生成式人工智能的内容安全风险：缘起、类型及其监管路径研究

郭佳楠[1,2]

（1.亚洲理工学院 发展与可持续性学院，泰国 空滦 12120；
2.中国人民大学 数字人文研究中心，北京 100872）

摘要：生成式人工智能技术以其强大的内容生成能力在各类信息生产场景中快速渗透，并深刻改变着数字内容的生成逻辑与传播机制。在释放巨大生产力的同时，其技术机制的不透明性、生成过程的不可控性以及内容输出的广泛性，也引发了前所未有的内容安全风险。立足生成式人工智能的发展现状，系统梳理其内容安全风险的多元缘起，聚焦模型层的数据训练与算法机制、应用层的主体行为虚置、内容层的技术异化与价值漂移，深入分析了虚假信息制造、信息污染扩散、算法偏见加剧与隐形权力嵌入等风险类型。溯源分析表明，风险的生成与扩散机制可从风险源生成、风险影响扩大、风险传播扩散等层面予以归纳。在此基础上，进一步提出从法律制度建设、分级分类监管、算法安全审计与用户素养提升等维度出发，构建适应性强、操作性高、前瞻性足的内容安全监管体系，以期实现生成式人工智能内容治理由被动响应向主动治理、由单一规制向系统协同的范式转型，从而确保技术赋能与社会安全的动态平衡，推动人工智能向善发展。

关键词：生成式人工智能；内容安全；风险类型；技术异化；算法监管

中图分类号：F49 **文献标识码：**A **DOI：**10.3969/j.issn.1003-8256.2026.03.011

随着人工智能技术的迭代升级与广泛应用，生成式人工智能（Generative Artificial Intelligence，GAI）作为新一轮科技革命的重要成果，正以前所未有的速度融入社会各个领域，重塑信息生产与传播范式。以大语言模型、图文生成系统为代表的生成式人工智能具备跨模态理解、语言生成、语境模拟等能力，不仅显著提升了内容创作的效率和表现力，也拓展了认知交互的边界。然而，技术赋能的同时也带来了前所未有的内容安全隐患，其生成内容的不可控性、不可预测性与不可追溯性，正成为数字时代信息治理面临的重要挑战。因此，研究生成式人工智能所带来的内容安全问题，厘清其风险类型与机制特征，探索其治理的制度逻辑与技术路径，既是维护国家网络安全和意识形态安全的现实需要，也是推动人工智能健康、可持续发展的重要保障。

生成式人工智能作为人工智能发展的重要分支，近年来因其强大的内容生成能力而受到学界与业界的广泛关注。尤其是在以大语言模型（如GPT系列、DeepSeek、Claude）为代表的技术推动下，生成式人工智能已广泛应用于新闻编辑、智能客服、教育辅助、舆情分析等场景。然而，随着其社会影响力的迅速扩大，内容安全问题日益突出，成为当前人工智能治理的重要研究议题。从国外研究来看，西方学界较早关注生成式人工智能所引发的伦理与安全风险。马塞尔·宾兹（Marcel Binz）等[1]指出，基于海量数据训练的大模型容易复制和放大其中的偏见与歧视，形成“有毒语言”的生成风险。尼古拉斯·斯蒂洛斯（Nikolaos Stylos）[2]进一步分析了生成式人工智能在虚假内容生成、操纵舆论、侵犯隐私等方面可能带来的社会危害，强调构建审慎、可控的发布机制与使用规范的重要性。此外，OpenAI、Google DeepMind等研究机构也逐步建立内容安全评估框架，尝试通过模型对齐、审查过滤等技术路径应对潜在风险。国内学界对该议题的研究起步稍晚，但发展迅速。在内容安全问题方面，研究主要集中在风险识别与规范治理两个方向。一类研究聚焦于内容生成机制与典型风险类型识别，鞠宏磊等[3]认为生成式人工智能存在“幻觉”内容、意识形态风险、恶意诱导等问题。另一类研究则聚焦于治理策略设计，乔喆[4]主张从法律、伦理、技术和平台自律等多个维度构建综合性治理体系，例如，提出“内容为核心、技术为手段、治理为保障”的三层逻辑治理框架；刘颖等[5]强调加强生成内容事前、事中、

基金项目：2021年度国家社科基金高校思政课研究专项（21VSZ031）；郑州工商学院科研创新项目（2023-KYZD-01）；2023年度郑州工商学院校级教育教学改革研究重点项目（GSJG2023003）

作者简介：郭佳楠（1991—），男，河北邢台人，亚洲理工学院发展与可持续性系博士研究生，中国人民大学数字人文研究中心学生研究员，研究方向：科学技术与公共政策。

事后的全过程监管,并建立统一的数据与算法审查制度。尽管已有研究从不同角度揭示了生成式人工智能内容安全问题的复杂性,但当前研究仍存在三个方面的不足:一是多数研究停留在“风险陈述”层面,缺乏对风险生成机制与扩散路径的深入剖析;二是法律、伦理、技术等治理路径之间的协同机制研究不足,缺乏系统性的整合分析;三是缺乏以中国治理语境为背景、结合平台生态与政策实践的本土化治理模式构建。上述不足表明,深入探讨生成式人工智能内容安全风险的结构性成因,并在此基础上提出可操作、可落地、可评价的综合监管路径,具有重要的理论价值与现实意义。基于此,本文拟在明晰内容安全风险内涵的基础上,从风险表现、诱发逻辑与扩散机制等维度切入,进一步探讨法律、技术、伦理三位一体的监管路径,力求为建立具有前瞻性、系统性与适应性的人工智能内容安全治理体系提供理论支持与实践方案。

