

应对 AI 时代健康科普新挑战

加快权威健康知识图谱建设专项报告

2026 年央视 3·15 晚会首次聚焦 AI 行业乱象，公开曝光利用生成式引擎优化（GEO）技术对 AI 模型实施信息“投毒”的黑色产业链，通过批量制造虚假软文、伪造信息证据、操控网络检索权重，诱导 AI 模型输出虚假、误导性内容，直接引发 AI 作为主流信息入口的信任危机。

在全民健康科普场景中，这一风险呈现致命性、广泛性、隐蔽性特征。当前，AI 问答、智能搜索、健康科普小程序已成为公众获取养生、疾病、用药、急救等健康知识的核心渠道，而针对健康领域的 AI 投毒行为，本质是通过污染训练数据、劫持检索系统、注入恶意提示，让大模型与 RAG 系统输出错误医学建议、伪健康科普、虚假养生理论，不仅摧毁 AI 健康服务的公信力，更会直接导致公众误信谣言、错误用药、延误诊疗，严重冲击健康科普权威性，威胁全民健康素养提升与公共卫生安全。

面对 AI 技术异化对健康科普阵地的侵蚀，依托权威医学资源构建标准化、结构化、可溯源、可验证的健康知识图谱，打造健康科普的“事实底座”与“防伪屏障”，成为守护健康科普阵地、保障公众健康知情权的紧迫任务与核心举措。

一、现状与风险：AI 健康科普领域的“投毒”核心隐患

1.投毒模式全链条覆盖，健康科普内容精准被污染

不法分子利用 AI 模型依赖公开网络数据、缺乏事实核验能力的漏洞，针对健康科普场景形成三层级攻击模式，直接破坏科普信息真实性：

训练数据污染：批量篡改健康科普知识库、疾病百科、用药指南等公开知识源，将错误养生方法、虚假疗效、伪医学理论植入模型底层，污染具有长期性、顽固性，修复成本极高。

检索上下文劫持：操纵 RAG 检索系统权重，让伪造的科普文章、虚假专家观点、违规产品宣传优先被调用，隐蔽性强、门槛极低，成为健康科普最常见的污染方式。

提示注入诱导攻击：通过构造特定健康问题话术，诱导 AI 输出符合不法目的的科普内容，如夸大保健品功效、传播偏方治病谣言、否定正规诊疗手段，直接误导普通民众。

2.GEO 技术深度异化，健康谣言形成“伪科普证据链”

生成式引擎优化技术被不法利用后，成为健康科普的最大污染源：攻击者借助大模型批量生产伪原创健康软文、虚构专家身份、伪造临床数据与科普结论，配合 SEO 技术提升网络曝光度。研究显示，少量伪造健康科普内容即可形成模型后门效应，大幅提升错误科普推荐概率。

伴随流量入口向 AI 平台迁移，传统搜索审核机制失效，健康谣言以“专业科普”“权威解读”“养生秘籍”等形式伪装传播，公众难以分辨真伪，导致伪科普取代真科普，严重扰乱健康科普传播秩序。

3.监管治理滞后，健康科普缺乏权威“对冲体系”

现有互联网监管聚焦直接虚假广告与违规信息，对间接 AI 投毒式伪科普界定模糊、溯源困难、惩处乏力，尚未形成针对 AI 生成健康科普内容的强制性审核、溯源与责任认定机制。

目前全球范围内的内容溯源技术、AI 水印技术在健康科普领域尚未落地应用，健康信息真假混杂、专业度参差不齐，公众缺乏统一、权威、可信的健康科普参照体系，面对 AI 生成的伪科普内容极易被误导，健康科普阵地面临失守风险。

二、核心对策：以健康知识图谱建设筑牢健康科普权威底座

以中国健康教育中心联合 CHTV 百姓健康频道成立健康知识图谱课题组为核心抓手，聚焦健康科普主场景，加快推进权威健康知识图谱建设，构建 AI 时代健康科普“防护网”。

1. 构建权威科普知识底座，打造健康信息“事实锚点”

摒弃大模型概率性、易被污染的统计关联模式，建立以权威来源为核心的结构化健康知识图谱，实现健康科普从“随机生成”向“依据推理”转变。

整合国家级权威科普资源：系统纳入循证医学指南、国家卫健委科普标准、疾控中心健康提示、药品说明书、临床诊疗规范、全民健康素养基本知识等核心内容，参照国家科学健身知识图谱建设标准，统一健康术语、规范科普口径。

