眼见未必为实

——人工智能生成内容引发的社会信任风险及其化解策略

Seeing May Not Be Believing:
Social Trust Risks Caused by Artificial Intelligence Generated Content and Their Mitigation Strategies

王张华 /WANG Zhanghua　李成成 /LI Chengcheng　曾玉芝 /ZENG Yuzhi

（湘潭大学公共管理学院，湖南湘潭，411105）
(School of Public Administration, Xiangtan University, Xiangtan, Hunan, 411105)

摘　要：社会信任是构建和谐人际关系和推动社会进步的基石。然而，在人工智能时代，人工智能生成内容（AIGC）潜藏着多重社会信任风险，严重威胁社会信任体系的稳定性，导致“眼见未必为实”的现实困境。基于技术－社会互构理论，遵循“风险源头－风险触发－风险放大－风险扩散”的逻辑框架，深入剖析AIGC引发的社会信任风险的生成机制，提出应从法治、技治、众治三个维度综合施策，以化解AIGC带来的社会信任危机。期望在人工智能时代，重新构筑坚实的社会信任体系，促进人与人、人与技术、政府与公民之间建立更加紧密的信任纽带。

关键词：人工智能生成内容　社会信任　深度伪造　信任风险

Abstract: Social trust is the cornerstone of building harmonious interpersonal relationships and promoting social progress. However, in the era of artificial intelligence, Artificial Intelligence Generated Content (AIGC) carries multiple social trust risks, seriously threatening the stability of the social trust system and leading to the reality of “seeing may not be believing”. Based on the theory of technology-society interaction, this article follows the logical framework of “risk source, risk trigger, risk amplification, risk diffusion”, deeply analyzes the generation mechanism of social trust risk caused by AIGC, and proposes comprehensive measures from three dimensions: rule of law, technology governance, and crowd governance to resolve the social trust crisis caused by AIGC. We hope to rebuild a solid social trust system in the era of artificial intelligence, and promote the establishment of closer trust bonds between people, people and technology, and government and citizens.

Key Words: Artificial Intelligence Generated Content; Social trust; Deep fake; Trust risks

中图分类号：TP18; C01　DOI: 10.15994/j.1000-0763.2026.07.011　CSTR: 32281.14.jdn.2026.07.011

ChatGPT、DeepSeek等生成式人工智能大模型的接连问世，引发了社会各界对人工智能生成内容（Artificial Intelligence Generated Content, AIGC）的高度关注和热烈讨论，主

基金项目：国家社会科学基金青年项目“数字治理视域下政府与平台型企业合作模式选择及其风险管控研究”（项目编号：22CZZ037）；湖南省学位与研究生教学改革研究重点项目“ChatGPT类生成式人工智能应用于研究生科研场景的负面效应与应对策略”（项目编号：2023JGZD037）。
收稿日期：2025年6月14日
作者简介：王张华（1990–）男，湖南株洲人，湘潭大学公共管理学院副教授，研究方向为数字公共治理、人工智能治理。Email: wangzhanghua@xtu.edu.cn
李成成（2001–）男，河南邓州人，湘潭大学公共管理学院硕士研究生，研究方向为数字公共治理、人工智能治理。Email: alluringcc@163.com
曾玉芝（1991–）女，湖南株洲人，湘潭大学公共管理学院博士研究生，研究方向为公共数据治理。Email: 317254287@qq.com

要涵盖视频、音频、图像、文本四种内容类型。AIGC通过运用神经网络、生成扩散模型和大型预训练模型等人工智能技术，探寻现有数据的规律，重塑内容生成方式，兼具内容与技术的双重特性。[1]生成式AI的迅猛发展和广泛应用正深刻改变着人们的生活、学习与工作方式，但其高度仿真性背后也潜藏着降低社会信任的风险。[2]随着人工智能生成技术在社会生活中的广泛渗透，算法黑箱、数据局限、AI“幻觉”、技术滥用等缺陷逐渐显现。公众在享受技术红利的同时，也开始怀疑所接触的视频、音频、图像和文本的真实性，“眼见未必为实”的疑云笼罩着人们的工作、学习与生活，长此以往可能导致社会信任的坍缩，增加社会运行成本和公共治理难度。需要注意的是，公众对AIGC的不信任可能扩展至社会层面生成式人工智能相关产品或服务，引发系统性的社会信任坍缩风险。社会信任是经济发展与社会稳定的前提和基础，如何有效化解AIGC引发的社会信任风险、重塑人工智能时代的社会信任，对于推动经济发展和社会稳定至关重要。

