基于行政执法责任界分的生成式人工智能内容安全监管研究*

盛小宝，刘晋名，姚　迁
（公安部第三研究所，上海　200100）

摘　要：生成式人工智能（Generative AI，GenAI）产业结构和内容生成机制的特殊性，导致其内容安全监管面临诸多挑战，特别是在行政执法责任的划分上。目前，公安网安部门、网信部门和工业和信息化部在 GenAI 内容监管中存在职责模糊与交叉的问题，进而影响监管效果。具体表现为技术手段与监管需求错配、现有法律法规适用性不足、公众认知不足和企业责任履行不力等具体面向。对此，亟需优化监管制度路径：一是以明确各执法部门的职责边界作为起点，建立健全跨部门的信息共享与协同机制；二是通过技术赋能，构建完善内容安全的分级治理框架；三是发挥企业主体积极性，协同赋能推动监管效率之提升。未来，应推动法律、技术与管理机制的深度融合，构建更加科学、规范的 GenAI 内容安全监管体系，为网络空间安全和社会秩序提供有效的制度保障。

关键词：生成式人工智能；内容安全；行政执法边界；责任界分
中图分类号：D922；TP399　　**文献标志码：**A　　**DOI：**10.19358/j.issn.2097－1788.2026.01.009

中文引用格式：盛小宝，刘晋名，姚迁. 基于行政执法责任界分的生成式人工智能内容安全监管研究［J］. 网络安全与数据治理，2026，45(1)：56－63.
英文引用格式：Sheng Xiaobao，Liu Jinming，Yao Qian. Research on the content safety regulation of generative artificial intelligence based on the division of administrative enforcement responsibilities［J］. Cyber Security and Data Governance，2026，45(1)：56－63.

Research on the content safety regulation of generative artificial intelligence based on the division of administrative enforcement responsibilities

Sheng Xiaobao，Liu Jinming，Yao Qian
（Third Research Institute of the Ministry of Public Security，Shanghai 200100，China）

Abstract：The uniqueness of the industrial structure and content generation mechanism of generative artificial intelligence（GenAI）has led to numerous challenges in content security supervision，particularly in the division of administrative enforcement responsibilities. Currently，there is ambiguity and overlap in the roles of public security cybersecurity departments，the Cyberspace Administration，and the Ministry of Industry and Information Technology（MIIT）in GenAI content regulation，leading to unclear enforcement responsibilities，and consequently，ineffective regulation. This issue manifests in mismatches between technical tools and regulatory needs，inadequate applicability of existing laws and regulations，insufficient public awareness，and lack of corporate responsibility. To address these challenges，the regulatory framework must be optimized：First，the clear delineation of responsibilities across enforcement agencies should be prioritized，with the establishment of a robust information-sharing and collaboration mechanism；second，leveraging technology to build a comprehensive，tiered governance framework for content safety；third，encouraging corporate responsibility and collaborative efforts to improve regulatory efficiency. Moving forward，it is crucial to integrate law，technology，and management mechanisms to create a more scientific and standardized regulatory system for GenAI content safety，providing effective institutional safeguards for cybersecurity and social order.

Key words：generative artificial intelligence；content safety；administrative enforcement boundaries；responsibility delineation

* 基金项目：2024LL36 项目

0 引言

在数字化技术蓬勃发展的背景下，生成式人工智能（Generative AI，GenAI）凭借其在文本、图像、音频等领域的强大内容生成能力，正在信息传播、内容创作和社会互动等诸多领域引发深刻而颠覆性的变革。从 GPT 系列的自然语言生成模型到 DALL-E、Stable Diffusion 等图像生成模型，GenAI 通过深度学习算法（如生成对抗网络 GANs 和变分自编码器 VAEs），基于海量数据训练，实现了高质量、与训练数据相似的内容自动生成。这一技术不仅广泛应用于娱乐、教育、新闻等行业，还通过其基础性支撑作用辐射至众多领域，推动产业创新发展、经济运行模式变革和社会进程迭代。然而，随着 GenAI 技术的快速迭代与广泛应用，其引发的内容安全问题日益突出，成为限制技术健康发展的关键挑战。首先，GenAI 技术的迅速发展加剧了虚假信息的传播，可能严重误导公众认知，扰乱社会秩序，甚至威胁国家网络安全。其生成的虚假内容不仅可能被不法分子利用操控舆论，还可能被用于制造社会对立，进而影响社会稳定。其次，内容版权侵权问题频发，侵犯创作者的合法权益，影响创作生态的健康发展。再次，GenAI 技术在处理海量数据时存在极高的隐私泄露风险，可能导致个人信息的非法获取和滥用，严重威胁个人权益保护，甚至对国家的数据安全形成潜在威胁。这些问题的出现呼应了“科林格里奇困境”（Collingridge Dilemma）。正如大卫·科林格里奇（David Collingridge）所指出，在技术发展的早期，其社会后果往往难以预见；而当不良后果显现时，该技术可能已深度嵌入社会经济结构，使得实施有效控制变得异常困难[1]。