遵循医疗信息通用标准：采用分层设计架构，对标 FHIR、LOINC 等国际医疗信息标准，确保知识图谱与各类 AI 科普平台、健康服务系统兼容互通。

构筑科普防伪屏障：将知识图谱作为 AI 健康科普的唯一可信检索源，所有 AI 生成的健康科普内容必须锚定图谱节点，实现来源可查、依据可验、内容可控，从源头阻断伪科普输出。

2. 依托课题组机制，强化科普知识权威审核与协同

建立跨学科、跨领域的健康知识图谱治理体系，保障科普内容专业性、准确性、时效性。

组建健康科普专家委员会：汇聚临床医学、公共卫生、健康传播、医学科普、信息技术等领域权威专家，专职负责健康科普本体设计、知识点审核、科普语言转化、内容更新把关。

建立高效协同落地机制：借鉴国家主动健康专项与全域联动工作模式，明确主管单位、医疗机构、科普平台、技术机构职责分工，实现健康科普知识从采集、审核、入库到应用的全流程闭环管理。

3.聚焦健康科普场景，实现知识图谱普惠化应用

坚持以科普为核心、以应用为导向，推动健康知识图谱深度融入全民健康科普全场景：

AI 智能健康科普标准化输出：为政务健康平台、媒体科普账号、智能问答助手提供统一权威知识库，自动生成标准化科普内容，涵盖疾病预防、合理用药、科学养生、急救知识、妇幼保健、老年健康等全品类科普信息。

全生命周期精准科普推送：基于知识图谱构建居民健康画像，结合年龄、性别、生活习惯、健康风险，实现个性化科普精准送达，让健康科普更贴合公众需求。

健康谣言智能识别与权威辟谣：依托图谱实时监测网络热门健康谣言、伪科普话术，自动生成权威辟谣内容，通过 AI 快速推送，形成“真科普对冲伪谣言”的高效机制，提升全民健康素养。

基层健康科普能力赋能：为社区卫生服务中心、乡村医疗机构提供标准化科普素材与智能科普工具，补齐基层健康科普专业短板，打通科普服务“最后一公里”。

三、 实施建议：保障健康知识图谱长效服务健康科普

1.技术融合创新，提升科普内容抗干扰能力

采用神经网络+知识图谱+规则引擎融合架构，通过 AI 技术高效抽取、更新健康科普知识，运用符号系统对知识点进行逻辑校验与异常检测，以权威图谱为标准微调大模型，彻底杜绝 AI 生成虚假健康科普内容，保障科普输出稳定、准确、可信。

2.建立动态更新机制，保持科普知识时效性

建立与国家卫健委、疾控中心、药监局等官方渠道实时对接的科普知识更新制度，自动同步最新诊疗规范、健康提示、用药信息、防疫知识等核心内容，通过智能检测与专家复核相结合的方式，确保健康科普知识始终与权威发布保持一致，杜绝过期、错误信息传播。

3. 构建多方共建生态，凝聚健康科普合力

打造政、产、学、研、媒、用一体化健康知识图谱生态：政府主导标准制定与监管；医疗机构提供权威科普内容；科研机构提供技术支撑；媒体与平台负责权威科普传播；公众参与反馈与监督。通过数据共享、隐私保护、激励机制，打破信息壁垒，形成全社会共同守护健康科普阵地的合力。

四、构建国家级健康科普指数图谱平台

打造兼具权威知识服务与智能治理能力的中枢系统，服务于全民健康素养提升与网络生态净化双重使命。平台将以医学知识图谱为基础，融合人工智能技术，建立覆盖全生命周期、重点疾病领域的结构化知识网络，支撑科学内容生产与精准辟谣响应，成为国家健康信息基础设施的核心组成部分。

（一）具体目标

- 知识集成目标：**建成覆盖全生命周期、重点疾病领域的结构化知识库，支撑95%以上科普内容的智能化生成与审核，确保知识来源可溯、逻辑严谨。
- 治理赋能目标：**建立“监测—识别—研判—处置—反馈”一体化治理闭环，年均发布辟谣报告不少于12期，重大舆情响应时间不超过4小时。
- 技术突破目标：**实现“大语言模型（LLM）+医学知识图谱”双引擎推理机制，对复杂型伪科普（如因果倒置、偷换概念、断章取义）的识别准确率 $\geq 90\%$ 。
- 机制创新目标：**建成跨平台协同治理网络，接入不少于5家主流内容平台API接口，实现线索移交、联合标注与分级处置。
- 人才培育目标：**完成千名医务人员科普能力认证培训，提升源头供给质量，推动医务人员成为科学传播的主力军。