本文以邱泽奇等人提出的技术–社会互构理论为理论基础，[3]强调技术和社会并非孤立存在，而是相互嵌套的整体，技术设计需要与社会结构相匹配。辩证地看，生成式AI的应用发展既受经济环境、政治制度、文化习俗等社会因素的深刻影响，又将反过来成为重塑与规范社会运行的变革力量，助推社会发展需求的实现。由此可见，AIGC引发的社会信任风险本质上是源于社会结构与技术设计之间的脱节。生成式AI的广泛应用确实为社会发展带来了驱动力，但其发展迭代速度却远超社会的调适与规范能力，有赖于强制手段的有效干预。技术–社会互构理论要求技术发展必须与责任伦理深度融合、技术与社会的协同演进以及协同治理。[4]为此，从技术与社会的互构逻辑出发探讨AIGC引发的社会信任风险的生成逻辑和治理路径。

一、人工智能生成内容引发的社会信任风险的表现与特征

社会信任风险是指一定社会或群体的道德准则和规范不被人们所遵守，人与人之间缺乏一种道德的联系和约束。[5]社会信任风险由来已久，只要信任关系还存在，信任风险就存在，并且有转化成实际损失的可能。AIGC引发的社会信任风险是指AIGC应用于社会各领域、各行业的过程中，由于AIGC技术缺陷或不当使用而引发的一系列社会信任问题，将蔓延至政治、经济、社会等多个领域。

1. 人工智能生成内容引发的社会信任风险的表征样态

第一，消解政府信任：政府权威形象受损。政府信任是公众与政府之间的一种互动关系，包括公众对公务人员、政府机构以及政府公共服务的信任。[6]在人工智能时代，不法分子可能利用生成式AI伪造官方公告、政策文件及领导人讲话等虚假信息，营造真假难辨的社会情境。这不仅误导公众对政策的理解，更将引发公众对政府的信任危机，阻碍公共政策的有效制定与执行。在2024年美国大选期间，民主、共和两党政客及其支持者利用深度伪造技术美化自身、抹黑对手、欺骗选民的事件层出不穷。[7]此类事件给美国总统选举和社会信任稳定带来了极大的负面影响，美国人民因此陷入恐慌之中，不禁发问当生成式AI成为总统选举、政治博弈的工具，那么民主和信任又该何去何从呢？公众一旦对政府产生不信任感，面对政府发布的信息，往往会多方查证其真实性。同时，政府在应对利用AI生成技术实施的违法犯罪时，受多重因素制约，常难以及时有效地回应与处理，导致政府公信力与形象受损。更为关键的是，政府信任的消解将动摇公共治理的根基，加剧治理难度。公众对政府机构、政策制定者及公共政策的信任缺失，往往引发其对新出台政策的怀疑或否定。例如，新冠疫情期间，疫苗副作用谣言的广泛传播加剧了公众对接种政策的抵触，不仅阻碍了防疫机制的高效运行，更增加了防疫政策推行的难度。