针对这些问题，GenAI 内容安全监管已成为维护国家网络安全、社会稳定与个人权益保护的重要任务。然而，目前我国在 GenAI 内容安全监管领域仍面临诸多挑战，尤其是行政执法责任划分不明确，导致监管效率低下，监管责任难以落实。现行监管体系下，涉及 GenAI 内容监管的部门众多，包括公安机关网络安全部门（以下简称“网安部门”）、国家互联网信息办公室（以下简称“网信部门”）、工业和信息化部（以下简称“工信部门”）等。由于 GenAI 内容的复杂性与多样性，各部门在实际监管过程中职责交叉和界限模糊的现象屡见不鲜。此外，现有监管制度在应对 GenAI 技术快速发展时也暴露出诸多不足，难以满足技术迭代与应用场景变化带来的治理需求。为实现 GenAI 内容安全的有效治理，应聚焦行政执法责任划分，深入分析 GenAI 内容安全的内涵及其监管现状，剖析当前面临的主要障碍，从风险导向出发，提出科学合理的制度设计与可行的监管措施，明确 GenAI 内容安全的行政执法责任，完善行政执法体系，构建良性技术应用生态，从而为维护国家安全、保障社会和谐提供有力支撑。

1 GenAI 内容安全的监管现状及面临的挑战

在 GenAI 模型即服务的新型产业模式下，生成内容的责任归属呈现复杂化特征，涉及《香港生成式人工智能技术及应用指引》所界定的 AI 技术开发者（指从事基础模型、算法的开发、训练及维护的主体）、AI 服务提供者（指作为技术开发者与用户之间中介、负责将技术部署为面向客户应用或服务的主体）、AI 服务使用者（指出于个人或专业用途使用服务的主体）等多元主体。我国《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）第九条明确规定服务提供者对网络信息内容负有生产者责任，但此规定未能清晰界定以下关键边界：一方面，AI 技术开发者在提供基础模型时对最终生成内容安全所承担的具体责任（《互联网信息服务深度合成管理规定》（以下简称《深度合成管理规定》）虽对类似角色“技术支持者”有所规制，但与《暂行办法》的衔接尚需明确）；另一方面，尤其是在高度人机交互的背景下，《暂行办法》《互联网信息服务算法推荐管理规定》（以下简称《算法推荐管理规定》）《深度合成管理规定》及《人工智能生成合成内容标识办法》（以下简称《标识办法》）均未明确使用者利用提示词操控生成过程时的法律角色与责任承担。这种主体责任界定的模糊性与使用者责任的缺位，不仅阻碍了行政执法机关对生成内容安全性的有效溯源与认定，也对监管执法效率产生了不利影响。为实现有序、高效的 GenAI 内容安全监管，亟需进一步厘清 GenAI 内容安全的内涵与技术风险特征，同时全面梳理当前各行政执法机关在 GenAI 内容安全治理中的分工现状及其所面临的主要困境，从而为完善监管机制、提升治理能力提供理论基础和实践指引。

1.1 GenAI 内容安全的内涵厘清

GenAI 技术的特性使其在内容生成领域展现了显著的优势，但也引发了独特的内容安全风险，主要表现为以下几个方面：一是内容真实性与误导性并存。GenAI 基于概率生成的机制使生成内容追求评价系统

的认可，而非真实准确。这种特性易导致内容失实，甚至吸纳训练数据中的社会偏见，例如性别、种族刻板印象及身份攻击等。此类信息的广泛传播可能误导公众认知，加剧社会矛盾。若此类内容缺乏明确的人工智能生成标识，用户将失去关键的信息来源判断依据，导致误导性与社会危害性被进一步放大。二是内容形式多样且隐蔽性强。GenAI 在生成深度伪造视频、合成音频、虚假文本等方面表现出极强的隐秘性和迷惑性，大幅增加识别与监管难度。一旦缺失显式且可追溯的技术标识，此类高度隐蔽的虚假或伪造内容的传播范围与欺骗成功率将显著提升，监管难度倍增。三是算法黑箱带来的不可控性。GenAI 的生成过程依赖深度学习算法，且常处于“黑箱”状态，其内部决策机制和模型权重对外界不可见，难以追溯。这使得 GenAI 的生成行为难以预测与监管，可能导致在未被察觉的情况下生成潜在有害内容，甚至被恶意利用进行网络攻击或其他犯罪活动。缺乏有效的生成内容标识机制，将导致有害内容的源头追溯与责任认定变得极其困难，使算法黑箱的不可控风险转化为实际监管失控的风险。

