

“人工智能 + ”网络生态治理体系变革： 机理、挑战与对策

孙 浩, 苏 竣

(清华大学公共管理学院、智能社会治理研究院, 北京 100084)

摘 要:在智能社会时代,人工智能推动网络生态治理从互联网生态治理范式向数智生态治理范式转型。本文基于“介质变革—结构重组—认知扰动—秩序适配—生态修复”五环分析框架,系统解释人工智能驱动网络生态治理体系变革的作用机理;识别中国情境下的主要挑战,提出相应对策建议。研究表明,人工智能首先改变内容与行动的生成介质,继而重组平台注意力结构,并进一步影响公众事实判断、情绪反应和价值认同,进而要求治理对象从末端内容扩展至数据、算法、模型、平台和智能体全链条,最终指向事实恢复、权益救济和制度学习意义上的生态修复。然而,我国也面临生成内容源头识别困难、平台结构性风险放大、公共事实基础受扰、制度规则衔接不足和生态修复韧性薄弱等挑战。针对上述挑战,我国的治理路径应在党的领导、人民至上、守正创新、法治护航和系统观念统摄下,吸收风险分级、系统性风险评估、组织风险管理、前端安全设计等国际经验,构建生成内容标识与模型测评、平台系统性风险评估、公共事实定性、分类分级监管、生态修复和国际规则合作相衔接的治理体系。

关键词:智能社会;数智生态;认知重塑;网络沉迷;虚拟现实迁移

中图分类号:C911 文献标识码:A 文章编号:1005-0566(2026)09-0068-15

Transformation of the AI + network ecosystem governance system: Mechanisms, challenges, and countermeasures

SUN Hao, SU Jun

(Institute for Intelligent Society Governance, School of Public Policy and
Management, Tsinghua University, Beijing 100084, China)

Abstract: In the era of an intelligent society, artificial intelligence drives the transformation of network ecosystem governance from the Internet ecosystem governance paradigm to the digital intelligence ecosystem governance paradigm. Based on the five-ring analysis framework of “medium transformation—structural reorganization—cognitive perturbation—order adaptation—ecosystem restoration”, this paper systematically explains the mechanism by which AI drives the transformation of the network ecosystem governance system; identifies the main challenges in the Chinese context, and proposes corresponding countermeasures. The study shows that AI first alters the medium for generating content and actions, then reorganizes platform attention structures, further influences public factual judgment, emotional responses,

基金项目:国家自然科学基金面上项目“虚拟化身、心理预期与行为遵从:基于社会实验的网络游戏对用户合作行为的影响机理与治理机制研究”(72374123);清华大学计算社会科学与国家治理实验室资助项目“基于计算实验的智能社会治理若干重大理论问题研究与数据库建构”(CGLab20250003)。

作者简介:孙浩(1986—),男,江苏徐州人,清华大学公共管理学院助理教授,硕士生导师,博士,研究方向为智能社会治理、赛博空间治理。通信作者:苏竣。

and value identification, and ultimately requires governance targets to expand from end-stage content to encompass data, algorithms, models, platforms, and intelligent agents, ultimately pointing toward ecosystem restoration in terms of factual recovery, rights relief, and institutional learning. However, China also faces challenges such as difficulties in identifying the sources of generated content, amplification of platform structural risks, disruption of the foundation of public facts, insufficient alignment of institutional rules, and weak resilience in ecosystem restoration. To address these challenges, China's governance approach should, under the guidance of Party leadership, people-first principles, adherence to principles while innovating, legal safeguards, and a systemic perspective, incorporate international experiences such as risk classification, systemic risk assessment, organizational risk management, and front-end security design to establish a governance system that links generated content identification and model evaluation, platform systemic risk assessment, qualitative public facts, classified and tiered supervision, ecosystem restoration, and international rule cooperation.

Key words: intelligent society; digital intelligence ecosystem; cognitive reshaping; internet addiction; virtual reality migration

习近平总书记关于网络生态治理的重要论述指出,网络生态治理是网络强国建设的重要任务,事关国家发展和安全,事关人民群众切身利益。抓网络生态治理必须坚持党的领导、坚持人民至上、坚持守正创新、坚持法治护航、坚持系统观念^[1]。《国务院关于深入实施“人工智能+”行动的意见》提出,加快推动人工智能赋能网络空间治理,强化信息精准识别、态势主动研判、风险实时处置等能力^[2]。这意味着,“人工智能+”网络生态治理是当前重大的理论命题与实践问题,亟须学界进行体系化学理化研究阐释。

在智能社会时代,人工智能正在改变网络生态运行的底层结构^[3],推动网络生态治理从传统的互联网生态治理范式向当前的数智生态治理范式转型。一般意义上的互联网生态,主要强调网络空间中的技术连接、平台应用、内容生产、信息流动和用户互动^[4];本文所论的网络生态,指的是智能社会条件下由主体生态、内容生态、技术生态、规则生态和价值生态交互形成的数智生态。它的关键特征在于:第一,主体不仅是自然人和组织,还包括数字人、智能体和自动化系统;第二,内容不只是由人工生产,还包括模型生成、多模态合成和算法再加工;第三,传播不再局限于用户转发,还包括算法排序、推荐机制和流量激励;第四,治理不能只关注平台处置和事后监管,需要延展到源头标识、结构干预、认知修复和制度学习。这正是当今网络生态治理区别于一般意义上的互联