1　生成式人工智能内容生成的技术机理

生成式人工智能依托于大规模神经网络模型的演进与算力资源的迅猛提升,在自然语言处理、图像生成、声音合成等领域展现出强大的内容生产能力。从技术架构来看,当前主流生成模型多基于Transformer结构,采用“预训练+微调”模式展开训练。在预训练阶段,模型通过对大规模文本语料的学习,掌握语言的词法结构与语义规律;而微调阶段则将模型应用于特定任务中,提升其语境匹配能力和生成质量。正如朗尼·李·胡德(Lonnie Lee Hood)[6]所指出:“生成模型的能力源于对语料世界的深度拟合,而非真正理解。”这说明了当前生成机制虽具语言仿真性,但其智能基础仍构建于“概率逻辑”而非“知识逻辑”之上。

同时,生成式人工智能内容生成的技术机理主要依赖于深度学习和自然语言处理等核心技术的协同作用。在实际运行过程中,具体生成流程主要包括输入解析、语义建模、序列预测与输出控制四个环节。模型首先接收用户输入(Prompt),通过Token嵌入技术将自然语言转化为可计算的向量表示;随后通过自注意力机制(Self-Attention)捕捉输入内部的语义依赖,进而形成上下文信息的深度建模;最后,基于最大似然估计策略,模型逐词预测可能的输出,形成连贯、逻辑一致的文本。以GPT为例,其“自回归”结构能够在每一步生成中融入历史信息,使内容更具延展性和语义连贯性。国内学者徐伟等[7]指出:“语言模型的生成能力实质是对人类语言经验的结构重组,是对语义可能性空间的技术性抽样。”因此,GAI的文本生成并非基于真理判断,而是基于高概率事件的排序组合。

尽管生成式人工智能的生成能力呈指数级增长,但其本质仍未摆脱“统计模拟”的范式局限。一方面,模型缺乏对事实真伪的判断能力,常产生所谓“AI幻觉”(hallucination),即逻辑通顺但内容失实的文本输出;另一方面,训练语料的偏差与噪声亦可能引发语言偏见、信息歧视等伦理问题。伊利亚·舒迈洛夫(Ilia Shumailov)等[8]直言:“大型语言模型的核心问题不在于它们不会说话,而在于它们不知道自己在说什么。”这一批评指出当前生成机制在认知层面上的空壳化风险。因此,在内容生成的同时,需通过对抗训练、人类反馈强化学习(RLHF)等机制进行结果调控,以提升模型输出的安全性、准确性与价值导向。

2　生成式人工智能的内容安全风险缘起

随着生成式人工智能技术的迅猛发展,其在文本撰写、图像创作、语音合成等领域展现出前所未有的内容生产能力,显著提升了信息生成的效率与多样性。然而,技术的跃迁也带来了日益突出的内容安全问题。从虚假信息的自动生成、敏感内容的扩散,到价值观的潜在误导和语料训练的偏见嵌入,生成式人工智能已从“工具性中立”的技术存在,转变为需要伦理规制与风险治理的复杂系统。在这一背景下,厘清生成式人工智能内容安全风险的成因机制,构建科学、系统的防控框架,已成为推动该技术健康发展的前提条件与治理重点。

2.1　模型层:数据训练与算法区分危及内容供给安全

生成式人工智能的模型层的安全风险首先源于其数据训练环节的内容结构与价值导向失衡。模型在大规模语料训练过程中,往往依赖互联网开放数据集,其中不可避免地包含虚假、偏激、低质或意识形态倾向性强的内容。这种“原始数据污染”会在深度学习的特征提取与参数拟合阶段被算法固化,进而以隐性方式影响生成内容的价值取向与语义导向。例如,一旦模型训练数据中存在偏向特定文化叙事或政治立场的文本,其生成结果便可能再现甚至放大这种偏向,削弱内容的公正性与可信度。此外,语料来源的地域差异与文化断层,使得模型在面对不同社会语境时易出现“语义失真”与“价值漂移”,在潜移默化中危及意识形态领域的安全。

此外,算法机制中的“区分逻辑”同样对内容供给安全构成潜在威胁。生成式模型在语言生成、图像合成等任务中,通过概率分布预测与语义优选机制生成内容,但这一机制本身建立在算法“区分”而非“理解”的逻辑之上。也就是说,模型并不具备真正的价值识别能力,而是在统计层面选择高频或高相关输出。这种算法层面的“形式理性”容易导致内容生成偏离真实语义或伦理规范,出现事实混淆、语义拼贴与逻辑失衡等问题。更为复杂的是,当模型在多轮优化中不断强化自身生成模式时,这种偏差会被反复放大并内化为模型参数的一部分,形成一种结构性的内容风险。因而,模型层的安全问题不仅是技术偏差的结果,更体现出生成式人工智能在语料、算法与价值三者之间尚未建立起稳定平衡机制的结构性隐忧。