基于上述特性，GenAI 技术引发的内容安全问题，尤其是在违法和不良信息的生成与传播方面，已成为不容忽视的严重威胁。这些问题不仅涉及个人权益保护和社会稳定，还可能对国家安全造成严重影响。根据《互联网信息服务管理办法》《暂行办法》《网络信息内容生态治理规定》及《网络安全技术 生成式人工智能服务安全基本要求》，违法和不良信息可分为以下几类（见图 1）：违反社会主义核心价值观类内容、歧视性内容、商业违法违规类内容、侵犯他人合法权益类内容以及无法满足特定服务类型的安全需求。上述违法和不良信息对社会秩序和公共安全构成了严重威胁，尤其在重大公共事件中，若再叠加内容标识缺失的漏洞，可能导致有害信息以“隐身”状态快速扩散，带来不可预估的后果。

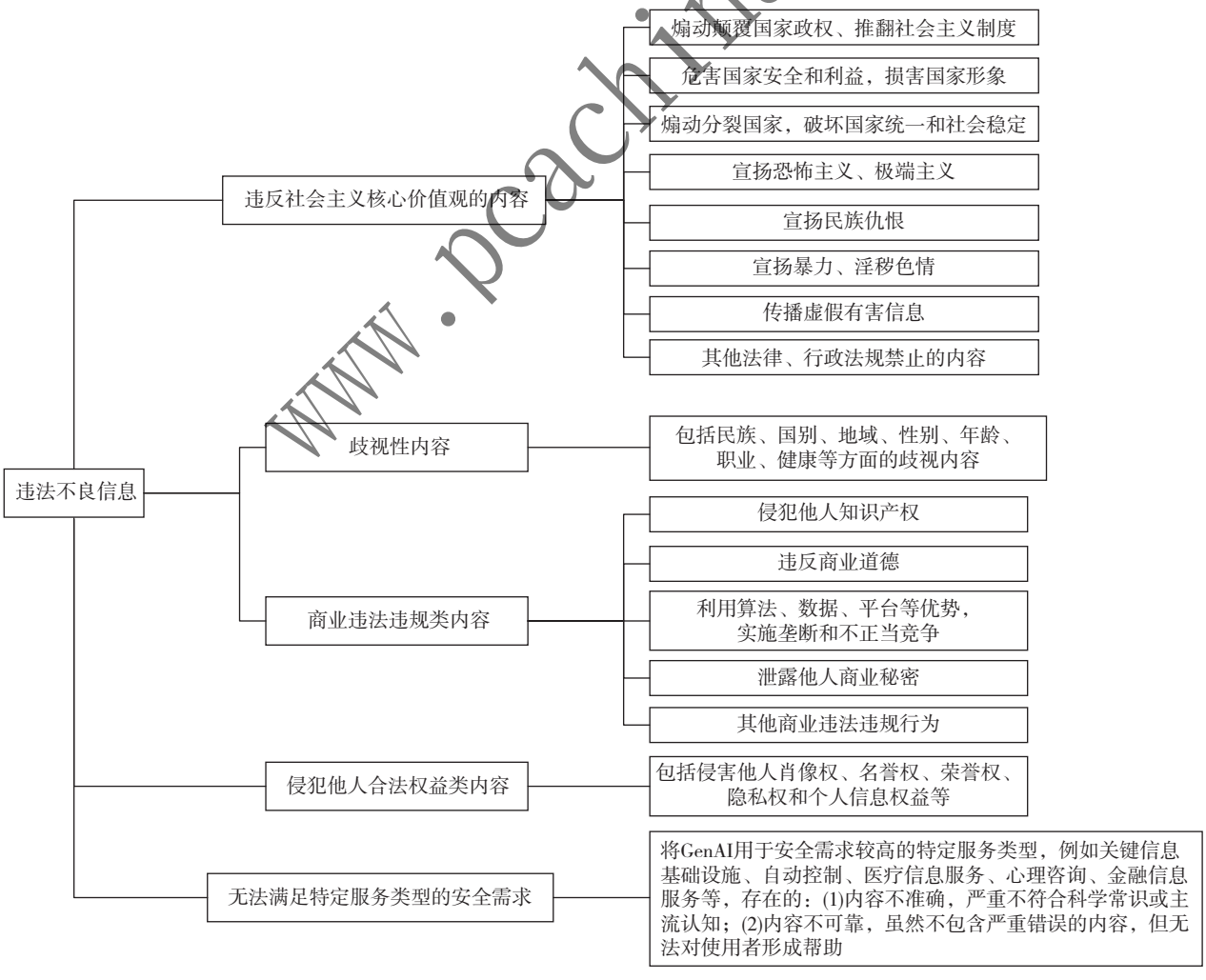

图 1　违法和不良信息内容类型图

1.2 行政执法机关对 GenAI 监管的职责分工现状

当前，我国 GenAI 内容安全监管领域涉及多个行政执法主体，呈现多部门协同监管格局。如何科学划分执法边界、避免职责交叉导致的“九龙治水”问题，已成为提升监管效能的关键挑战[2]。现阶段主要监管主体包括网安部门、网信部门和工信部门等。表 1 列出了 GenAI 内容监管部门的职责界面。

