理信号瑕疵；还有一些方法利用Xception等预训练视觉模型中蕴含的自然图像先验，通过微调的方法将通用视觉理解模型转化为生成内容检测模型，但生成内容逼真度的大幅提高正在不断缩小先验上的差异。② 从生成和编辑过程提取特征的检测方法。此类方法关注频域统计特性、压缩特性方面的差异。例如，有研究发现生成模型的上采样模块可能在生成图像中留下稳定的隐藏痕迹及纹理信息，因此可以通过提取隐藏痕迹用于检测[7]；对于利用人工智能修图产生的区域编辑图像[8]，还可以通过比较生成区域和原图区域在像素排列逻辑、光学噪声、重压缩痕迹实现更精细的区域定位。

（3）生成音频检测技术。生成音频检测可进一步分为全局生成检测和生成片段定位2个任务。其核心是通过考虑语音信号、声纹特征和频谱分布等特征进行鉴别。① 全局生成检测。此任务的基础特征包括原始波形和功率谱、幅度谱、相位等频谱特征。对于特定人的生成语音检测，还会提取与说话人身份有关的特征。近年来，大规模自监督预训练模型HuBERT的输出也成为检测模型采用的特征，其泛化性高于传统特征。② 生成片段定位。此任务用于应对语音篡改行为，更加注重建模帧级别特征，通过侦测真假语音的波形边界识别被替换为生成语音的片段。

（4）生成模型溯源技术。生成模型溯源的目的是从内容识别其来源模型，其基本假设与生成内容检测类似，都是认为生成内容中蕴含着某种具有模型特异性的特征。不同的是，溯源技术关注如何区分不同的AIGC模型。溯源方法根据是否可以获得模型内部信息，分为白盒方法、黑盒方法和灰盒方法。① 白盒方法。采用白盒设置的溯源方法通过获取给定内容在候选模型上推理的统计指标（如文本词频分布）作为特征，衡量模型对内容的“熟悉度”以判断来源。② 黑盒方法。采用黑盒设置的方法主要以数据驱动的思路构建溯源模型，通过挖掘同源生成内容的共性获得其中只与来源模型有关的特征实现溯源，提取出的特征也被称为“模型指纹”[7]。③ 灰盒方法。针对白盒方法无法用于闭源AIGC大模型的问题，近期学者开始研究灰盒溯源方法，即使用内部信息更方便获取的开源大模型作为代理估计闭源大模型特性，再利用白盒方法的思路做出判断，在生成文本溯源任务上取得了介于黑盒和白盒方法之间的溯源效果[9]；由于多数现有方法只能追溯到训练阶段已知生成模型，无法识别未知生成模型，近期一些学者也开始探索将未知模型归入“其他”类的开集模型溯源[10]和支持新生成模型发现的零样本模型溯源技术[11]。

3.2 实战场景AIGC检测技术面临的挑战

尽管目前AIGC检测技术和工具都已具备，但面对大模型应用的快速大规模普及，生成与检测的持续对抗仍在升级。未来的AIGC监管实战将面临3项关键挑战。

（1）如何提升检测模型针对新出现AIGC模型的泛化能力。AIGC技术迭代更新很快，生成质量的提升、模态的扩展、技术方案的升级、从闭源到开源生态的构建，往往是在几个月内完成的。例如，美国人工智能公司OpenAI在2024年5月展示了可语音交互的多模态大模型GPT-4o，9月就出现了Llama-omni[12]等跟进工作。随着新的生成模型不断出现，原有检测模型可能性能降低甚至失效，需要构建具有更强泛化能力的基座检测模型。

（2）如何在强对抗的犯罪场景下进行高精准的鉴伪。对于诈骗等强对抗、高风险犯罪，造假者会采取各种手段逃避检测。例如，造假者可能利用私有模型重述生成文本，抹除文本中来源模型的痕迹，使溯源手段失效；对于图像视频可能采取压缩手段，在仍保留语义信息的前提下减少检测模型依赖的其它信息，导致模型漏检。

（3）如何兼顾新技术的安全与发展，在大量无害生成中精准识别出有害伪造，降低对正向生成应用的影响。生成式人工智能作为新质生产力的代表，未来会催生大量正向生成应用。但从技术层面来说，正向应用和违法犯罪应用依赖的算法、模型在本质上没有区别。影视创作、智能客服等合理应用生成的内容依然会被检测模型识别，既影响这些内容正常传播的权利，也为监管系统造成了更大的负担。