2.2 应用层:主体虚置扰动内容生成的生态安全

生成式人工智能大模型的广泛部署,使“人人皆可创作”的技术理想成为现实,但也在实际应用中引发了“主体虚置”的安全隐患,即内容生产的责任边界模糊、行为动机不可识别、主体身份难以追溯。在传统传播生态中,内容创作者需承担明确的法律与伦理责任,媒体平台对信息发布拥有审校机制,内容流通具有“可定位性”。而在生成式人工智能应用场景下,创作者往往只是“触发者”,内容的实际来源是模型生成系统,这种“算法代理性”造成了责任主体的转移甚至失效,内容“去主体化”特征显著。例如,在B站、抖音等平台出现AI生成的拟真视频,利用换脸和合成技术制造与事实不符的画面,虽最终被澄清,但其传播链条中虚假信息扩散,反映出责任主体缺失带来的治理困境。

随着主体角色的日益弱化,主体虚置进一步演化为“生成泛滥”现象,严重干扰网络内容生态。生成式人工智能具备低成本、大规模、强表达能力,用户仅需输入关键词便可迅速产出文章、图像、评论,导致社交媒体、知识平台、新闻评论区充斥着“模板化”“伪原创”内容。这种“内容复制化”与“情绪煽动型生成”的泛滥,降低了公共信息质量,压缩了真实创作者的表达空间。李旭光等[12]指出,AI生成内容已广泛应用于虚假新闻撰写与网络舆论操纵,形成了一种“深度自动化信息污染”现象。而当生成内容被伪装为“真实用户发声”,并通过分发算法快速传播时,便可能制造“信息茧房”、加剧社会认知撕裂,对舆论治理和民主协商秩序构成长期干扰。

2.3 内容层:“技术异化”的辐射引发广泛的社会域风险

随着技术异化现象的扩展,其在传播层面的渗透正逐步动摇社会的认知框架和文化价值基础。一方面,AI生成的内容高度拟真且快速迭代,已突破了传统信息筛选和辨别能力的承受范围。尤其在灾难事件、国际冲突、公共健康等舆论敏感领域,大模型生成的不实信息、偏颇叙事或情绪煽动内容,极易激化公众情绪、扭曲群体判断。例如,2023年在以色列与巴勒斯坦冲突报道中,社交平台上广泛流传的生成式假视频与图像,曾引发多起误判和公众恐慌①。其根源即在于算法对内容进行“拟态生成”,而非基于真实情境再现,从而重塑了公共事件的叙事逻辑。

另一方面,“技术异化”还通过平台算法与商业模式的绑定机制,加剧了内容的消费主义与碎片化倾向。在商业逻辑驱动下,生成式AI更多以“用户偏好最大化”为目标进行内容生产,迎合注意力经济与情绪化传播,从而催生“流量为王”“热点优先”的内容生态。这种趋势不仅压缩了严肃内容的生存空间,也导致公众对“知识”与“价值”的认知发生变异,形成“快读即真”“算法即权威”的新型信息偏见。例如,OpenAI前研究员保罗·克里斯蒂安(Paul Christiano)指出:“未来数年内,AI生成内容将在网络上远远超越人类内容,而我们甚至难以察觉两者的边界[9]。”当真实与虚构、理性与煽动、深度与浅表难以分辨,社会的认知秩序与文化信任便陷入结构性危机。

3 生成式人工智能内容安全风险的主要类型

随着以ChatGPT、DeepSeek等为代表的生成式人工智能大模型迅速迭代,其在文本、图像、音频、视频等内容创作中的广泛应用,为信息传播方式带来了深刻变革。然而,伴随着技术边界的持续拓展,其所引发的内容安全风险问题日益凸显,逐渐演化为人工智能治理中的核心议题。当前的研究表明,这些风险可被划分为多个典型类型,既包括直接危及用户认知与信息真实性的“虚假内容风险”,也包括挑战社会伦理底线与公共安全的“有害信息传播风险”,还涉及算法操控、情绪煽动、数据污染、信息歧视等深层次问题②。因此,系统梳理生成式人工智能内容安全风险的主要类型,对于厘清风险成因、识别治理重点、构建科学的技术监管与伦理干预体系,具有重要的理论价值与现实意义。

3.1 批量制造虚假信息的风险

生成式人工智能大模型在文本生成、图像合成、语音克隆等方面的高度逼真性和低门槛应用,使其逐渐成为虚假信息制造的“新引擎”。不同于传统造假行为依赖人工编辑与媒介加工,生成式人工智能通过深度学习模型对海量数据的训练,可以在数秒内生成具备高度语义连贯性和风格模仿能力的内容,极大提高了虚假信息的生产效率和“伪装性”。以ChatGPT为例,在未设定限制条件的测试环境中,其可在用户提示下迅速编造新闻、虚构人物评论、制造学术引证内容,并以正式语言组织结构,欺骗性极强。据《麻省理工科技评论》2023年发布的数据,在一项“AI生成vs人类写作”的辨别实验中,有近48%的受访者未能准确识别由AI生成的虚假新闻标题,表明生成式内容正呈现“假中有真、真中有假”的高度混淆状态③。

从操作层面看,虚假内容的生成与传播已形成一条“产业化”路径。例如,2025年“AI伪造邓紫棋视频”事件④,某商家利用深度合成技术生成虚假视频内容,

① 资料来源:巴以冲突的“信息战”,别被这些消息误导了.https://www.163.com/dy/article/IH82KM280514R9P4.html.

② 资料来源:Generative AI Security Risks:8 Critical Threats You Should Know.https://keepnetlabs.com/blog/generative-ai-security-risks-8-critical-threats-you-should-know.