表 1　GenAI 内容监管部门的职责界面

	网安部门	网信部门	工信部门
监管权源	负责网络空间的公共安全。	负责网络生态治理。	对信息化领域的监督和执法。
核心监管职责	预防、打击与遏制网络犯罪，保护公民个人信息及维护国家安全等。打击利用 GenAI 的违法犯罪行为。	统筹协调、宏观指导；建设内容安全和信息生态，引导、规范、审查生成内容。	拟订实施行业规划、产业政策和标准；企业合规指导。
监管现状	以网络诈骗、虚假信息传播、深度伪造等技术型犯罪为聚焦，犯罪技术手段复杂、行为隐蔽性高，需强化技术水平和侦查能力。	以内容的真实性、健康性为聚焦，面临海量内容审查与资源调配压力。	以日常监督管理、企业合规性为聚焦，需完善适应技术迭代的弹性监管机制。
执法实践	建立专项执法流程与证据固定标准。	实施生成内容标签化管理与平台责任监督。	推行技术标准与政策以促企业合规。

首先，网安部门以维护网络公共安全为核心职能，侧重打击利用 GenAI 实施的网络诈骗、深度伪造、虚假信息传播等违法犯罪活动。依托既有网络犯罪侦查经验，网安部门已初步形成针对 GenAI 违法行为的执法流程与取证标准。

其次，网信部门负责统筹网络生态治理，负责 GenAI 生成内容的导向审查与真实性管理。通过压实平台主体责任、制定内容安全规范，保障信息传播符合社会主义核心价值观。面对海量生成内容，其现有技术手段与人力资源面临识别效率压力。

最后，工信部门负责履行行业监管与技术标准制定职能，监督企业在 GenAI 数据处理、模型训练及应用部署中的合规性，重点推动技术研发安全标准落地，但需动态调整政策以适应技术迭代速度。

因 GenAI 生成内容的隐蔽性、传播快速性及社会影响广泛性，各部门在履职过程中易出现职责交叉或界面模糊现象[3]。具体而言，GenAI 内容的多维度特性使得监管责任难以单一界定。首先，虚假信息的生成与传播既属于网信部门的信息内容规范管理范畴，也涉及网安部门针对信息犯罪的侦查职能[4]。其次，企业在数据使用中的行为需同时满足工信部门的技术安全标准和网信部门的内容管理要求。此类交叉问题导致执法资源分散与效率损耗。厘清部门间责任接口，建立权责明确、协作高效的执法机制，成为当前治理体系优化的核心任务。

当前监管框架虽初步建立，但在执法协同层面仍存在显著缺陷。首先是违法内容处置中权限交叉导致协调不畅，易出现推诿或重复执法。例如虚假信息案件需网信与网安双重介入。其次是跨部门动态监测数据未能有效整合，制约风险预警与应急响应效能[5]。最后是技术设备与专业人才分散配置，难以应对 GenAI 内容的指数级增长与跨平台风险扩散。此类缺陷不仅影响了执法效能，还可能带来监管盲区。尤其在违法内容引发舆情危机时，协同机制缺失可能放大社会风险。亟需通过制度重构破解多头监管困境，以匹配技术演进下的治理需求。

1.3 GenAI 内容监管面临的主要挑战

GenAI 的颠覆性特征对全球监管体系构成严峻挑战[6]。当前各国的监管普遍呈现应急性特征，各国通过技术审查、风险评估等临时性措施遏制数据安全与隐私风险。我国建立以法律治理为核心、技术治理为支撑的复合监管体系，但在技术迭代加速的背景下仍面临三重结构性矛盾：技术能力与监管需求的鸿沟持续扩大、法律规则与执法实践的脱节日益显著、多元主体责任缺位导致治理效能衰减。这些矛盾制约监管效能的深度释放，威胁国家信息生态安全。

1.3.1 技术迭代与监管能力的结构性断层

监管技术的滞后性构成首要挑战。GenAI 依托 Transformer 架构生成内容，其概率统计特性导致输出以合理性而非真实性为导向[7]，系统性产生虚假信息及歧视性内容。与违法有害信息相比，GenAI 生成的深度伪造视频、虚假文本及合成音频，常以高度仿真