网治理和网络安全治理的结构性特征。

面对上述变化,传统互联网生态治理模式面临三重矛盾。首先,风险生成方式已经模型化,而治理对象仍偏重自然人内容发布;生成式大模型、多模态合成和智能体行动使内容来源、主体身份和责任链条变得复杂,单纯沿着“发布者—平台—内容”的路径追责已经难以充分覆盖风险源头。其次,风险传播机制已经算法化,而治理工具仍偏重末端审核和事后处置;平台推荐、热榜排序、社群扩散和商业投放共同塑造注意力结构,政策干预如若不进入平台结构和算法机制,治理只能在风险结果端被动补救。最后,风险后果已经生态化,而治理评价仍偏倚短期乱象整治指标和专项行动成效;删除、封禁、限流可以快速降低可见风险,但难以自动恢复被破坏的事实秩序和价值秩序^[5]。

现有研究为理解这一问题提供了重要基础,但仍存在进一步探究的空间。互联网治理研究关注网络主权、多利益相关方机制、全球技术协调和网络空间规则^[6-8];平台治理研究揭示平台如何通过规则、接口、审核、推荐和商业模式塑造公共表达^[9-12];算法治理研究强调算法并非中立工具,而是嵌入资本逻辑、权力结构和价值选择的社会技术装置^[13-15];人工智能治理研究围绕可信人工智能、伦理原则、风险管理、算法问责、模型透明、审计评估和全球协调展开^[16-18];信息生态研究揭示虚假信息、错误信息、情绪动员、回音室和认知

偏差对公共事实基础的侵蚀^[19-22]；风险治理与数字政府研究则提示，复杂风险治理需要协同、韧性和制度学习^[23-26]。这些研究各有贡献，但总体上仍存在三点不足：一是更多围绕内容、平台、算法或 AI 系统等单一对象展开，缺少对当今具有数智特征的网络生态整体运行逻辑的刻画；二是较少解释人工智能如何沿着生成、组织、认知、约束和修复等环节连续改变网络生态的过程机理；三是国际经验讨论较多停留在制度介绍层面，对中国情境中的转化边界与适配条件分析不足。

基于此，本文试图回答四个问题：其一，“人工智能+”如何推动网络生态从互联网生态转向数智生态？其二，人工智能通过哪些环节改变网络生态治理体系的运行基础？其三，中国情境下这种变革面临哪些主要挑战？其四，如何在吸收国际经验基础上提出具有适配性的治理对策？

一、理论溯源与文献基础

围绕“人工智能+”网络生态治理，既有研究大体形成五条脉络：互联网治理、平台治理、算法治理与人工智能治理、信息生态与虚假信息治理、公共风险与数字政府治理。它们分别回答网络空间如何协同治理、平台如何塑造公共表达、算法与 AI 如何被问责、错误信息如何影响认知、复杂风险如何被管理等问题。

首先，从互联网治理理论看，早期研究主要围绕互联网资源分配、域名系统、协议标准、跨国治理和多利益相关方机制展开。Ostrom^[27-28]的公共池资源治理、多中心治理和社会—生态系统分析，为理解分散主体参与网络规则形成提供了基础。Denardis^[29]、Mueller^[30]、Lessig^[31]与 Dutton^[32]的研究强调，互联网是一种规则空间，代码、协议、域名和国家权力共同塑造网络空间秩序。这些研究主要面向早期的互联互通和技术协调的问题，对于人工智能所带来的大模型生成内容、智能体行动和认知操控风险解释力有限。当前的数智生态不同于早期互联网生态。早期互联网生态主要关注连接问题，比如，信息传播、平台应用和用户互动等；数智生态关注的核心则是生成与协同，强调模

型、算法、数据、智能体和人类主体共同参与网络生态运行。由此，网络生态治理的范式发生转变：治理对象从“人发布信息”扩大到“人—模型—平台—智能体共同生成内容的生态过程”；治理目标从“内容合规”上升到“生态质量”；治理策略从“事后处置”延展为“源头识别、结构干预、认知支持、制度约束和生态修复”。

其次，平台治理研究的重心则从开放互联网转向大型数字平台。随着智能社会与智能经济的加速到来，平台的属性发生变迁，它不仅是信息通道，还是通过内容审核、社区规则、接口设计、推荐机制和账号管理塑造网络表达秩序的新治理者^[33-35]，这种制度安排正在重构公共价值^[10,13]。这些研究对本文的重要启示为，人工智能赋能平台以后，平台的结构性权力进一步增强。因而，网络生态治理不能只关注内容末端，还需要关注平台结构本身这一环节。

再次，算法治理与人工智能治理研究则进一步挖掘平台治理技术发生作用的内部机制。算法治理研究强调，算法排序、搜索推荐、自动审核、用户画像和行为预测并非单纯计算过程，而是具有社会选择功能的制度化机制^[36-37]。人工智能治理研究则围绕可信 AI、伦理原则、可解释性、风险分级和生命周期治理展开^[38-41]。那么，是否仅靠伦理原则就能保证 AI 向善，这是一个具有争议的问题。Mittelstadt^[39]指出，原则本身不能替代制度安排；NIST AI RMF、欧盟 AI 法案和 G7 广岛 AI 进程则分别从组织风险管理、风险分级和先进 AI 系统责任等方面提供了制度化路径^[42-44]。不同国家在法律约束、技术标准、组织责任和协同治理方面形成了不同规则组合，未来关键在于形成规则体系之间的衔接与协同^[45]。因此，需要把秩序如何约束和适配作为一个重要环节加以考虑，用以分析数据、算法、模型、平台和智能体全链条制度责任。