③ 资料来源:Weird rat,AI Fake Science-trust the what? https://www.theaioptimist.com/p/weird-rat-ai-fake-science-trust-the.

④ 资料来源:邓紫棋被AI视频祝福开业大吉,后援会发声:禁止任何形式的AI邓紫棋.https://www.yangtse.com/news/wy/202510/t20251007_273671.html.

并在网络上进行大规模传播,致使公众产生误解,严重损害了艺人名誉与个人形象。类似的“AI换脸”“AI合成声音”也常被用于制造虚假新闻或色情内容,侵害当事人的人格权与隐私权。事件虽被快速辟谣,但造成的恐慌与社交恐惧效应却未能完全消除。这种“内容+媒介+算法”协同放大的造假模式,使AI生成的虚假信息不再局限于某一类型文本,而是具备了跨媒体、跨平台、跨语境的快速扩散能力,严重冲击信息传播的可信性基础。

3.2　大规模“信息污染”的风险

生成式人工智能所引发的信息污染问题,正在突破传统媒介污染与信息冗余的边界,呈现出内容生成无限膨胀、语义价值高度稀释、信息环境系统性退化等复杂化风险。所谓“信息污染”(Information Pollution),是指信息内容的大量冗余、低质、失真、失序、重复和干扰,致使个体在信息识别、理解、判断与决策中遭遇障碍。与传统信息污染多由人工生产、重复发布所致不同,生成式人工智能以其近乎无限的内容生产能力,将这一问题放大为结构性、生态性挑战。

当前,以大语言模型(如ChatGPT、Claude、DeepSeek等)为代表的生成式人工智能,其内容生成已突破“有问有答”的被动模式,向自主生成、自我学习、自我演化方向迈进。尤其是在搜索引擎、内容分发平台、社交媒体等场景中,生成内容不仅混杂于人类生产的自然语言文本中,而且常常被算法推荐系统优先推送,形成内容“堆叠”与“劣币驱逐良币”的效应。在丹尼斯·任(Dennis Ren)等开展的调研中,超过72.5%的人认为被推荐的研究摘要符合伦理道德,他们只能依靠直觉和猜测来做出判断,因此难以判断其背后是由人类创作还是AI生成[10];而一旦AI生成的内容被采纳并反哺至公共知识平台(如维基百科、知乎等),将进一步加剧“信息失真-信任危机-平台滑坡”的负面循环。此种“低质泛滥-主流稀释-知识污染”的信息生态污染模式,不仅削弱了公众认知的基础能力,也威胁着数字公共领域的理性秩序。

更为严峻的是,生成式人工智能生成内容在多语言适配、多平台传播与多模态融合的协同驱动下,正逐步演化为具有跨地域、跨语境特征的全球性信息污染源,对国际舆论环境与认知安全构成系统性挑战。以AI自动生成营销、评测、教学、咨询类内容为例,一方面推动了“内容农场”式伪知识扩张;另一方面则出现大量同质化、语义空洞的“水文式”文本充斥搜索结果。2024年4月,谷歌宣布将针对由AI生成的“低质量SEO优化内容”进行新一轮算法清洗,彼时检测出近40%的网页存在AI痕迹,且其中85%以上内容存在“信息重复、数据脱节、逻辑混乱”等问题[11]。此外,视觉生成AI(如Midjourney、Sora)所制作的图像和视频内容,也因缺乏可信标签或验证机制,频频成为虚假图像、伪科学展示和不实引用的主要来源。例如,2023年,国内某电商平台曾曝出利用AI生成商品评论和虚假用户反馈的现象。大量由生成式人工智能自动撰写的好评、测评与用户体验文章,在短时间内充斥平台,不仅使消费者难以辨别真实信息,还扭曲了市场的正常竞争秩序。

3.3　算法偏见带来的认知风险

在人机交互的过程中,看似客观中立的回答中其实早已渗入了大量原始数据所拥有的理念与价值观,这些数据所携带的价值观决定了什么样的内容被看见,什么样的内容被隐藏。具体来看,一方面,由于生成式人工智能高度依赖大规模语料进行模型训练,而此类语料本身通常嵌套着历史叙事、文化立场与社会结构等固有因素,因此不可避免地携带某种程度的历史性、文化性乃至社会性的偏见,使模型在内容生成过程中潜在地复制甚至放大原始语料中的意识形态倾向与价值判断。这种算法偏见一旦大量生成内容并作用于社会实践,极易造成认知偏差的扩散,导致使用者潜移默化地接受某种“默认设定”,加剧了数字空间中的结构性不平等。

另一方面,算法偏见还体现在认知结构的单一化与语义生成的路径依赖上。大模型往往基于概率统计与相似性嵌入机制生成内容,即“越常见的组合越容易被生成”,导致非主流叙事、边缘话语、个体经验被系统性弱化甚至抹除。这种机制虽然提升了语言生成的流畅度与语法正确性,但也强化了文化与知识的“平均主义”逻辑。以高校思想政治教育为例,若在设计AI自动讲稿、思政问答机器人或政策解释系统时,模型训练数据主要来源于官方文献与主流媒体,则在输出中可能过度强调“政治正确性”,忽视了学生的真实关切、历史争议话题和多样化表达方式。久而久之,教育者与学习者都可能陷入“同质认知回路”,丧失思辨能力、批判精神与对复杂现实的感知能力,从而弱化了思想政治教育本应具有的多元视角、理性表达与公共讨论功能。