形式伪装为可信信息，使得公众难以辨别真假。深度伪造技术在多模态融合支持下，其内容仿真度已突破传统检测工具的有效识别阈值。研究表明，当合成质量达 90% 以上时，检测工具的误判率显著上升。同时，黑灰产团伙利用对抗样本技术持续优化攻击手段，迫使监管部门识别模型效能呈现递减趋势。更具危害性的是训练数据污染的风险传导机制，恶意数据投毒通过扭曲模型参数，显著提升有害内容的生成概率。例如开源社区对 Llama2 模型的定向投毒实验，使其输出错误医疗建议的概率倍增。这些内容的复杂性和隐蔽性大大增加了现有技术手段的识别成本。在执法资源配置层面，目前的监管技术多依赖特征提取或基于规则的算法，基层监管部门普遍面临技术能力薄弱、动态数据库缺失与专业人才匮乏三重困境，难以快速适应 GenAI 生成内容的快速演变，从而使得虚假信息在社交媒体平台上的传播速度和影响力远超传统模式[8]。

1.3.2 法律体系与实践需求脱节

当前，我国已经形成以《暂行办法》《算法推荐管理规定》《网络信息内容生态治理规定》《深度合成管理规定》《标识办法》《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》（以下简称《安全评估规定》）为核心的监管框架，但存在显著实施困难。首先，现有法规缺乏具体场景的操作性指引。例如，《暂行办法》虽然涵盖了技术治理和服务合规，但在虚假信息、深度伪造等问题上缺少明确的处理标准，导致执法人员难以依据法律快速行动。其次，监管职能分散与协同机制缺位形成制度瓶颈。网安、网信、工信三部门在数据权限封闭与技术标准脱节的交互作用下，形成碎片化的蜂窝状监管结构。各部门因业务范畴、技术系统建设以及管理模式的差异，处于“执法信息孤岛”状态[8]，缺乏常态化的信息共享和应急响应机制，进而无法形成实时、高效的联合行动，因跨部门信息传递延误将致使经济损失扩大。更值得关注的是法律惩处力度与违法收益失衡的深层矛盾，最高 200 万元的行政罚款相较于技术滥用带来的经济收益威慑效能有限。

1.3.3 公众认知缺陷与企业责任缺失

公共认知缺陷与企业责任缺失形成治理薄弱环节。当前普通用户对 GenAI 技术机理的有限理解严重制约其辨识深度伪造内容的能力，这种认知缺陷不仅导致社会风险抵御能力弱化，更使公众在参与技术治理时难以形成有效协同。当部分企业为追求技术领先优势与市场份额，在关键领域削弱内容安全控制措施时，例如未建立完备审核机制即开放高风险语音合成功能——实则将技术伦理让位于商业逻辑。这种工具理性优先的发展模式持续消解社会信任基础，形成技术风险向社会领域的外溢传导。更深层次的系统性症结在于技术伦理共识建构的集体缺失：开发者迭代效率至上的技术逻辑与公众认知滞后之间的矛盾，既阻碍了风险沟通机制的有效运行，也持续削弱监管框架的实施效能，最终导致技术治理成本的非理性扩张。

2 GenAI 内容安全监管的制度优化路径

当前，GenAI 的快速发展对我国行政执法体系和法律框架提出了新的挑战。由于 GenAI 相关的治理和监管仍处于初步阶段，各执法部门在实施过程中缺乏统一的监管机制，职责边界模糊，且 GenAI 带来的内容合规性和安全性问题日益严峻。因此，需要在现有法律框架下设计切合我国国情、具备操作性的解决路径。此路径应加强跨部门合作、信息共享、内容治理标准化、技术赋能以及公众监督，构建多层次的 GenAI 治理体系。为实现这一目标，应从法律依据和执法框架出发，完善部门间协作机制，明确职能划分，推动形成系统的治理模式。同时，要确保这些路径的实际可操作性，并从法律和社会治理的角度进行分析，确保其有效性和可持续性，以促进 GenAI 技术的良性应用与发展。

2.1 厘清职责边界，构建触发式协同机制

针对当前 GenAI 内容安全监管中行政执法责任模糊与交叉的问题，亟需从组织法维度重构职权配置，明确各监管部门的职责边界，并在此基础上构建高效的跨部门协同治理体系，以破解“九龙治水”困境。在《暂行办法》修订中，应进一步细化并明确各监管部门在 GenAI 内容安全监管中的核心职能与责任边界，避免职责重叠和推诿。网安部门应聚焦利用 GenAI 实施的网络诈骗、深度伪造、虚假信息传播等违法犯罪行为的侦查、打击与遏制。网信部门应主导 GenAI 生成内容的合法性审查与导向管理，统筹网络生态治理。工信部门应专司技术安全标准制定、安全主体责任的监督检查与产业发展引导。

在明确职责边界的基础上，应建立一套高效的风险阈值触发机制和常态化协同机制，确保各部门在 GenAI 内容安全事件中的无缝衔接与快速响应。风险阈值触发机制应设定明确的风险评估标准和触发条件，例如当虚假信息传播量级突破预设标准（如传播范围、点击量、社会影响程度等）或深度伪造技术应用于特