第二，侵蚀技术信任：技术抵触情绪蔓延。技术信任是指对技术本身或者技术功能的信赖。[8]技术是人类为满足自身需求和解决各类

问题所凝聚的智慧结晶，是以人类福祉为旨归、驱动社会进步与发展的方法与手段。然而，任何事物天然具备双重属性，生成式AI技术在赋能经济腾飞和社会发展的同时，也不可避免地暗藏着风险与隐忧。技术悲观主义认为，技术本质上具有非人道的价值取向，可能导致人类面临灾难性的后果。[9]换言之，他们视技术本质为一种威胁，忧心生成式AI将觉醒自我意识、挣脱人类束缚，进而取代人类的主宰地位。当下，生成式AI已在代码生成、艺术创作、客户服务等多个领域展露出超越人类的潜能。更深层地看，若人类过度依赖AI进行思考与决策，自身的批判性思维与主观能动性将日渐销蚀，最终恐沦为技术的“附庸”。此外，生成式AI自身亦存在诸如“幻觉”等难以察觉的漏洞，可能在内容生成时杜撰文献或虚构事实。加之“AI马斯克”诈骗案、虞某涉黄AI换脸案以及美国大选违规使用深度伪造技术等事件，暴露了技术滥用的风险。可以肯定的是，生成式AI对人类主体性的消解及其滥用事件的频发，正在不断侵蚀公众技术信任的根基，引发公众对AIGC及其相关衍生应用和产品的抵触情绪。这种抵触情绪的扩散将会成为阻碍技术创新与应用、威胁社会公平正义的桎梏。

第三，解构人际信任：社会交往中无端猜疑。信任是社会交往的基石。个体在社会交往中形成的信任，在极大程度上建立在对他人的诚实的信任基础之上。[10]在社会交往中，借助生成式AI炮制的大量真假难辨的视频、音频、图像及文本，正不断冲击个体的认知边界，推高人际沟通成本，瓦解社会信任基础。一方面，社会交往面临沟通对象及内容的真实性挑战。在网络交往中生成式AI能够轻松模拟人类的语音、文字风格乃至体态特征，使公众难以准确辨别交流对象和内容的真实性。这种不确定性会削弱公众在网络交往中的信任度，引发担忧与警惕，使得建立信任变得极为困难，甚至可能使亲密关系因相互猜疑而疏远。另一方面，虚拟世界的人际信任风险可能通过认知迁移等途径渗透到现实社会。网络中的人际信任风险将演变为公众日常交往中的“防御性冷漠”和“普遍化不信任”。为此，公众在社会交往中将不得不花费更多时间和精力去核实对象和信息的真实性，增加信任建立与社会合作的成本。对于个体而言，当社会信任度普遍下降时，其在人际交往中的自我保护意识会增强，变得愈发谨慎甚至多疑。这种过度的自我保护可能导致个体逐渐封闭自我，抵触深入交流和接触，进而影响个人情感体验和社交满足感。对于社会而言，信任度下滑将降低公众对他人与社会的责任感，促使人们聚焦私利而忽视公益，从而动摇社会凝聚力，阻碍有效团结协作，最终危及社会和谐稳定发展。

2. 人工智能生成内容引发的社会信任风险的基本特征

AIGC引发的社会信任风险与其技术特性及社会环境深度关联，呈现出技术特性与社会应用情境相互耦合的双重性特征。

首先，仿真性与隐蔽性交织。AIGC引发的社会信任风险在传播路径上呈现仿真性与隐蔽性交织的特点。借助深度学习、生成对抗网络、大语言模型、扩散模型等技术架构，生成式AI能够迅速生成真假难辨的虚假信息，形成“AI幻觉”。然而，这一复杂的生成过程对公众而言却是不可见和难以解释的“黑箱”。大量AIGC虚假信息往往隐匿于真实信息之中，公众在缺乏专业技术手段和知识储备的情况下，难以直观辨识其真伪。更甚者，算法推荐机制构筑的“信息茧房”，使公众难以接触全面、真实、客观的信息，导致对AIGC虚假信息的辨识能力持续弱化，用户无意中成为虚假信息传播的“推手”。AIGC虚假信息可能在短时间内被海量用户转发分享，引发严重负面影响。例如，“AI数字人”其口型、声音、口音均高度逼真，令普通人难以辨别，使受害者在毫无察觉中身陷骗局。

其次，破坏性强与定责难相互叠加。AIGC引发的社会信任风险在危害呈现上兼具破坏性强与定责难的双重特性。一方面，AIGC引发的风险往往涉及多个领域和行业，一旦爆发将造成严重破坏。在公共治理层面，可能损害政府权威形象，破坏社会治理体系的协同有效性；