3.4　隐形权力生成的风险

生成式人工智能在重新定义信息生产方式的同时,也悄然重构了权力的分布与运作逻辑,其所衍生的“隐形权力”不同于传统社会结构中明晰的行政权、法律权或市场权,而是一种以算法为中介、以技术系统为通道、以内容生成为手段的“看不见的权力”。这一权力通过内容选择、语义建构与行为诱导等机制潜移默化地影响着用户的知识认知、价值判断与社会行为,进而对公共领域中的言说空间、意识形态安全乃至民主治理模式构成深远挑战。

无论国内还是国外,生成式人工智能平台大多依赖由技术企业构建的封闭性数据体系和内容呈现规则,用户所接触的信息、知识甚至事实都由算法“代劳”过滤与再组织,导致信息接收的单向性与去批判性。这种由系统控制的“信息过滤”并非中立过程,而是隐含了平台

方、模型训练者甚至投资方的利益考量与意识形态偏好。例如，当模型在输出关于社会议题、政治理念或文化价值观的内容时，往往会自动优先呈现“安全”“稳定”“主流”的选项，而将边缘话语、另类观点“算法性消音”。这一操作在形式上看似“理性中立”，实际上却造成了认知资源分配的非对称性，让个体无法充分理解复杂性现实。此外，语义生成的自动化进一步扩展了这一隐性控制力。大模型所生成的文本不仅内容精准、逻辑通顺，而且具备语言权威感与技术中立性，其输出往往被用户默认为“更专业”“更可信”。然而，这种技术中立的表象掩盖了其背后的话语选择与价值编码。

4 生成式人工智能内容安全风险的生成与扩散机制

生成式人工智能的内容安全风险并非孤立出现，而是内嵌于其技术逻辑、应用场景与社会结构之中的动态过程。因而，生成式人工智能的内容风险表现为由点到面、由内而外、由技术到社会的链条式演化，其生成与扩散机制需要从多维度进行系统性分析，以揭示其深层逻辑与治理难题。

4.1 风险源生成：数据和算法不确定的内生局限性

生成式人工智能的内容安全风险首先源于其技术本身的不确定性。大模型的性能依赖于海量数据与复杂算法，而数据的采集与处理往往存在偏差与噪音。例如，训练数据中隐含的刻板印象、虚假信息或错误标注，会在模型生成内容时被放大并持续复制，形成“数据放大效应”。同时，深度神经网络的高度复杂性使得模型难以解释与预测，其输出结果具有概率性和不稳定性。这种“黑箱化”特征意味着即便开发者在输入端设置了严格的规则，也难以确保生成内容的绝对安全与可控，从而使风险在源头阶段就具备了潜在的内生性。

另一方面，技术开发者的价值观与商业逻辑也会成为风险生成的重要环节。开发者在算法设计与模型调优过程中，往往追求效率与性能最大化，而忽视了对价值导向和伦理责任的平衡。部分企业在激烈的市场竞争中，为抢占先机而弱化风险预防机制，使得模型的开发带有明显的功利化倾向。这种观念异化导致技术成果更多服务于资本逻辑与流量逻辑，而非公共利益和社会安全。一旦开发者在训练语料选择、内容审核规则制定等环节缺乏责任意识，模型生成的内容极易出现虚假化、低俗化甚至极端化，从而在无形中加剧了风险源的生成。

4.2 风险影响扩大：技术创新中伦理审查缺位及技术伦理和向善文化缺失

生成式人工智能的高速迭代，使得伦理规约在实际应用中往往出现失效局面。在我国，人工智能伦理指引虽然已陆续出台，如《新一代人工智能伦理规范》等，但在落地执行环节仍存在抽象化、原则化的问题。部分企业在实践中仅将伦理规范作为“合规展示”的工具，而非内生的研发准则，导致伦理约束在技术创新中的“纸面化”。例如，深度合成技术在短视频与社交平台的泛滥应用，往往缺乏对真实性和公共安全的有效把关，虚假视频、合成语音等内容迅速扩散。这表明，伦理规约在面对强大技术驱动力时往往被边缘化，导致技术发展的“价值中立化”倾向不断增强，使潜在风险逐渐积聚并扩散。

同时，机构监管的失能与法律约束的失范同样是风险影响扩大的重要原因。近年来，我国已在国家层面设立相关监管机制，并出台如《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等制度性文件，但在实践中，监管仍面临技术复杂性与实施可操作性不足的双重挑战。地方监管部门在专业人才储备、技术检测能力、跨部门协同等方面存在明显短板，导致监管往往滞后于技术创新。除此之外，法律约束的失范是风险扩散的制度性根源。虽然我国在人工智能治理中逐渐引入法律规制思路，但相关法律条文仍以原则性表述为主，缺乏对具体应用场景和责任分配的细化规定。在技术跨平台、跨领域应用的复杂环境下，这种制度安排的模糊性使得开发者、平台与用户之间的责任边界难以明确，形成事实上的“责任真空”。

4.3 风险传播扩散：社交媒体、移动终端等传播媒介扩大了内容安全风险的传播范围和破坏力

在生成式人工智能内容安全风险的扩散过程中，社交媒体的加速作用尤为显著。作为信息传播的主要平台，社交媒体具有低门槛、强互动和高扩散等特征，使得人工智能生成的内容在极短时间内实现规模化传播。与传统媒体相对稳定的编辑与把关机制不同，社交媒体的信息流主要依赖于算法推荐与用户分享，其逻辑往往偏向于“流量最大化”，而非“真实性优先”[12]。这种传播机制放大了虚假、失真乃至恶意信息的扩散速度，使得风险在极短时间内实现跨平台、跨人群的迅速扩展。同时，社交媒体的社群化特征与“信息茧房”效应进一步加剧了风险的聚集性与叠加性。