定高危场景（如涉及国家安全、公共安全、金融诈骗等）时，自动激活跨部门应急响应程序，同时确保响应的及时性和精准性。常态化信息共享与应急响应平台建设应依托现有技术平台和机构，建立全国性的GenAI风险信息共享平台，推动监管信息的高效流通与协同响应[9]。该平台可以借鉴已有的信息共享平台经验，例如国家政务服务平台或公安信息资源库，在此基础上进行扩展，整合网安部门、网信部门、工信部门等相关部门的数据资源与技术能力，打破信息孤岛，实现监管数据的动态交互与全流程共享。平台建设应严格依据我国《网络安全法》《数据安全法》《个人信息保护法》等相关法律法规进行，突出合规性与安全性，确保共享过程中信息合法、准确和高效。具体操作上，平台应开发统一的数据接口标准[10]，并结合我国网络安全等级保护制度，制定涵盖数据存储、传输与使用的规范化要求，确保数据流通符合国家信息安全标准。为确保数据的安全性，应对敏感数据采取加密存储、链路加密和访问控制等技术手段。此外，可以引入零知识证明技术和联邦学习等隐私保护计算方式，在保障隐私的前提下，促进信息价值的共享和流通。平台还应设计实时监控与应急响应功能，通过算法对GenAI应用中的风险数据进行实时监控与智能分析，及时发现潜在风险。并建立高风险事件的预警系统，一旦检测到重大风险，能快速通过反馈通道协调各部门进行任务分配和响应行动，确保风险处置的高效性和针对性。

为确保上述机制的有效运行，需要强化顶层设计和法律保障，确保跨部门协同治理的权威性和可持续性。建议由中央网信办牵头设立人工智能治理协调办公室，赋予其风险态势评估、执法资源调度以及跨部门协调的职权。该办公室应负责统筹GenAI治理战略与方向，确保跨部门合作的顺畅进行，并定期对协同机制的运行效果进行评估和优化。针对现有法律法规在GenAI监管中存在的适用性不足和惩戒力度有限的问题，应推动相关法律法规的修订与完善，明确GenAI违法行为的法律责任，并提高违法成本，使其与违法收益相匹配，形成有效震慑。同时，加强《暂行办法》《算法推荐管理规定》《深度合成管理规定》等法规之间的衔接，形成统一、协调的法律体系。最后，鉴于当前跨部门信息共享面临的体制障碍与法律协调难题，应进一步探讨“数据治权”模式，厘清数据的所有权、使用权与保护权的划分，确保各部门在法律框架下高效协作。通过引入区块链等去中心化存储技术，提升数据共享的安全性和透明性[11]，增强跨部门协作的法律保障与技术支撑。

2.2 健全内容审核标准，构建风险分级治理框架

为有效应对GenAI内容带来的复杂合规挑战，有必要建立统一、科学的内容审核标准和风险分级治理体系[12]，确保生成内容符合法律法规和道德规范要求，并实现监管资源的优化配置。工信部门联合多部门发布的《国家人工智能产业综合标准化体系建设指南》，已为人工智能标准体系提供了全面框架，包括基础支撑、关键技术、安全治理等多个部分，为构建内容治理标准奠定了基础。在此基础上，建议进一步细化和完善内容审核机制，形成分级管理和分类审核的统一操作规范。该机制需与《标识办法》形成协同，通过植入可追溯元数据有效破解隐蔽性难题，为行政执法提供技术支撑。

借鉴“分层治理”模式[13]和欧盟《人工智能法案》的分级监管思路，对GenAI生成内容进行风险分级管理，以实现精细化和精准化的治理目标。一级风险内容指对国家安全、公共利益、社会秩序和公民合法权益造成严重危害的内容，例如严重违反社会主义核心价值观、煽动暴力、恐怖主义、大规模虚假信息传播、深度伪造国家领导人或重要公共人物等。对此类内容，应直接适用国家最高标准，实施最严格的审核与管控措施，包括但不限于强制事前审核、即时删除、永久封禁相关账号、追究刑事责任等。二级风险内容指对社会秩序和公民合法权益造成一定危害，但危害程度相对较低的内容，例如歧视性言论、商业违法违规、侵犯他人合法权益（非严重）、误导性信息等。对此类内容，可在行业规范框架下采取适度的管控措施，如提示性警告、限制传播、要求整改、行政处罚等，并赋予服务提供者一定的灵活处置空间，鼓励其通过技术手段进行自查自纠。三级风险内容指对社会影响较小，主要涉及伦理道德或轻微不当的内容。对此类内容，可主要依靠行业自律、用户举报和平台自治进行管理，监管部门进行引导和监督。

针对不同风险等级的GenAI生成内容，需结合现有平台资源与技术手段，建立细化的审核规则和分级管理机制，以实现内容治理的精准性与高效性[14]。应依托国家政务服务平台、公安信息资源库等现有平台，拓展其功能至内容治理领域，实现资源整合与跨部门协同，减少重复建设成本。对于特定行业的数据与风险信息共享，可由各行业主管部门负责，依据相关法律法规将分级审核任务落实到位，确保治理工作有序