3.3 构建AIGC全流程检测体系

按照公共安全事件“事前—事中—事后”的分阶段管理机制，围绕AIGC生成内容的制作和传播过程，有必要构建“生成时可赋标、传播中可鉴别、案发后可溯源”的AIGC内容检测技术体系。其具体内涵可总结为3个部分。

（1）事前治理：生成时可赋标。针对文本、图像、音频、视频等不同模态生成内容，在模型输出时，通过算法主动植入带有信息的数字水印，水印中包含模型型号、用户身份标识号（ID）等隐式的身份指示信息，在内容可视区域添加用户可明显感知的标识，方便用户识别。

（2）事中治理：传播中可鉴别。针对网络空间中传播的大量未标识内容，使用AIGC检测技术自动识别疑似AIGC，进行标识提醒，对恶意伪造内容进行及时预警。

（3）事后治理：案发后可溯源。针对已经识别到有害的AIGC，开展追查溯源工作。对于带有数字水印的内容，通过显式标识识别、元数据抽取或隐式水印提取等方式，得到生成内容的来源模型名称；对于不带有数字水印的内容，使用生成模型溯源技术，根据内容从候选模型寻找疑似的生成模型；针对未收录的模型生成内容，支持归入“其他”类的开集设置。

4 AIGC检测发展展望与建议

AIGC安全风险治理是一项世界各国共同关心的课题。作为生成式人工智能应用大国，探索和构建AIGC检测技术体系既是维护我国公共安全、引导推动我国人工智能技术健康发展的必要举措，也是为全球人工智能治理积累中国经验、贡献中国智慧的重要契机。中国有望成为世界范围内“人工智能与公共安全”方向的引领者，而率先构建AIGC检测技术体系将成为其中的关键一步。

AIGC检测能力决定着AIGC应用的安全边界，AIGC全流程检测体系的有效建立是AIGC应用蓬勃发展的前提。建立涵盖事前、事中、事后的检测体系不是单纯的技术问题，需要监管部门、科研机构、AIGC服务者紧密合作。面向公共安全实战需求，从技术层、机理层和应用层同步发力，在检测技术与能力不断提升的同时优化制度要求、技术水平和应用场景的适配程度。开展检测能力验证计划，大力推动实战演练，从真实场景中发现痛点问题，达到用技术解决技术问题的效果。

4.1 技术层面

推动AIGC检测能力基座化，实现AIGC检测高效可泛化。面对AIGC技术快速迭代导致的广谱检测和快速响应难题，需要摒弃“来一个打一枪”的事后思维，重视检测能力的基座化。① 构建AIGC检测的基座大模型，提升针对不同来源生成内容的检测泛化能力，突破面向检测大模型的持续学习，实现有限样本下的可扩展模型训练，使模型快速具备新出现AIGC的检测能力；② 提高检测基座的推理效率，通过软硬协同设计，使模型推理与算力基础设施特性相适应，更好地应对大批量AIGC检测需求。

4.2 机理层面

探索生成过程的逆推溯源，促使AIGC检测结果可解释。随着AIGC应用场景日趋复杂，其制作过程往往由多重伪造操作叠加，对鉴伪取证和责任界定构成了严峻挑战。因此，需要探索伪造操作叠加条件下的生成过程逆向解离和原始内容复原。① 全面分析伪

造操作类型，构建覆盖常见伪造工具的特征库，深入解析伪造过程对最终内容施加的影响；② 构建伪造失真分级量化体系，挖掘伪造手段本质模式，增强伪造痕迹的消除和原始特征的还原效果。

4.3 应用层面

面向受众提供多种形式的伪造检测工具，实现“人人可鉴伪”。随着大模型轻量化部署能力的快速发展，生成内容安全风险逐渐转移到终端，每一个普通民众都是“认知战”的受众主体。为应对安全风险终端化的趋势，应从2个方面入手：① 提升大众人工智能技术素养是抵御认知干扰最好的方法，要加大科普力度，提高民众对生成式人工智能技术的认识；② 要给民众提供简单易用的鉴伪服务和鉴伪工具，让普通用户在身份验证、内容鉴定等日常场景中有工具可用。例如，杭州中科睿鉴科技有限公司发布的“终端AI鉴伪大师”将鉴伪服务深度融入终端系统，已在手机、平板电脑、笔记本电脑等消费级终端设备上部署，实现对视频通话、会议、直播等场景下伪造内容及时告警，及时保护终端用户安全。

参考文献

1 王祺, 李冬露, 张云, 等. 2023年中国AIGC产业全景报告. 北京: 艾瑞咨询, 2023.
Wang Q, Li D L, Zhang Y, et al. 2023 China AIGC Industry Panorama Report. Beijing: iResearch Center, 2023. (in Chinese)

2 Jakesch M, Hancock J T, Naaman M. Human heuristics for AI-generated language are flawed. PNAS, 2023, 120(11): e2208839120.