同时，移动终端则在风险扩散中提供了高频、便捷的传播载体。随着智能手机、平板电脑等移动设备的普及，用户可以随时随地接触、生产和传播由人工智能生成的内容。这种“全天候、全场景”的信息接触方式，使得风险传播突破了时间与空间的限制，形成了高度碎片化与即时化的扩散格局。在我国，移动终端与社交媒体的高度绑定更是放大了风险的外溢效应。用户在移动端的高黏性使用习惯，使得AI生成内容通过即时通讯、短视频和多功能应用程序实现无缝传播，形成多渠道、多层级的叠加传播链条。此外，移动终端的个性化推送与精准推荐机制，使得风险信息能够以“定制化”的形式触达不同群体，导致风险在扩散过程中不仅范围更广，

而且影响更深。这种高度依赖移动终端的传播方式,使得内容安全风险具备了持续性与隐蔽性,最终对社会治理、公共安全乃至国家信息安全带来更为复杂和深远的挑战。

5　保障生成式人工智能产品内容安全的监管路径

人工智能治理需要在技术进步与防范失控性风险中寻求平衡。科学监管是促进生成式人工智能创新发展与风险控制的重要内容。监管路径的探索应充分发挥中国特色社会主义制度优势,深入贯彻党的二十大和二十届二中、三中全会精神,严格遵守相关法规政策要求。由此,面对生成式人工智能内容生成特征的高度自主性、复杂性与不可预测性,亟须在法律法规、行业自律、平台责任与技术手段等多维度展开有效的监管路径设计,以实现对内容安全的前瞻性预警、过程性干预与系统性治理,为技术健康发展与社会稳定提供制度保障。

5.1　持续完善人工智能基础性法律法规,使内容安全监管具有延续性和适应性

生成式人工智能的迅猛发展催生了内容生产与传播范式的深刻变革。与传统媒介相比,其内容生成具有高度自动化、规模化、即时性与跨域性的特征,使得风险传播更为隐蔽、扩散更为迅速。为实现对生成式人工智能内容安全的有效管控,亟须推动法律规范从“点状规制”走向“体系治理”,从“滞后反应”迈向“动态适应”。

具体来看,健全生成式人工智能法律法规体系,首要任务是推进基础性立法,确立人工智能产品内容生成的权责边界与行为规范。应加快制定《人工智能法》《算法法》《智能内容管理法》等综合性立法,将算法生成内容纳入法定监管范畴,明确生成者、开发者、平台方及使用者之间的法律责任划分。同时,法律制度的完善不仅需要明确规则,还需强化与伦理规范、行业自律及社会治理的协同。在中国治理体系下,推动生成式人工智能内容安全的法律监管,应将社会主义核心价值观与技术治理相结合,以“向善文化”引导技术发展方向,避免法律的刚性约束陷入单一的工具化逻辑。另外,还可以通过建立国家层面的人工智能伦理审查与评估机制,将伦理规范作为法律实施的重要参照,使企业在开展技术创新时必须进行伦理评估和安全影响预测。

5.2　细化生成式人工智能产品分类分级监管标准,平衡内容安全和技术创新

在生成式人工智能快速发展的背景下,如何在确保内容安全的同时激励技术创新,已成为我国治理语境下面临的重要课题。在我国人工智能治理框架下,这一监管路径可从明确分类维度、设立分级标准与完善配套监管机制三个方面展开。

在分类维度上,可以依据技术类型(文本生成、图像合成、视频生成、语音合成等)、应用场景(教育、娱乐、金融、医疗、政务等)、潜在风险(涉及国家安全、意识形态、社会稳定、个人隐私等)进行综合划分,使不同类别的人工智能产品纳入差异化的监管路径。在分级标准设立方面,必须结合我国治理实践,形成“风险导向、动态分级”的监管逻辑。具体而言,可以将生成式人工智能产品划分为低风险、中风险与高风险三个层次,依据产品功能、用户规模与社会影响力进行动态评估。另外,配套监管机制的完善是分类分级监管有效落地的关键环节。从我国人工智能治理实践出发,建立“法律规制-行业自律-社会监督-技术赋能”相结合的多元共治格局[13]。法律规制方面,应通过专门立法或配套规章明确分类分级监管的基本框架与法律责任,确保制度有法可依;行业自律方面,可以推动人工智能企业和产业联盟制定细化的自律标准与技术规范,在法律框架下形成自我约束机制;社会监督方面,应引导公众、媒体与学术机构参与监管评价,提升透明度与社会信任;技术赋能方面,需借助人工智能本身的能力,开发智能化监测与风险预警系统,提升监管的实时性与精准性。

5.3　完善生成式人工智能产品的算法备案和安全审计标准和流程,加强政策工具的系统化

生成式人工智能作为驱动新一轮产业革命的关键技术,其背后的算法逻辑、训练机制和应用模式往往具有黑箱化、复杂化和动态迭代的特征。这种特性虽然促进了技术的迅猛突破,但也在数据来源、算法设计和内容输出等环节潜藏着安全风险。要实现生成式人工智能的健康发展,就必须在算法备案和安全审计的标准、流程与政策工具的系统化衔接方面进一步强化制度设计,推动形成统一、透明、可操作的监管体系。