推进。采用“自动审核 + 人工复核”的多层次审核体系[15]，自动审核阶段利用自然语言处理（NLP）、多模态分析等技术对生成内容进行初步筛查，高效过滤潜在风险内容。特别是对于方言识别等技术短板，应加大研发投入，提升自动检测工具的准确性。人工复核阶段由内容审核专家对高风险内容进行深入甄别，确保审核的全面性与准确性。尤其是针对一级安全风险，应通过严格的审核与管控措施确保不发生监管盲区。压实 GenAI 服务提供者的内容安全主体责任，要求其建立健全内部审核机制，配备专业审核团队，并定期向监管部门提交内容安全报告。对于未能履行主体责任的企业，应依法予以严厉处罚。

2.3 强化技术赋能，深化政企协同治理

提升执法部门在识别、监控和追溯 GenAI 内容方面的能力，并在此基础上推动监管提质增效，关键在于强化技术赋能与企业协同。行政执法部门应积极引入前沿技术，提升对违法不良内容的精准识别与追溯能力，并针对现有技术短板进行重点突破。具体而言，深度伪造检测技术方面，应加大研发投入，特别是针对多模态融合、高仿真度伪造内容的识别，支持产学研合作，开发具有自主知识产权的检测工具，并建立动态更新的伪造样本数据库，以应对对抗样本攻击和技术迭代带来的挑战。同时，针对 GenAI 内容形式多样且隐蔽性强的特点，发展多模态分析技术至关重要，以实现对文本、图像、音频、视频等多种模态内容的综合分析与交叉验证，提高识别的准确性和效率，并着力解决方言识别资源匮乏和误判问题，开发针对不同方言的识别模型。此外，应探索利用区块链技术构建 GenAI 内容溯源平台，记录内容的生成、传播路径和修改历史，实现全链条可追溯。针对基层部门算力不足和数据隐私保护的需求，推广联邦学习等隐私保护计算技术，在不共享原始数据的前提下，实现跨部门、跨机构的模型协同训练和风险信息共享，从而提升监管效率。

在技术赋能的同时，行政执法机构应与企业加强合作，明确各自职责边界，并通过技术手段确保企业责任的落实，避免“技术依赖”和“成本转嫁”。为此，建议引入强制合规认证制度，确保 GenAI 技术符合伦理和安全要求。同时，要求 GenAI 服务提供者开放必要的技术接口，便于监管部门进行内容安全检测、风险评估和数据审计，但需注意避免过度干预企业技术创新。强制要求 GenAI 系统内嵌内容溯源功能，将生成内容的来源信息与元数据集成到执法平台。借鉴欧盟《数字服务法案》第 24 条关于透明度报告的要求，要求 AI 开发商披露模型训练数据来源及生成内容标识信息，为违法行为追溯提供法律与技术支持。为激励企业积极参与，可设立专项资金或税收优惠政策，鼓励企业投入 GenAI 内容安全技术研发，并建立行业自律联盟，共同制定行业标准和行为准则，推动企业履行社会责任。值得注意的是，监管部门在引入企业技术工具时，应确保对关键技术的主控权，警惕因技术手段对执法行为形成掣肘，规避因依赖第三方工具滋生技术垄断或执行效率滑坡问题。同时，应合理分担内容安全治理成本，避免将全部成本转嫁给企业，影响其创新积极性。溯源功能作为 GenAI 内容安全监管的必要环节，其增加的成本应视为企业履行社会责任的合理投入。为避免过度技术依赖，应采取“技术中立”原则，即不强制企业采用特定技术实现溯源，而是要求其达到溯源效果[16]。监管部门应提供多种技术路径选择（如数字水印、元数据嵌入、区块链存证等），并鼓励企业自主创新，开发更高效、低成本的溯源方案。同时，通过税收优惠、项目补贴等方式，激励企业在溯源技术研发和部署上的投入，将“增加成本”转化为“创新动力”。

国际经验表明，企业技术赋能与行政执法责任界分的有机融合，需要技术、法律与机制的协同联动。一方面，构建技术治理框架至关重要，例如欧盟《人工智能法案》明确提出加强对 AI 系统的监管工具研发，并推动跨机构技术能力的共享，尤其是在行政执法责任的划分方面[17]。另一方面，“算法公平性”理论指出，GenAI 内容监管应在技术创新与法律约束之间寻求平衡，并明确行政执法的责任。例如，美国国会在推动 AI 监管立法时特别关注算法偏见和歧视问题，确保技术使用中的公平性与透明性；英国数据伦理中心（CDEI）也在研究如何通过法规和标准规范 AI 在执法领域的应用，避免因算法歧视导致执法结果不公。