在经济发展层面，可能降低社会普遍信任，影响科技创新应用和经济稳定增长；在社会交往层面，可能阻碍信任建立和社会合作，影响公众的社会交往体验及社会和谐稳定。另一方面，AIGC的生成过程涉及算法、数据、用户指令等多方参与，其开发者、运营者或使用者的责任边界难以清晰界定。底层算法的复杂性与不可解释性导致内容生成决策过程难以追溯，而海量且来源多样的数据也使得数据采用者或标注者的责任难以有效认定。此外，用户输入的自然语言指令与模型处理之间可能存在偏差，进一步导致用户责任难以直接追溯。

最后，复杂性与挑战性并存。AIGC引发的社会信任风险在防治路径上呈现复杂性与挑战性并存的属性，其风险样态具有覆盖范围广、成因复杂、破坏性强、治理难度高等特点。这种全新且持续演变的复杂风险形态，无疑为社会风险治理带来了前所未有的巨大挑战。作为一种全新的社会信任的风险形态，AIGC社会信任风险具有独特的表现形式、生成路径与内在特征，传统风险化解的策略和路径难以发挥作用。因此，必须深入生成式AI的技术属性，聚焦AIGC风险的表现形式、核心特征及生成逻辑，制定切实有效的防治策略。此外，生成式AI技术的快速迭代也推动着信任风险样态的持续演变和更新，要求防治策略保持一定的动态性，才能确保风险治理的长期有效性和一致性。

二、人工智能生成内容引发的社会信任风险的生成逻辑

基于对AIGC引发的社会信任风险的表现与特征的系统梳理，本文构建“风险源头–风险触发–风险放大–风险扩散”的风险演化脉络，从全链条视角深入剖析AIGC引发的社会信任风险的生成逻辑。

1. 风险源头：算法黑箱与数据局限

算法和数据是AIGC的基础与前提。算法黑箱与数据局限为AIGC埋下了诱发社会信任风险的隐患和源头。一方面，AIGC底层算法逻辑的不透明性和不可解释性，加剧了其“黑箱”属性。算法作为生成式AI技术的核心，常被视作模拟人类智能行为的数学模型。通过处理和分析大量数据，算法助力AI大模型决策、学习和解决问题，进而实现AI内容生成进阶。然而，算法黑箱源于生成式大语言模型内部结构的复杂性、数据输入的多样性和巨量参数规模，使得模型决策过程难以被外部理解，成为智能算法价值偏向的“隐蔽所”与“遮羞袍”。[11]换言之，在AI的内容生成过程中，算法的运行原理对公众而言是不透明且难以解释的，公众难以洞悉AIGC的内部机制。随着生成式AI能力的提升，黑箱困境就愈发严重，模型越强大，其内部运行原理越复杂难懂。从自然语言处理模型到神经网络与深度学习的广泛应用，模型参数规模已突破人类认知边界，公众认知水平与技术迭代形成巨大落差，为AIGC蒙上神秘面纱，侵蚀着公众的信任根基。

另一方面，训练数据可能存在原始偏见和污染，威胁AIGC的可信度。生成式人工智能大模型依赖海量训练语料和高质量的微调语料进行学习与理解。[12]AIGC所采用的数据喂养生成范式在提升效率的同时亦潜藏社会信任风险，这种风险根源于数据喂养方式的固有局限。由于AIGC对训练数据高度依赖，一旦数据存在污染或偏见（如虚假信息、违法信息、种族歧视、文化偏见等），将直接影响AIGC的准确性、真实性和客观性。目前，大量数据源自受商业利益、政治倾向和文化偏见影响的互联网平台。以ImageNet数据集为例，超过45%的数据来自美国，而中国和印度合计仅贡献3%。[11]大模型在持续训练中继承数据中的统计主导特征，可能导致英语国家互联网平台在训练数据中占据隐性主导地位，进而边缘化其他文化特征。这种带有特定国家、政治、文化、民族和语言偏见的AI生成，潜在诱发公众信任危机的可能。

2. 风险触发：AI“幻觉”与技术滥用

第一，AI“幻觉”是AIGC引发社会信任风险的直接触发机制。AI“幻觉”是指人工智能系统在生成内容时，产生的看似合理但实际上错误、虚构或不存在的信息。[13]AI“幻觉”产生技术内部，关系公众使用AIGC的体验感与获