在算法备案方面,关键在于实现备案内容的规范化、备案流程的标准化和备案管理的动态化。一方面,应建立国家级算法备案信息库,将备案信息按照风险等级进行分层管理,并实现与地方监管平台的互联互通,避免出现“信息孤岛”。另一方面,还应推动“备案即核验”机制,即在备案过程中同步启动技术核验程序,由专业评估机构对算法的公平性、安全性和内容生成的合规性进行抽样检测,确保备案信息的真实性。在安全审计方面,重点是建立权威统一的审计标准、健全分层次的审计机制,并推动审计结果与政策工具之间的系统衔接。在实践操作中,应建立“企业自查-第三方独立审计-政府抽查”三重机制,形成多维度的合规验证体系。同时,应推动审计结果与政策工具的制度化对接,例如将审计报告纳入企业信用体系,作为市场准入和政府采购的重要参考条件;将审计结论与平台治理挂钩,对存在严重违规的产品实行限流、下架或联合惩戒。在政策

工具系统化方面，应着力实现行政监管、行业自律和社会监督的有机整合，构建全方位的治理格局。从行政监管角度看，应推动跨部门联动，建立网信、工信、公安、广电等部门的协调机制，形成信息共享和协同执法模式，避免出现监管空档或重复执法。在行业自律方面，应发挥行业协会和头部企业的示范效应，推动制定统一的行业标准和自律公约，倡导透明、公平与向善的技术文化。

5.4 强化对终端用户人工智能素养和行为的规范教育，确保交互性内容安全

在生成式人工智能快速普及的背景下，终端用户已经成为交互性内容安全的关键变量。人工智能系统通过人机互动不断生成、扩散和调整信息，用户的认知水平、价值取向与使用习惯直接决定了技术能否在可控范围内发挥积极作用。在我国，强化人工智能素养教育不仅是单纯的技术使用培训，更是培育数字文明的过程，旨在帮助用户理解人工智能的能力边界与潜在风险，进而在日常交互中形成责任意识与自觉行为，从根本上降低内容安全隐患。

推进用户人工智能素养的提升，需要从多层次、多环节系统展开。教育体系应将人工智能素养纳入终身学习框架，通过中小学、大学到职业教育的贯通式培养，建立起技术认知、伦理意识与批判性思维的完整体系。此外，行为规范的构建同样是确保交互性内容安全的重要环节。结合我国人工智能治理实践，用户行为的管理不仅依赖法律法规的强制性约束，更需要通过社会规范与平台责任实现多维协同。政府部门可以推动制定针对终端用户的《人工智能交互行为准则》，对滥用生成式人工智能制造虚假内容、侵犯他人权益等行为予以明确界定和惩戒，从法律制度上提供底线保障。同时，平台应强化算法推荐与用户行为的联动约束，建立内容生成与交互的实时监测机制，对违规行为进行精准干预和分级处置，形成“事前预防、事中监控、事后问责”的闭环管理模式。

6 结论与启示

生成式人工智能的迅速发展在推动技术创新和产业升级的同时，也带来了显著的内容安全风险。尤其是立足于我国国家治理体系与结构下，内容安全不仅关乎网络空间的清朗秩序，更直接涉及国家意识形态安全、社会治理效能以及公众利益保护。内容安全不再是传统意义上的“技术附属问题”，而是关涉治理体系现代化、文化安全与社会稳定的重要议题，亟须纳入人工智能整体治理战略的核心层面。通过从技术、应用与内容三个层级系统梳理了生成式人工智能内容安全风险的生成逻辑，厘清了算法模型、用户行为与生成机制等多因子耦合互动下的风险链条，揭示了虚假信息制造、认知干扰、伦理侵蚀与规避监管等多维风险类型的特征、机制与影响。同时，在风险生成与扩散机制上，不仅涉及技术内生局限、开发者观念异化和技术与社会的单向互动，也与伦理审查缺位、机构监管失能及法律约束不足等制度性因素相关。社交媒体和移动终端等新兴传播渠道的广泛应用，更加速了风险的扩散和影响范围，使内容安全管理面临更高的不确定性和复杂性。由此，应对生成式人工智能内容安全风险，通过结合国内外监管经验与发展趋势，并结合我国实际情况采取多维度治理措施，进一步从法律法规、监管标准、技术审计、用户教育等维度提出系统性应对策略，强调构建包容审慎、动态适配、多元协同的监管机制，推动内容安全治理从“事后补救”向“事前预防”与“过程嵌入”转型。可以预见，生成式人工智能仍将以指数级速度演进，其内容能力的边界尚难预测，而治理能力的提升却远未同步。未来，应持续推进跨学科交叉研究，加强伦理法规与技术发展的同步耦合，构建具有中国特色、全球视野的生成式人工智能内容治理体系，真正实现技术进步与人类福祉的双重跃迁。

参考文献：

[1] BINZ M, SCHULZ E . Using cognitive psychology to understand GPT-3[J].Proceedings of the National Academy of Sciences of the United States of America, 2023, 120(6): e2218523120.

[2] STYLOS N, OKUMUS F, ONDER I. Beauty or the borg: agentic artificial intelligence organizational socialization in synergistic hybrid transformative dynamic flows[J].Tourism Management, 2025(111): 33-35.