综上所述，通过明确行政执法责任界分，并利用企业技术赋能推动执法创新，不仅能够提升对 GenAI 生成内容的监管效率，还为构建责任明确、标准统一的追责机制提供技术保障。我国可以结合自身实际需求，制定符合国情的技术监管政策，吸纳国际先进经验，探索技术与治理相融合的监管路径。企业技术赋能与行政执法责任明确相结合的模式，将有效推动 GenAI 治理的科学化与体系化进程，为全球 AI 治理贡献中国智慧与方案。

3 结束语

综上所述，伴随着 GenAI 技术的快速发展，其在内容安全监管中的潜在风险和挑战日益显现。行政执法部门在应对 GenAI 带来的新型内容安全监管需求时，需要克服现有职责划分模糊、技术与监管需求错配、法律法规滞后等问题，以适应不断变化的技术和社会环境。

针对这些挑战，可以通过以下路径优化治理机制：第一，明确跨部门执法职能边界，构建信息共享和协同机制，提升执法效能；第二，依托技术手段制定内容审核标准和分级治理框架，平衡监管效率与规范性；第三，与企业建立合作关系，充分利用其技术优势助力行政执法责任的合理划分，从而形成多方共治格局。这些措施旨在实现 GenAI 内容安全的高效监管，维护国家安全、社会秩序以及公民合法权益。

GenAI 内容安全的治理不仅是技术创新的命题，更关乎法律规范的调整与社会治理模式的优化。通过强化技术赋能、促进企业协同和完善公众参与，可以构建一个多方共治、协同高效的 GenAI 内容安全监管体系，从而提升监管效能，保障 GenAI 技术的健康发展。在未来的动态治理过程中，还需注重跨学科研究和多维度实践，构建科学、有效且可持续的监管体系，以确保 GenAI 在技术进步与社会责任之间取得良性平衡，为全球数字治理提供借鉴与经验。

参考文献

[1] COLLINGRIDEG D. The social control of technology [M]. New York: St. Martin's Press, 1980.
[2] 王猛，王红莉. 双重整合与重心下沉：中国行政执法体制改革的内在逻辑与实现路径 [J]. 中国行政管理，2024，40 (6)：59 – 69.
[3] 徐远. 论生成式人工智能与国家创新体系的现实契合与应然互动 [J]. 河北经贸大学学报，2024，45 (5)：52 – 61.
[4] 张素华，李凯. 生成式人工智能虚假信息风险与治理研究 [J]. 学术探索，2024 (7)：129 – 140.
[5] 戚建刚，杨丰合. 论公共卫生数字信息共享机制的行政法规制 [J]. 思想战线，2024，50 (3)：120 – 131.
[6] 陈禹衡. 生成式人工智能中个人信息保护的全流程合规体系构建 [J]. 华东政法大学学报，2024，27 (2)：37 – 51.
[7] 张旭芳. 生成式人工智能的算法安全风险及治理路径 [J]. 江西社会科学，2024，44 (8)：90 – 100.
[8] 刘春年，陈梦秋，易岚. 深度伪造视频中的信息特征萃取及其关联计算 [J]. 情报杂志，2024，43 (8)：92 – 101.
[9] 张铤. 人工智能的伦理风险治理探析 [J]. 中州学刊，2022 (1)：114 – 118.
[10] 郑曦，杨宇轩. 司法领域关键信息基础设施安全保护问题研究 [J]. 学习与探索，2024 (7)：76 – 86.
[11] 董柞壮. 数字货币、金融安全与全球金融治理 [J]. 外交评论，2022，39 (4)：133 – 155.
[12] 禹卫华. 生成式人工智能数据原生风险与媒介体系性规范 [J]. 中国出版，2023 (10)：10 – 16.
[13] 毕文轩. 生成式人工智能的风险规制困境及其化解：以 ChatGPT 的规制为视角 [J]. 比较法研究，2023 (3)：155 – 172.
[14] 胡泳，张文杰. 数据标注治理：可信人工智能的后台风险与治理转向 [J]. 云南社会科学，2024 (6)：29 – 36.
[15] 王燕. 超大型平台数字治理风险与欧盟法的应对 [J]. 国际经贸探索，2024，40 (2)：91 – 105.
[16] 张吉豫. 数字法理的基础概念与命题 [J]. 法制与社会发展，2022，28 (5)：47 – 72.
[17] 苏可桢，沈伟. 欧盟人工智能治理方案会产生“布鲁塞尔效应”吗？——基于欧盟《人工智能法》的分析 [J]. 德国研究，2024，39 (2)：66 – 88.

(收稿日期：2025 – 08 – 26)

作者简介：

盛小宝 (1979 –)，男，硕士，副研究员，主要研究方向：数据安全。

刘晋名 (1995 –)，男，硕士，研究实习员，主要研究方向：数据安全与合规。

姚迁 (1997 –)，女，硕士，研究实习员，主要研究方向：数据安全与合规。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部