得感，直接影响公众对AIGC的信任。具体来说，在面对不会答或者不知道的问题时，以ChatGPT为代表的生成式AI会“说谎”，在生成文本的过程中“一本正经地胡说八道”。[14]更具体地看，对于那些可能将生成式AI用于辅助工作和学习的公众而言，这种AI的“幻觉”现象潜藏着致命的危害，因为“如果大语言模型（Large Language Model, LLM）始终未能消除幻觉，而用户习惯于消费AI‘幻觉’、利用‘AI’幻觉甚至创建‘AI’幻觉，作为国民应用的AI则持续地协同人生产后真相，那么，一个‘奇幻社会’便会来临，带来文化、生活、社会信任危机。”[15]显然，AI“幻觉”现象将会影响公众对相关信息的准确认知，导致其逐渐丧失辨别信息真伪的能力，可能会盲目迷信AIGC，将其置于理性判断之上。如此，一旦AI“幻觉”破灭，便会触发AIGC的社会信任危机。

第二，技术滥用是AIGC引发的社会信任风险的间接触发机制。技术滥用是指将生成式AI用于非法、不道德或其他不当用途，对公众的生命财产安全造成威胁的应用过程。生成式AI具有极强的创造性、仿真性和互动性，不法分子可以利用其输出内容（AIGC）实施极具迷惑性的犯罪行为。[16]生成式AI已成为不法分子或有心之人用以欺骗受害者的工具。例如，基于AI生成技术制作出极具迷惑性的视频、图像、音频和文本，用以诓骗当事人或从事色情交易，从而获取大量非法收入。这种现象如果不加以制止和约束，必然消耗公众对AIGC及其衍生应用的好感。无论是出于风险防范的自我保护，还是对成本收益的权衡，公众都可能会对AIGC及其衍生应用产生不信任，间接引发社会信任风险。

3. 风险放大：认知失调与制度失灵

认知失调与制度失灵会放大AIGC引发的社会信任风险的影响。一方面，AI技术的持续迭代使得传统社会的认知验证体系面临崩解和失效，技术对“真实”的重新定义正在不断地消解社会信任的底层逻辑，削弱个体对社会风险的应对能力。随着深度学习、生成对抗网络、大语言模型、扩散模型等技术的迭代，AIGC在感官上高度接近甚至超越人类创作的真实性水平，已突破人类感官验证的生理阈值。传统社会中个体习惯于依赖感官经验、生活常识与社会共识来构建现实认知，从而达到辨别信息真伪的目的。数字时代AIGC已能模拟人类的表情、神态、语气、音色、文笔等，甚至生成毛孔级别的皮肤纹理，传统验证手段难以辨别真伪。在美国大选中，伪造的“总统演讲”视频，演讲者喉结滚动的节奏与声波频率完全同步，已突破人类肉眼鉴伪的极限。当传统验证体系全面失效，社会将被迫转向技术中介化信任（区块链溯源、数字水印检测等），进一步弱化了公众风险认知和应对能力。这种技术中介化信任仅是将信任对象从AIGC转移至验证技术，而验证技术本身的可靠性尚未可知。

另一方面，生成式AI的发展呈现指数级增长，而规范层面的制度更新却仍然滞后，导致AIGC引发的社会信任风险难以得到有效化解。具体来看，AIGC依托复杂的算法、海量的数据训练及特殊的程序设定，使其在创作主体、过程和结果等多方面与传统媒体和人类创作存在显著差异，现行基于传统情境制定的法律法规难以有效地对其进行约束。在处罚利用AI生成技术产生的违法犯罪行为时，由于缺乏明确的法律依据，难以对相关责任主体进行追责和处罚。与此同时，AIGC“数字巴别塔”[17]滋生监管套利、技术套利等问题。AIGC“数字巴别塔”是指由于不同国家或地区在经济发展、政治倾向、文化包容性及民众接受程度等方面的差异，导致相关的法律法规和政策方针往往存在显著差异，使得AIGC的全球治理呈现出碎片化、差异化和本土化的鲜明特征。这可能引发监管套利和法律冲突问题，导致AIGC虚假信息在国际间肆意传播，威胁全球信息安全与社会信任的形成和稳定。