再者，虚假信息、错误信息和信息生态研究也构成本文研究的网络内容分析基础。Aïmeur 等^[46]通过文献评估了虚假信息和错误信息在社交

媒体充斥的严重性,Vosoughi 等^[19]发现虚假新闻在社交网络上具有更快、更广的扩散特征,Del Vicario 等^[20]则尝试揭示了错误信息在社交网络中的传播机制,McLoughlin 等^[22]进一步说明,错误信息会利用愤怒等情绪获得传播优势,Ecker 等^[47]则强调错误信念形成后往往具有纠偏阻力。这些研究将网络生态治理的焦点从信息真假拓展到事实判断、情绪反应、社会认知与信任等层面。这启示我们,面对信息环境带来的认知扰动,要在维护公共事实基础、保障理性表达环境、促进健康社会心态和积极价值秩序等方面铆足力气,进行社会公众层面的认知重构。

最后,公共管理和风险治理文献则指出,现代风险具有不确定性、跨域性、系统性和反馈性,不能依靠单一部门、单一工具或单次处置完成治理^[25-26,48]。公共价值理论进一步提醒,技术可以提升效率,但公共治理需要同时回应公平、透明、合法性和公众信任^[49]。这意味着,追求更快识别、更快删除和更强拦截固然重要,但是也要形成能够前瞻识别、吸收冲击、修复损害和制度学习的治理体系,进行网络生态修复建设。

基于以上分析,表1从文献脉络、核心议题、主要争议与重要启示等四个方面进行总结,以便下文把这些学理研究资源转化为面向当今网络生态自身运行逻辑的解释框架。

表1 文献脉络、争议焦点与本文回应

文献脉络	核心议题	主要争议	重要启示
互联网治理	网络空间如何形成规则	开放协同与网络主权、多中心治理之间的张力	提出数智生态概念,说明AI使治理对象从连接空间转向生成生态
平台治理	平台如何塑造公共表达	平台是中立通道还是结构性权力	把平台系统设计纳入结构组织环节
算法与AI治理	算法和模型如何被问责	伦理原则、软法框架与硬法责任如何衔接	以秩序约束解释全链条制度安排
信息生态	虚假信息如何影响公众认知	事实暂时纠偏是否足以恢复公共信任	以认知建构分析事实、情绪和价值秩序
风险治理	复杂风险如何持续治理	事件暂时处置还是长期修复和制度学习	以生态修复回应修复、救济和复盘问题

二、分析框架、案例选取与作用机理

基于上述文献对话,本文提出“介质变革—结构重组—认知扰动—秩序适配—生态修复”五环分析框架,如图1所示。该框架把人工智能视为改变网络生态运行逻辑的内生变量:它首先改变内容和行动的生成介质,继而重组平台结构和注意力分配,进一步影响公众事实判断、情绪反应和价值认同,由此要求制度安排向数据、算法、模型、平台和智能体全链条延伸,并最终通过生态修复实现事实恢复、权益救济和制度学习。五环分析框

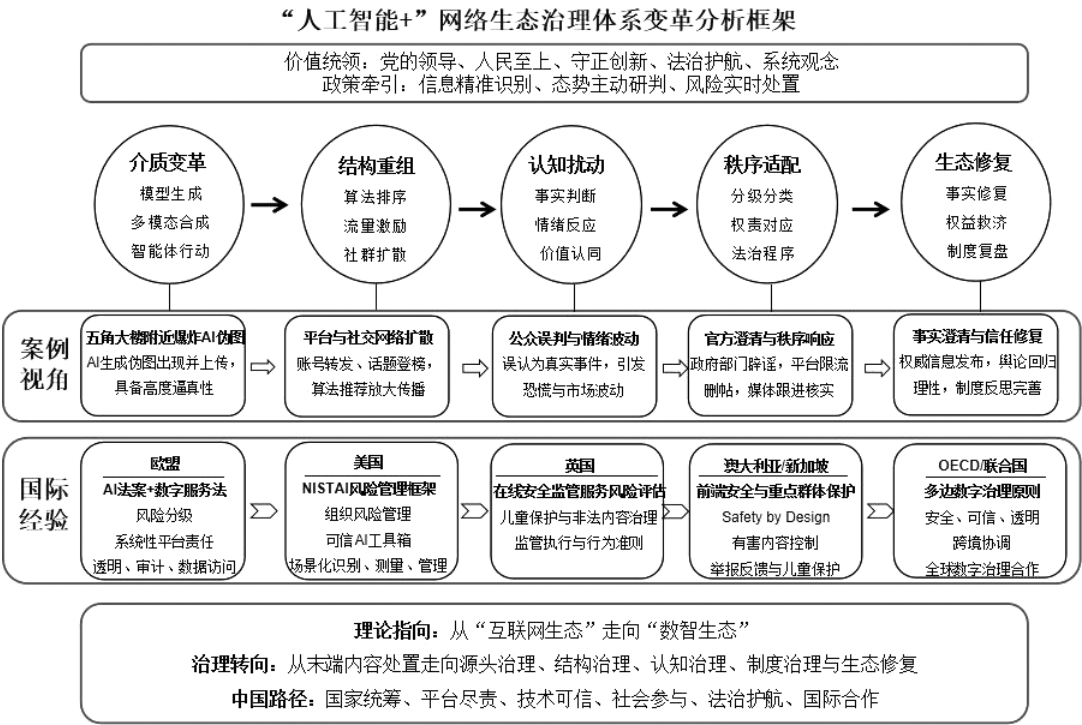


图1 五环分析框架

架与习近平总书记关于网络生态治理“五个坚持”重要论述形成内在对应。党的领导为五环分析框架治理提供统筹力量,人民至上要求介质、结构和认知治理最终服务人民群众切身利益,守正创新要求在鼓励人工智能应用的同时坚持正确方向和技术向善,法治护航要求秩序环以法律程序和责任规则保障治理正当性,系统观念则要求把介质、结构、认知、秩序和生态作为整体来治理。《国务院关于深入实施“人工智能+”行动的意见》提出的精准识别、主动研判与实时处置等要求也为五环分析框架提供政策牵引。从逻辑上讲,五环之间具有连续传导关系,任何一个环节失灵,都可能导致治理效果不彰。