（二）核心架构设计

（1.模式层（本体构建）

在原有医学实体类型基础上，进一步扩展治理相关实体，包括“谣言类型”“辟谣证据链”“发布者信用等级”“误区类别”“高风险人群标签”等，增强图谱对伪科普内容的表征能力。同步构建新型语义关系，如“被驳斥”“来源于”“相似于”“高风险人群易感”“商业利益关联”等，支持逻辑推演与溯源分析，为智能识别提供本体支撑。

2.数据层（资源汇聚）

整合多源异构数据，构建四大核心数据库：

- **权威知识库：**纳入WHO全球健康建议、国家诊疗规范、循证医学数据库(如Cochrane Library)、中华医学会指南等，作为科学判断的基准依据；
- **伪科普案例库：**收录经专家确认的典型伪科普样本，标注其传播形式、话术特征、影响范围与危害程度，形成可计算的负面知识集合；
- **辟谣资源包：**预置标准化反驳话术、短视频脚本模板、图文素材与常见问答

集，支持快速响应与内容再生产；

- **创作者信用档案**：记录内容科学性评分、用户投诉率、辟谣响应速度、违规历史等指标，实施动态管理与分级分类。

3.技术层（系统支撑）

部署三大核心技术模块：

- **AI 审核引擎**：基于多模态特征库，实现对文本、图像、音频、视频内容的自动初筛，识别关键词变异、视觉误导、语音伪造等异常模式；

- **双引擎推理模型**：结合大语言模型的语义理解能力与医学知识图谱的逻辑验证能力，进行深层语义解析与科学性校验，提升对半真半假信息的辨识精度；

- **信用评分系统**：采用加权算法动态计算健康传播信用分，实施红黑名单管理，对优质创作者给予流量扶持，对违规主体实施梯度惩戒。

（四）治理支撑层

为强化平台的主动干预能力，增设治理支撑层，包含以下功能模块：

- **舆情预警中心**：实时推送高传播风险伪科普内容，按紧急程度分级告警；

- **联合辟谣指令下发系统**：一键向合作平台发送处置建议，支持限流、标注警示或下架操作；

- **公众自助查询接口**：开放小程序/API 供第三方调用，支持用户输入内容片段查验真伪；

- **治理成效评估仪表盘**：可视化展示辟谣覆盖率、误判率、响应时效、公众满意度等关键绩效指标，支撑持续优化。

五、治理功能定位与实施路径

（一）科学辟谣与误区澄清的核心功能

功能	实现方式
科学溯源	图谱节点链接原始研究、临床路径、官方指南，实现一键追溯权威依据
逻辑验证	双引擎检测因果倒置、偷换概念、相

	关误作因果等逻辑谬误
快速响应	调用预置模板自动生成辟谣文案、短视频脚本与社交媒体推文
公众教育	推送个性化辟谣内容至老年人、慢性病患者等高风险人群

上述功能依托于动态更新的“健康科普乱象特征库”，涵盖文本、图像、行为三维度标签体系，确保对新型伪科学话术的持续识别能力。

（二）“五步闭环” 治理流程

- 1. **监测采集**：利用 AI 爬虫持续抓取主流社交平台、短视频平台、新闻客户端中的健康类内容，建立全天候监测网络；
- 2. **智能识别**：通过特征库匹配与双引擎模型分析，筛选出疑似伪科普内容，生成初步风险评级；
- 3. **专家研判**：由医学专业人员与传播学者组成的专家组进行复核确认，确保定性准确；
- 4. **协同处置**：通过 API 接口联动合作平台，实施限流、添加警示标识或下架处理，并向社会公开处置结果；
- 5. **反馈优化**：将典型案例反哺至知识图谱与特征库，优化识别模型参数，实现系统的自我进化。

该闭环机制确保治理过程科学、透明、可追溯，具备持续迭代能力。

（三）三大机制创新

- 1. **健康传播信用积分制度**：建立全维度创作者信用档案，围绕内容科学性、合规性、辟谣响应度等设置评价指标，实施动态评分与分级管理。对信用良好者给予资质认证、流量倾斜等正向激励，对低信用主体采取警告、限流、账号联惩等约束措施。

2. **跨平台协同治理协作机制**：推动与抖音、快手、微信公众号、微博、B 站等至少 5 家主流平台建立 API 级数据互通，实现线索移交、联合辟谣、账号管理等方面的常态化共治流程，打破信息孤岛。

3. **定期发布《健康科普传播力指数报告》**：作为持续观测产品，按季度发布行业报告，追踪优质内容传播效能、医生个人品牌影响力及治理成效，引导行业健康发展，为政策制定提供数据支撑。