4. 风险扩散：算法推荐与去中心化传播

AIGC所引发的社会信任风险隐匿在虚假信息之中。那么，AIGC虚假信息的传播实际上就是社会信任风险的扩散过程。一方面，AIGC虚假信息能够依托算法推荐实现精准推送。算法推荐机制依据用户的行为轨迹将AIGC虚假信息

精准推送给可能感兴趣并倾向于转发分享的用户。一旦用户首次接触并互动AIGC虚假信息，算法便会持续推送相似内容，甚至反复推送相关变体，最终形成“信息茧房”，导致公众对AIGC虚假信息的判断力显著下降。此外，众多社交平台为提升用户活跃度或追求流量，可能优先推荐情绪极化、争议性强的内容。AIGC虚假消息因其标题夸张、内容震撼、图片仿真等显著特点，更易被算法选中并推广。被推广的AIGC虚假信息利用用户的固有偏见或极端标题刺激情绪，用户因“情绪共鸣”而忽视对信息的审查，甚至可能再次传播AIGC虚假信息。总的来说，在算法推荐机制的加持下，AIGC虚假信息能够迅速实现精准推送并广泛传播。

另一方面，去中心化传播加剧了AI生成的虚假信息的泛滥。人工智能时代信息传播权力被不断下放，信息传播不再依赖传统的传播手段与途径，呈现出鲜明的去中心化传播特征，人人都可以成为信息的制造者和传播者。去中心化传播意味着AI生成的信息不再通过传统的中心节点传递，而是通过无数个体节点扩散，最终产生裂变效应。在这种传播模式下，每个接收者都可能成为新的传播源（通过点赞、分享、转发等方式），导致AI生成的虚假信息传播速度更快、范围更广、危害更大。以新闻发布为例，过去新闻通常由专业机构审核后发布，信息源明确且可信度高。如今，任何人利用AI生成技术迅速生成大量逼真的虚假新闻，并通过社交媒体、网络平台等媒介传播扩散。此外，去中心化传播使得AI生成虚假信息难以识别。在传统的中心化平台，用户可通过查看信息来源辅助判断真伪，如照片胶卷底片，文件纸质原件，视频拍摄记录。然而，在去中心化传播过程中，信息常以“去标签化”形式存在，且AI生成技术还能伪造来源信息（如虚假的官方通告、伪造的学术论文等），使得AI生成的虚假信息难以被及时有效识别，加剧了AIGC虚假信息的泛滥。

三、人工智能生成内容引发的社会信任风险化解策略

信任的复归不仅仅取决于个体的道德自律，还需要通过完善的制度保障机制来指引人们的行为。[18] 为此，应该从法治、技治、众治三个维度综合施策，引导AIGC技术朝着健康、良性的方向发展。

1. 法治：构建权责边界清晰、规范化的AIGC法律框架

第一，加快完善AIGC相关法律法规。一方面，以法律法规的更新推动数据来源标注常态化。规定AI大模型在生成内容时需标注来源，使用户在获取信息时能够清晰了解其出处，增强对内容真实性的判断能力。另一方面，明确划分开发者、数据来源方、用户及内容发布平台在AIGC虚假信息传播、侵权等复杂问题中的法律责任。开发者需在技术开发过程中确保技术向善、可控，并持续提升算法透明性，承担因算法缺陷导致的AIGC失真责任；数据来源方必须确保训练数据合法合规，避免使用未经授权或存在侵权风险的数据，并承担因数据局限导致的AIGC失真责任；用户在使用AIGC时应遵守法律法规及道德规范，若故意生成、利用虚假信息谋取私利，须承担相应法律责任；内容发布平台应采用先进技术手段严格审查上传内容，及时识别并拦截AIGC虚假信息，若大量虚假信息扩散并对公众造成心理、身体或财产损害，平台亦需承担相应法律和社会责任。