基于此框架,结合“五角大楼附近爆炸 AI 伪图事件”作为案例(以下简称“案例”),分析“人工智能+”网络生态治理体系变革的作用机理。据有关美国媒体报道,从 2023 年 5 月 22 日开始,一张声称美国五角大楼附近发生爆炸的图片在社交媒体快速传播,多个认证账号转发相关内容,传言一度使标普 500 指数由涨转跌。随后,美国国防部发言人和阿灵顿消防部门相继否认发生爆炸,称五角大楼及附近没有事故或公共危险。媒体分析称,该图具有疑似 AI 生成特征,如灯柱不对称、围栏位置异常等;五角大楼附近爆炸并未发生,但伪图仍快速扩散,集中呈现了生成式 AI 虚假内容经平台传播后,对公共认知、市场情绪和治理秩序产生外溢影响的典型过程^[50]。

(一) 介质变革

介质变革是五环分析框架的起点。案例中首先体现的正是介质层面的变化,那张五角大楼附近爆炸图片让股民信以为真,引发股市恐慌。多方机构介入核查证伪后,图片被认定由人工智能技术生成。具体幕后黑手及图像原始来源尚未查明,难以锁定主要的责任主体。在介质层面,人工智能对网络生态最直接的影响,是改变内容生成的基础形态。传统的网络内容主要由自然人或组织主体以文字、图片、音视频等形式生产,其责任链条沿着发布者、账号、平台等路径传播。生成式

人工智能出现以后,内容生产由模型生成、提示词引导、多模态合成或者自动化再加工。尤其是,智能体更加先进,它不仅可以生成内容,还能实现检索信息、调用工具、执行交易、控制接口并与用户持续交互等自动执行功能。介质变革具有以下特征。

第一,内容生产规模化。过去生成大量相似但略有差异的虚假信息、营销内容或低质文本,需要较高人力成本;现在通过大模型可以快速低价生产大量同类内容。内容生产规模化后,当低成本生产与流量激励结合时,网络生态可能出现信息噪声上升、低质内容泛滥和优质内容被挤出等问题^[3,13,46]。在治理上,传统的人工审核难以应对。

第二,内容真实性边界被重构。深度合成技术可以模拟面孔、声音、动作和场景,生成虚假内容的外观真实感越来越强。“眼见为实”等经验判断变得越来越不可靠^[46-47]。由此,在治理上,不仅要判断内容真假,还要判断内容来源、生成方式、标识状态和传播语境,针对不同环节设置不同风险等级标识。

第三,责任链条复杂化。一段 AI 生成视频可能由用户输入提示词,由第三方模型生成,由应用平台包装,由数字人账号发布,由推荐算法分发,由商业机构获利。生产与传播链条长就意味着责任链条复杂^[40]。在治理上,要求制度设计从单一发布者责任扩展到模型开发者、应用部署者、平台运营者、内容发布者和传播获利者的分层责任结构。

总之,介质变革推动政策注意力从内容发布后的审核和处置前移到数据、模型、标识和工具层。介质环的治理能力,决定了整个网络生态治理能否在风险形成前获得基本识别和证据基础。

(二) 结构重组

结构环揭示平台和算法对网络生态的组织作用。案例中伪图之所以产生强大的外溢影响,除了因为图片本身真假难辨外,还因为它进入了平台化传播结构。公开报道显示,图片在数分钟内

被一批社交媒体账号,包括部分认证账号分享,进一步放大混乱,平台结构赋予其更高可见度和更强传播势能。当前的智能平台通过推荐算法、热榜排序、话题聚合、商业推广和界面设计等手段“引流”和分配注意力,并在技术指标、商业目标和用户行为反馈之间不断优化。平台的组织权力意味着它可以塑造公共议题。

结构变革带来的核心问题是,被算法化和规模化的平台系统设计具有不可知性。传统媒体时代,信息或者议题的发布与排序有明确的负责人与规则流程;平台时代,算法系统根据特定的商业目标有选择性地推荐某些内容、使某些话题登上热榜,并为某些群体塑造回音壁^[51]。其社会后果是,平台可能会带来持续放大低质、极端、沉迷和情绪化内容等一系列问题。在治理上,仅从内容端着手无法较好解决问题,需要把平台系统设计纳入治理对象。

平台结构还改变了网络内容的传播速度和传播形态。在算法时代,平台内容传播扩散不再取决于用户主动转发,而是取决于平台算法推荐。风险扩散由点状传播转向链式、网状和跨平台传播。虚假信息研究表明,情绪化内容和低可信内容在此背景下可能获得更强扩散优势^[52]。对监管或者规制来讲,平台结构的审查将成为执法重心。