3 Pocol A, Istead L, Siu S, et al. Seeing is no longer believing: A survey on the state of deepfakes, ai-generated humans, and other nonveridical media// Computer Graphics International Conference. Shanghai: Springer, 2023: 427-440.

4 Mai K T, Bray S, Davies T, et al. Warning: Humans cannot reliably detect speech deepfakes. PLoS One, 2023, 18(8): e0285333.

5 Liang W X, Izzo Z, Zhang Y H, et al. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews// Proceedings of the 41st International Conference on Machine Learning. Vienna: ML Research Press, 2024: 29575-29620.

6 Mitchell E, Lee Y, Khazatsky A, et al. DetectGPT: Zero-shot machine-generated text detection using probability curvature// Proceedings of the 40th International Conference on Machine Learning. Honolulu: ML Research Press, 2023: 24950-24962.

7 Yang T Y, Huang Z Y, Cao J, et al. Deepfake network architecture attribution. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(4): 4662-4670.

8 Sun Z H, Fang H P, Cao J, et al. Rethinking image editing detection in the era of generative AI revolution// Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne: ACM, 2024: 3538-3547.

9 Shi Y H, Sheng Q, Cao J, et al. Ten words only still help: Improving black-box AI-generated text detection via proxy-guided efficient re-sampling// Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Jeju: IJCAI, 2024: 494-502.

10 Yang T Y, Wang D D, Tang F, et al. Progressive open space expansion for open-set model attribution// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 15856-15865.

11 Yang T Y, Cao J, Wang D D, et al. Model synthesis for zero-shot model attribution. arXiv preprint, 2024, doi: arxiv.org/abs/2307.15977.

12 Fang Q K, Guo S T, Zhou Y, et al. LLaMA-Omni: Seamless speech interaction with large language models. arXiv preprint, 2024, doi: org/10.48550/arXiv.2409.06666.

Challenges on public security system in AI era

—Preface for special column "Artificial Intelligence and Public Security" in *Bulletin of Chinese Academy of Sciences*

CAO Juan[1,2*] SHENG Qiang[1] LI Guojie[1]

（1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;

2 School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China）

Abstract The rapid development of artificial intelligence generated content (AIGC) technology has triggered new public security risks, posing a serious threat to national security and social stability. This study exhibits the recent advances of artificial intelligence (AI) content generation and detection techniques, points out the challenges of detection techniques in real-world scenarios, and advocates that it is necessary to develop AIGC detection technology for public security needs and build a whole-process detection technology system from generative models to online platforms, which supports AIGC to be labeled at the generation phase, identifiable during dissemination and source-traceable after the incident occurs.

Keywords artificial intelligence generated content (AIGC), generated content detection, public security, whole-process governance

曹 娟 中国科学院计算技术研究所前瞻研究实验室主任、数字内容合成与伪造检测实验室主任、研究员。主要研究领域：人工智能内容安全。E-mail: caojuan@ict.ac.cn

CAO Juan Professor of Institute of Computing Technology, Chinese Academy of Sciences (CAS), Director of Prospective Research Laboratory and Media Synthesis and Forensics Laboratory. Her research focuses on artificial intelligence generated content (AIGC) safety. E-mail: caojuan@ict.ac.cn

李国杰 中国工程院院士，发展中国家科学院院士。中国科学院计算技术研究所原所长、研究员。《中国科学院院刊》副主编。主要从事并行算法、高性能计算机、未来网络、人工智能、大数据和技术发展战略等领域的研究。E-mail: lig@ict.ac.cn

LI Guojie Academician of Chinese Academy of Engineering, Fellow of the World Academy of Sciences for the advancement of science in developing countries (TWAS). Professor of Institute of Computing Technology, Chinese Academy of Sciences (CAS). He serves as Associate Editor-in-Chief of *Bulletin of Chinese Academy of Sciences*. He is the former Director of the Institute of Computing Technology, CAS. He mainly engages in research on parallel algorithm, high performance computer, future network, artificial intelligence, big data, and technology development strategy. E-mail: lig@ict.ac.cn

■责任编辑：文彦杰 梁小星

*Corresponding author