第二，强化国际法治合作与交流。一是亟需加快建立国际信息共享机制，打破国与国之间的数据隔阂，促进各国在AIGC相关资讯上的互联互通。通过共享AIGC违法犯罪事件及虚假信息方面的情报、经验和教训，提升各国应对利用AI生成技术违法行为的综合能力。二是针对利用AI生成技术跨国犯罪，强化联合调查与执法。鉴于AIGC虚假信息的传播常跨越多个国家和地区，单一国家的执法力量难以全面追踪和惩治违法源头。借助国际合作，各国执法机构可以整合资源，共享调查线索和证据，构建全球性的执法协作网络。三是制定统一的国际AIGC监管标准。各国应积极在国际层面展开协商，确立涵盖数据标注规范、内容生成准则、审核标准及法律责任界定等多方面的统

一框架。此举旨在有效填补监管套利带来的“真空地带”，为全球AIGC的健康发展营造一个规范、有序且值得信赖的环境。

2. 技治：建立全方位全过程的AIGC技术防护体系

第一，强化事前预防工作。持续改进和优化AI模型的训练机制，提升可靠性和安全性。一是增强AI模型的语义理解能力。通过增加预训练数据的多样性，如使用包含多种语义表达、不同领域文本的数据，使模型接触各类语义场景，在持续训练中逐步提升其语义理解能力。引入Transformer架构中的多头注意力机制，帮助模型更精准地聚焦文本中的不同语义部分，理解上下文信息，从而全面理解和处理用户指令。二是提升AI生成模型的逻辑推理能力。设计层次结构的神经网络，使模型能分层处理信息，类似人类分步骤解决问题的逻辑。同时，增加专门的逻辑训练任务，让模型在完成大量逻辑任务中不断学习推理规则和模式。三是建立严格的数据筛选标准，确保训练数据质量可靠、使用合规，减少因数据局限导致的内容质量问题。

第二，完善事中监测工作。加快创新AIGC监测与验证技术，防止AI生成的虚假信息泛滥。一是开发更先进的数字水印技术。专门的检测工具通过预设的数字水印提取算法，从图像、视频或文本中提取隐藏的数字签名，解析后验证AIGC的来源及其在传播过程中是否被篡改。二是加快建立AIGC监测系统。平台应建立高效的AIGC实时监测系统，依托强大的机器学习算法，对海量AIGC进行不间断扫描与分析。通过与已知AIGC样本库比对及运用语义理解技术评估文本逻辑合理性，确定AIGC是否为虚假信息。若判定为虚假信息，系统将迅速发出警报，通知平台运营方采取相应措施，遏制其进一步传播。

第三，落实事后救济工作。一方面，要最小化风险损失。对已造成恶劣影响和严重后果的AIGC虚假信息或违法犯罪行为，应及时通报相关监管部门或新闻媒体，以便后续调查处理。通过官方通告和媒体报道披露AIGC违法犯罪的真实信息，遏制虚假信息传播，限制风险扩散。另一方面，建立AIGC虚假信息及违法犯罪数据库，将已出现的虚假信息和违法犯罪信息按类别、等级记录在案，可依据数据库信息对平台内容进行比对，提升识别虚假信息的准确率和效率。

3. 众治：构建多元主体共同参与的AIGC协同监管体系

第一，国家应该加强宏观调控。一是以主流价值观引领AIGC的创作方向，鼓励平台和用户生成高质量、有深度、富有创意的AIGC内容，确保其发展契合社会需求。二是加快设立AIGC监管机构。凭借公共机构的权威和资源调配能力，依据国家战略、法律法规及社会整体利益，为AIGC监管制定总体方针与战略规划。三是加大对AIGC监管领域的技术、资金、人才等资源投入，鼓励科研院所、高校、科技企业深化AIGC监管研究。

第二，平台需要强化内容审核。作为信息传播的关键节点，社交媒体平台肩负着维护信息真实性、合法性与道德性的重任，强化平台审核的能力机制，遏制AIGC虚假信息的传播。内容发布平台需建立一套严谨的审核流程，覆盖预发布和已发布的所有AI生成内容。采用自动化审核工具与人工审核相结合的方式，对AIGC进行全面、多层次的检查，力求从源头拦截虚假信息。在审核过程中，对于涉嫌违法、侵权、违反社会道德的AIGC虚假信息，平台可视情节对违规账号实施不同程度的封禁措施，提升平台对AIGC的审核质量和监管水平，为用户营造健康、安全、可信的网络信息环境。