平台组织结构变革机制带来启示:平台在彰显权力的同时也应该负有更多责任,不仅要对内容管理负有责任,还要对生态结构负有责任。欧盟《数字服务法》要求超大型平台开展系统性风险评估,把平台视为网络社会结构性风险的重要来源^[53]。我国网络生态治理也需要推动超大型平台从流量组织者转变为生态责任主体,通过算法影响评估、透明报告、外部审计和社会监督,防止平台风险呈现结构性放大趋势。

(三) 认知扰动

认知环是网络生态治理区别于一般网络安全治理的深层环节。案例中伪图呈现出的爆炸、恐怖袭击、军事设施等内容天然具有高风险和高情绪唤起特征,一旦图像具有现场感,用户就会形成

情绪反应,然后再等待事实核验。公共事实基础在短时间内被动摇,社会公众认知被扰动,权威机构需要消耗治理资源进行澄清和信任修复。在认知层面,人工智能改变了公众形成事实判断、情绪反应和价值认同的过程,具体表现为以下特征。

首先,影响公众判断。合成图片、仿真声音、伪造视频和机器生成文本越来越逼真,普通用户仅靠经验判断已经难以区分真假。即使平台事后辟谣或删除虚假信息,用户心理阴影仍可能存在。有研究表明,事后纠正并不总能完全消除错误信念,特别是在信息与身份认同、政治立场或情绪体验结合时^[21]。

其次,诱发情绪动员。平台算法和生成式大模型可以根据用户偏好生成更能触发恐惧、愤怒、同情或对立的内容。在“引流”为王的时代,人们的情绪可能就是商品,容易被商业化与算法化,就此,网络公共空间的理性讨论可以瞬间转变为情绪动员^[25]。尤其是,当情绪动员出现在重大公共事件、灾害事故、公共卫生、未成年人保护、民族宗教和国际冲突等敏感场景中,很容易造成舆情事件或者危机事件,增加公共治理负担。

最后,冲击价值秩序。网络上低俗、炫富、极端、攻击性和虚假内容的持续推送,将对广大用户尤其是未成年人的人生观、价值观与成功观带来不可磨灭的影响。习近平总书记^[1]强调要坚持以正面声音、主流价值、时代新风塑造网络空间,使互联网成为思想引领、道德培育、文化传承的重要阵地。这说明,认知层面的治理重在加强优质内容供给、主流价值传播和网络文明建设,重在坚持长期观与整体观。

对于网络生态治理而言,认知环要求建立事实核查、权威发布、平台提示和公众媒介素养之间的联动机制。

(四) 秩序适配

秩序环强调制度安排对网络生态运行的约束和引导。案例中所呈现的风险链条至少包括疑似AI生成图片、账号发布、认证标识误导、平台扩散、媒体跟进、官方澄清和市场反应等多个环节。在

事件的风险处置中,各方主要关注最初发布者,没有协调一致地完整回应风险形成过程。这要求制度上需要进一步区分生成者、发布者、转发者、平台分发者和获利者的责任,同时要求平台对高风险公共事件类内容设置更高的来源核验、即时提示和传播限制标准。通常情况下,这种全链条的制度安排需要做好以下适配调整。

首先需要分类分级。不同人工智能应用产生的风险具有较大差异性。普通娱乐性内容生成与公共事件新闻生成不同;个人创作辅助与未成年人陪伴型智能体不同;平台推荐普通生活内容与推荐公共政治议题不同等等。从国外经验来看,NIST AI 风险管理框架强调按组织、场景和影响管理风险;欧盟 AI 法案采用风险分级思路^[42-43]。在此背景下,我国也应在包容审慎原则下形成网络生态 AI 风险分类分级目录,至少对一般风险、中等风险、高风险和特殊风险采用差异化制度安排。

其次需要权责对应。模型开发者、应用部署者、平台运营者、内容发布者、商业推广者、技术检测机构和监管部门之间的责任要明晰^[40]。模型开发者应承担训练数据和模型安全基础责任,应用部署者应承担具体场景风险责任,平台应承担分发、审核、申诉、报告和生态修复责任,技术机构应承担检测认证责任,监管部门应承担规则制定、监督执法和公共利益保护责任。智能体责任方还需进一步明确任务授权、工具调用、行为记录和人类监督责任等。

最后需要法治护航。法治是智能社会治理获得社会信任的条件。人工智能赋能的治理手段可以缩短处理流程、提升治理效率,也可能削弱法规程序正当性。比如,自动识别、自动限流、自动封禁和自动标识技术可能会通过忽略合法表达进而削弱了程序正当性等^[13]。所以,为了防止平台私权力或算法权力的盲目扩张,依赖人工智能赋能的平台治理需要法治护航,需要在理由说明透明化、申诉渠道便捷化与复核机制人工化等方面下足功夫。

(五)生态修复

生态环是五环框架的终点,也是保证网络生

态实现长期韧性规范治理的体现。案例所呈现的处置过程体现了生态修复的重要性。事件发生后,官方机构及时澄清事实,媒体和其他事实核查机构随后进行核验,股票市场波动也在澄清后逐步恢复。但是,该案例也暴露出事件处置中的短板,即依赖事后辟谣的末端处置方法是被动的生态修复策略,容易陷入“整治—反弹—再整治”的恶性循环。这启示我们,生态修复应该从末端处置和专项整治转向长期韧性建设,进一步把预防、吸收、处置、修复和学习连成闭环。