[3] 鞠宏磊，申欣雨.生成式人工智能产品的内容安全风险及监管路径[J].经济问题，2023(12)：16-19，27.

[4] 乔喆.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[J].电信科学，2023，39(10)：136-146.

[5] 刘颖，卢超男.生成式人工智能引发的内容风险及治理对策[J].机器人产业，2024(5)：14-20.

[6] LONNIE LEE HOOD. Experts say that soon, almost the entire internet could be generated by AI[EB/OL].(2023-03-05)[2025-07-05].https://futurism.com/the-byte/ai-internet-generation.

[7] 徐伟，韦红梅.生成式人工智能训练数据风险治理：欧盟经验及其启示[J].现代情报，2025，45(5)：89-98.

[8] SHUMAILOV I, SHUMAYLOV Z, ZHAO Y, et al. The curse of recursion: training on generated data makes models forget [J]. ArXiv, 2023, abs/2305.17493.DOI: 10.48550/arXiv.2305.17493.

[9] OUYANG L, WU J, XU J, et al.Training language models to follow instructions with human feedback[J].Neural Information Processing Systems, 2022(35): 27730-27744.

[10] REN D, TAGG A J, WILCOX H, et al. Identification of human-generated vs AI-generated research abstracts by health care professionals[J].JAMA Pediatr, 2024, 178(6): 625-626.

[11] 腾讯云.谷歌宣布清理搜索结果中的AI生成垃圾[EB/OL].(2024-03-27)[2025-10-11].https://cloud.tencent.com/developer/news/1323124.

[12] 李旭光，胡奕，王曼，等.人工智能生成内容研究综述：应用、风险与治理[J].图书情报工作，2024，68(17)：136-149.

[13] 苏辰阳.舆情风险管控视角下人工智能生成内容可信度评价研究[D].杭州：浙江大学，2024.

Content Security Risks of Generative Artificial Intelligence: Research on Origin, Types and Regulatory Paths

GUO Jianan[1,2]

(1.Department of Development and Sustainability, Asian Institute of Technology, Kong Luan 12120, Thailand;
2.Digital Humanities Research Center, Renmin University of China, Beijing 100872, China)

Abstract: Generative artificial intelligence technology has rapidly penetrated into various information production scenarios with its powerful content generation capabilities, and has profoundly changed the generation logic and dissemination mechanism of digital content. While releasing huge productivity, the opacity of its technical mechanism, the uncontrollability of the generation process and the extensiveness of content output have also caused unprecedented content security risks. Based on the current development status of generative artificial intelligence, this paper systematically sorts out the multiple origins of its content security risks, focusing on data training and algorithm mechanisms at the model layer, the virtualization of subject behavior at the application layer, and the technological alienation and value drift at the content layer. It deeply analyzes the risk types such as false information creation, information pollution diffusion, algorithm bias intensification and invisible power embedding. Tracing back analysis shows that the risk generation and diffusion mechanism can be summarized from the aspects of risk source generation, risk impact expansion, and risk transmission and diffusion. On this basis, it further proposes to build a content security regulatory system with strong adaptability, high operability and sufficient foresight from the dimensions of legal system construction, hierarchical and classified supervision, algorithm security audit and user literacy improvement, so as to achieve the paradigm transformation of generative artificial intelligence content governance from passive response to active governance, and from single regulation to system coordination, so as to ensure the dynamic balance between technology empowerment and social security and promote the good development of artificial intelligence.

Keywords: generative artificial intelligence; content security; risk type; technological alienation; algorithmic supervision

(上接第 72 页)

The Dynamic Evolution and Integrative Framework of Thriving at Work Research

LI Zhengdong, DENG Anru

(School of Humanities, Shanghai Institute of Technology, Shanghai 201418, China)

Abstract: With the rapid acceleration of global digitalization, workplace challenges have become increasingly pervasive. Thriving at work, as a positive psychological state, helps mitigate job burnout and enhance organizational effectiveness, emerging as a cutting-edge topic in organizational behavior that has garnered widespread academic attention in recent years. Based on bibliometric methods and knowledge mapping techniques, this study systematically analyzes thriving-at-work research literature from the China National Knowledge Infrastructure (CNKI) database (2013—2024), revealing the developmental trajectory of this field through dimensions such as research hotspots, current status, evolutionary trends, and framework integration. The findings indicate that: (1) Current research has shifted from conceptual definition and measurement tool development to in-depth exploration of multi-level influencing factors and mechanisms, with a focus on the interplay of individual traits, leadership behaviors, and organizational environments; (2) Thriving-at-work research is in a phase of rapid expansion, with an exponential increase in publications in recent years. However, fragmented collaboration among core authors and insufficient institutional cooperation suggest significant untapped potential in this field; (3) The evolutionary trajectory exhibits four distinct phases: from conceptual construction to mechanism exploration, then multidimensional expansion, and finally contextual deepening. Research hotspots are increasingly shifting toward emerging management practices such as job crafting and shared decision-making, reflecting a dual emphasis on theoretical refinement and practical application; (4) Future research should prioritize five key directions: cross-cultural comparative studies, influence pathways of emerging leadership models, the double-edged sword effect of thriving at work, occupational heterogeneity mechanisms, and longitudinal tracking designs. This study provides valuable insights for advancing thriving-at-work research and offers important references for subsequent theoretical innovation and practical applications.

Keywords: thriving at work; dynamic evolution; influence mechanism; integrative framework; knowledge mapping