第三，科技企业应坚持负责任的创新。科技企业在追求生成式AI研发和推广盈利的同时，亦应兼顾社会效益，承担相应社会责任。[19]一方面，需加快提升AIGC的安全性和可靠性，通过加大在模型优化、数据清洗、算法更新等方面的人力、物力、财力投入，降低AIGC致幻风险。另一方面，积极响应政府号召，加强如数字水印技术、区块链技术、AIGC监测技术的研发。

第四，公众应该不断提升数智素养。一方面，公众应主动通过互联网、学校、媒体等渠

道学习AIGC相关知识，不断提升自身数智素养，掌握AIGC虚假信息的基本原理、常见类型及判断方法，增强辨别能力。另一方面，公众需合法合规使用AIGC技术，确保AI生成的内容符合法律法规和道德规范，当好内容发布平台的“监督者”，及时揭露AIGC虚假信息及违法犯罪行为。

［参 考 文 献］

[1] 杨敏然、张新兴、陶荣湘. 现状与趋势：国内人工智能生成内容（AIGC）研究透视 [J]. 图书馆理论与实践，2024,（2）：56–65.

[2] 邓建鹏、赵治松. 文生视频类人工智能的风险与三维规制：以Sora为视角 [J]. 新疆师范大学学报（哲学社会科学版），2024, 45（6）：92–100.

[3] 邱泽奇.技术与组织的互构——以信息技术在制造企业的应用为例 [J]. 社会学研究，2005,（2）：32–54; 243.

[4] 张宁、高鹏程. 生成式人工智能情感模拟的伦理风险与治理路径：基于技术–社会互构理论框架的分析 [J]. 科学决策，2025,（2）：123–134.

[5] 朱贻庭. 伦理学大辞典 [M]. 上海：上海辞书出版社，2010, 22.

[6] 姜晓秋、陈德权. 公共管理视角下政府信任及其理论探究 [J]. 社会科学辑刊，2006,（4）：41–44.

[7] 袁莎. 深度伪造技术对美国选举政治的影响 [J]. 当代美国评论，2024, 8（4）：102–123.

[8] Mcknight, D. H., Carter, M., Thatcher, J. B., et al. 'Trust in a Specific Technology: An Investigation of Its Components and Measures' [J]. *Acm Transactions on Management Information Systems*, 2011, 2(2): 1–25.

[9] 赵建军. 追问技术悲观主义 [M]. 沈阳：东北大学出版社，2001, 154.

[10] 郑也夫. 信任论 [M]. 北京：中国广播电视出版社，2001, 16.

[11] 马忠、高怡英. 生成式大语言模型的社会认知风险与应对 [J]. 浙江社会科学，2025,（2）：95–105.

[12] 钱力、刘熠、张智雄 等. ChatGPT的技术基础分析 [J]. 数据分析与知识发现，2023, 7（3）：6–15.

[13] 邱元阳. AI的幻觉 [J]. 中国信息技术教育，2025,（9）：22.

[14] 经羽伦、张殿元. 生成式AI幻象的制造逻辑及其超真实建构的文化后果 [J]. 山东师范大学学报（社会科学版），2024, 69（5）：113–126.

[15] 杜骏飞. 奇幻社会的来临——DeepSeek幻觉与后真相递归 [J]. 探索与争鸣，2025,（3）：11–14.

[16] 盛浩. 生成式人工智能的犯罪风险及刑法规制 [J]. 西南政法大学学报，2023, 25（4）：122–136.

[17] 杜骏飞. 数字巴别塔问题 [J]. 当代传播，2024,（1）：71–76.

[18] 王天夫. 重建社会信任是社会管理的首要任务 [J]. 行政管理改革，2012,（11）：42–46.

[19] 王张华. 政府与平台型企业合作模式及其风险管控研究论纲——基于数字治理的视角 [J]. 湘潭大学学报（哲学社会科学版），2023, 47（4）：61–69.

［责任编辑　李斌］