一般来讲,生态修复层面包括以下能力:第一是预防能力,即在产品/设计、模型上线、平台规则和内容供给环节嵌入安全和价值强制标准与要求。比如,澳大利亚 Safety by Design 强调在问题发生前通过设计预判、检测和消除在线伤害^[54]。第二是吸收能力,即当风险发生时,平台系统以及整个监管系统能够快速识别、降低扩散、保护受害者并保持各方沟通畅通^[25]。第三是处置能力,即依法采取标识、澄清、限流、阻断、证据固定、联合执法等措施^[13]。第四是修复能力,即对事实、名誉、心理、信任等方面的不良影响进行修复^[46]。第五是学习能力,即把重大事件处理过程复盘,并根据成功经验进行规则调整、技术补强和责任改进等^[7]。

尽管生态修复无法做到将风险完全消除,但是它强调在风险冲击中保持基本秩序、及时恢复并不断学习,把正向内容供给、风险识别、平台责任、公共沟通、受害者救济、制度复盘和国际合作连接起来,由此在保持网络社会表达活力的同时,最大程度防止违法侵权、虚假操纵、情绪极化和价值失序,形成长期稳定的治理。

值得指出的是,五环分析框架之间具有耦合关系。介质变革改变风险源头,结构重组改变风险扩散,认知扰动改变风险后果,秩序适配革决定治理边界,生态修复决定治理体系的韧性。与既有研究框架相比,它具有三个优势。第一,它是从网络生态运行过程出发,回答了人工智能如何改变网络生态的介质、结构、认知、秩序和生态。第

二,它能够同时容纳国际经验和中国情境,在学理与经验方面均具有一般性。第三,它具有政策可转化性,每一环都能对应具体治理工具,从而避免空谈。

三、中国情境下的主要挑战

尽管我国已经形成较为丰富的网络生态治理实践和制度基础,专项整治、平台责任、算法规制、数据安全、个人信息保护、生成式人工智能服务管理等制度工具不断完善,但是人工智能驱动的网络生态风险具有生成快、扩散快、跨域强、修复难等特征,具有数智特征的网络生态整体运行仍然面临一些关键挑战。

(一)网络内容生成源头识别不足

习近平总书记^[1]强调,要深化党的创新理论上宣传,大力弘扬社会主义核心价值观,推出更多有内涵、有温度、有影响的网络作品。网络空间既是信息传播和社会交往空间,也是思想引领、道德培育、文化传承和文明建设的重要阵地。我国互联网生态治理长期强调内容生产与价值引领之间的平衡。

生成式人工智能突破了人们对于内容创作的传统认知,逐渐成为写作助手、检索工具和知识来源,其广泛应用重塑了内容生产模式,改变了受众的信息接收和获取方式,网络传播环境发生显著变革^[55]。生成式人工智能降低了内容生产门槛,使用者只需输入自然语言提示词即可生成文本、图片、音频、视频等内容,推动网络内容生产进入模型化生产阶段。与此同时,低成本、规模化、标准化的生成能力也可能被用于制造低质文本、伪造图片、自动化营销、网络水军内容和情绪化叙事。

尽管当前的制度设计已经开始重视生成式人工智能、深度合成和内容标识问题,但在现实治理中,源头识别仍存在三个不足:一是标识机制尚未充分覆盖二次剪辑、截图搬运、转码传播和跨平台扩散等场景;二是模型测评、日志留存、内容溯源与平台处置之间尚未形成稳定衔接;三是智能体的权限边界、工具调用和行为记录仍需要更明确

的规则。

总之,若源头信息建设、事实核验和优质内容供给不能同步增强,网络生态则可能出现“内容丰富”与“价值引领不足”并存的矛盾状态。

(二)平台商业激励与网络生态公共价值存在张力

平台不仅是内容存储和传播通道,还是内容流动、注意力分配和公共表达的组织者。平台通过规则、接口、审核、排序和商业模式塑造公共表达。我国超大型平台深度嵌入新闻资讯、网络电商、社交互动、本地生活、搜索问答和政务服务等日常场景,平台既组织信息流,也组织消费流、服务流和社交流。平台的算法生产内嵌于算法工程师、产品经理、运营人员、资本驱动、社会需求和政府规制等多重力量之中^[56]。

一方面,平台依靠流量、广告、用户时长、打赏、带货和数据积累实现商业收益。另一方面,低质化、标题党、情绪化和极端化内容通过算法获得可见度,极易带来公共价值风险。这种风险具有以下特征:一是扩散快,短视频、直播和社交平台的推荐机制能够在短时间内把某一内容推送给大量用户;二是链条长,一个话题可能先在短视频平台形成情绪动员,再进入社交平台发酵,随后被新闻资讯平台、搜索平台和电商平台二次利用;三是外溢强,平台上的注意力竞争可能影响线下消费、市场预期、社会情绪、公共服务压力甚至公共安全。

尽管现有政策已经持续压实平台主体责任,但是如若平台算法生产的社会机制不能被有效观察和评估,促进平台责任从自我说明、形式合规和内容管理责任向生态治理责任转型依然艰难。

(三)认知扰动带来公众认知保护压力加大

人工智能介入网络生态后,公众形成事实判断、情绪反应和价值认同的环境发生变化,认知扰动成为网络生态治理的核心环节。数智生态下的治理目标,需要公共信息环境支持多样接触、支撑理性判断和锚定事实真相。理想的信息

生态强调多元流动、主动探索、协作共建,具备多元信息入口、强用户主动性、协作共建机制与生态互联性等特征^[57]。然而,平台算法把不同群体带入差异化信息环境,使公共讨论容易被情绪化、片段化和立场化叙事牵引。现实治理的困难在于,保护公众认知的智能传播环境建设仍然存在一定的挑战。

首先,公共事实判断成本提高。过去,公众识别信息真伪主要依赖于信息来源、官方证据、社会常识或者权威解释等渠道。现在,人工智能生成内容“以假乱真”并随处可见,普通用户判断信息真假无从下手,成本或显著提高。另外,未经事实核查的伪造信息一旦进入情绪传播链条,即使后续被辟谣、删除或者澄清,其后果也已经影响了群体态度与认知。这就意味着,删除虚假内容不等于恢复事实秩序。

其次,情绪动员的识别与处置更加困难。公共事件中公众的愤怒和焦虑,既可能反映公众的真实关切,也可能被部分账号利用来吸引流量、制造对立。这使治理面临两难:只审核内容真假,可能遗漏借助真实事件煽动对立的行为;仅根据情绪倾向采取限流、删帖等措施,又可能限制正常批评和合理诉求。因此,如何形成有依据、有区别的处置机制,是公众认知保护环节面临的现实挑战。

最后,未成年人和数字弱势群体的认知保护压力上升。不同群体在媒介理解、风险识别和自我保护能力上存在差异。当前,未成年人进入短视频、网络游戏、直播和智能陪伴等网络场景的年龄提前,其价值判断、审美趣味和社会认知更容易长期受到网络上有关伪造信息、情感诱导、金融诈骗和算法推荐的影响。

(四) 全链条治理制度设计较难覆盖责任主体、贯穿责任链条

当前,数字化、网络化、平台化、智能化的发展大势正在使传播生态逐渐脱离人类中心的交往模式,转向人与机器等多样主体共生的语境^[58]。智能工具时代正在迈向智能主体时代,智能体由“能

说会写”走向“能干会做”,具备任务分解与规划、记忆与知识管理、工具调用与执行、多智能体协同能力^[59]。生成式人工智能和智能体进入网络生态以后,治理对象不仅只是自然人与组织发布的内容,还包括模型生成的内容、算法分发的内容、智能体执行的行为以及人机混合形成的网络互动。

人工智能应用场景中,模型开发者、数据提供者、算力服务者、插件开发者、应用部署者、平台运营者、账号主体、商业推广者、技术检测机构和监管部门都可能参与风险形成。另外,不同场景中的人工智能应用风险差异很大,同一强度的监管要求普遍不能适用诸如创作辅助、娱乐互动、商业客服、新闻生成、教育陪伴、医疗问答、金融营销、公共事件传播、未成年人服务和智能体自动执行等各种场景。人工智能驱动的网络生态风险具有链条性和跨域性^[60]。因此,单项制度之间如果衔接不足,就可能出现规则覆盖不完整、责任边界不清晰和执行标准不统一等问题。网络生态治理前端预防能力的形成,需要宣传、网信、公安、市场监管、教育、文化、民政、金融等部门实现常态化跨部门协同,而不是事后启动协同机制。

总的来看,秩序环的关键在于既有规则如何转化为覆盖源头、过程、传播与处置的整体性安排,使得制度体系真正适应人工智能驱动下网络生态风险的链条化演变。

(五) 网络生态长效机制仍需完善

党的十八大以来,我们把网络生态治理摆在重要位置,不断壮大网上主流价值、主流舆论、主流文化,有效打击遏制网络乱象,网络生态总体向上向好^[1]。同时,当前人工智能驱动的网络生态风险具有工具更新快、主体迁移快、形态变换快和传播外溢快等特点,网络谣言、网络暴力、算法滥用、数据安全、资本无序扩张等乱象仍时有发生,并呈现顽固性、反复性和新异性特征^[61]。中国情境下的主要挑战是“人工智能+”网络生态治理所要求的体系化能力与既有治理惯性之间存在适配压力。

首先,预防能力尚待提升。一些风险往往在引发舆情后才被看见,这说明先前的治理对模型上线、产品设计、平台规则、商业投放和内容推荐等环节的前置评估还不充分。尤其是在新应用、新工具、新玩法快速上线的环境中,如果没有将安全、价值和责任等要求嵌入产品设计前端,治理就只能在后续环节不断补救。

其次,社会伤害修复能力不够系统。网络空间带来的危害具有持久性特征,并不总是随着内容消失而消失。对于被深度伪造侵害的个人、被谣言影响的组织而言,权益和信任方面的修复往往需要更长时间。风险发生后,监管部门快速开展处置,平台迅速删除内容,但是受害者名誉、公众事实判断、群体情绪和社会信任如何修复,往往缺少系统性的方案。

最后,制度学习能力仍需加强。重大网络生态事件处理之后,若缺少公开透明的复盘机制、可共享的案例库、可更新的技术标准和可追责的责任链条,经验就难以转化为制度资产。平台、监管部门、技术机构和研究者之间的数据与知识如果不能有效共享,则会削弱治理体系的学习能力。网络生态治理韧性建设需要在一次次风险冲击后做到更早识别、更稳吸收、更准处置、更好修复。

值得指出的是,以上五个主要挑战是一组连续的治理适配难题。介质变化使风险源头前移,结构变化使风险扩散加快,认知变化使风险后果加深,秩序变化使制度衔接承压,生态缺位则使治理成效难以持续。风险贯穿内容生成、平台分发、公众认知、制度约束和生态修复全过程。这种连续性约束始终制约着智能治理能力生成。

四、对策建议

针对上述挑战,结合中国网络生态治理的制度条件和实践场景,充分分析国际经验,正视中国和其他国家的区别,取长补短,提出中国情境下的具体建议。

(一) 国际经验

从国际经验看,发达国家和地区围绕人工智

能安全、在线平台责任和数字治理已经形成若干制度工具,为我国提供了经验参照。第一,欧盟经验突出风险分级和系统性平台责任。欧盟《人工智能法案》采用风险分级方法,对不同风险水平的AI系统设置差异化义务^[43];《数字服务法》则要求超大型在线平台和搜索引擎识别、评估和缓解系统性风险,并接受透明度、审计和数据访问等要求^[53]。这种治理策略对我国具有启发意义:人工智能网络生态治理不能只盯着末端内容,也要把模型风险、平台结构和系统性影响纳入制度框架。其局限也较为明显:欧盟模式建立在特定的法律传统和市场环境基础上,合规成本较高,面对超大规模平台生态,可能难以兼顾创新速度、治理效率和公共价值。

第二,美国经验更强调风险管理框架、行业标准和组织合规。NIST AI 风险管理框架(AI RMF)把可信性要求嵌入AI产品、服务和系统的设计、开发、使用与评估过程^[42],强调由组织根据场景识别、测量和管理风险。我国在模型测评、算法审计、风险报告和技术标准建设中可以借鉴它的工具箱。但其不足在于,自愿性框架较依赖企业内生合规动力,面对具有公共性、舆论性和安全性的网络生态风险时,组织自律形成的约束力具有局限性。

第三,英国在线安全监管经验重在服务风险评估和监管执行。英国通信管理局(Ofcom)围绕非法内容和儿童保护等领域推进风险评估、行为准则和合规要求,体现出其将平台服务特征、用户群体和危害类型结合起来进行治理的思路^[62]。这一经验说明,在线安全不仅要按内容类别治理,还要分析服务设计造成伤害发生的过程。对我国而言,可以适度借鉴其风险评估方法。

第四,澳大利亚与新加坡经验更重视前端安全设计和重点群体保护。澳大利亚提出Safety by Design,强调在产品阶段预判、检测和减少在线伤害^[54];新加坡在线安全实践守则突出社交媒体服务对有害内容扩散的控制,尤其强调儿童保护^[63]。这些做法提示我国,应将进一步强调未成

年人保护、用户赋权、举报反馈和前端安全嵌入平台产品设计,减少把风险治理后置于投诉举报和行政处置环节。其局限在于,这类面向特定平台和重点伤害类型的制度,还需进一步拓展到智能体行动等更复杂的场景中。

第五,OECD 和联合国等多边机制强调安全、可信、透明和国际协调。OECD AI 原则、联合国全球数字契约等文件为跨国人工智能治理提供了原则框架^[64-65]。这些经验为我国带来启示,要在跨境数据安全和国际网络犯罪治理等方面加大参与力度,乃至主导国际规则建设。虽然多边机制通常约束力有限,但是需要把国际规则合作转化为提升网络生态治理能力和国际话语权的重要渠道。

综上,国际经验可以概括为五类可转化工具:风险分级、系统性风险评估、组织风险管理、前端安全设计和跨国规则协调。值得注意的是,不同规则体系的制度重心、主体责任、治理工具和协同方式并不相同,必须结合本国制度条件加以转化^[45]。我国与上述国家和地区存在重要差异。我国网络生态具有用户规模大、平台应用深、线上线下联动强、公共价值目标明确、制度协同能力较强等特点。如果将这些政策工具借鉴应用于我国,必须坚持党的领导、人民至上、守正创新、法治护航和系统观念,将这些工具嵌入技术可信、平台尽责、社会参与、制度保障和生态修复相结合的治理体系,服务于网络强国建设、网络文明建设、未成年人保护和国家安全等综合目标。

(二) 对策建议

1. 以源头信息建设牵引介质环的治理,构建生成内容标识、模型测评与智能体权限管理体系

面对介质变革带来的挑战,需要把风险识别从内容发布之后前移到模型、工具使用与内容生成环节。首先,应完善人工智能生成合成内容标识制度,推动显式标识、隐式标识、元数据记录和跨平台识别标准相互衔接。特别针对未成年人服务等特殊群体应用场景,应提高标识要求和人工复核要求,引导用户辨别真伪。

其次,应建立与网络生态场景相适配的大模型测评制度,不仅考察大模型通用能力和安全性,还要考察其在谣言生成、深度伪造、极端内容、诱导消费、隐私泄露和未成年人保护等方面的风险表现。

最后,还应特别关注智能体的自动执行边界。对可自动发布、自动交易、自动调用接口或者自动执行任务的智能体,应建立权限分级、任务授权、行为日志和异常终止机制。智能体不能以用户授权为由无限扩展行为边界,也不能让平台和模型开发者以工具中立为由回避安全义务。

2. 以平台生态责任牵引结构环的治理,推动平台系统性风险评估、算法审计与透明报告

结构环的治理的重点应该是推动平台责任从内容管理责任转向生态治理责任。首先,应当把超大型平台的推荐机制、热榜规则、广告投放、直播分发和社群组织纳入风险评估范围。平台不仅要违法内容负责,也要对系统设计可能造成的风险放大负责。建议按照平台规模、用户类型、内容影响和公共属性,建立分层分类的平台生态风险评估制度。对影响面广、用户规模大、涉及公共议题和未成年人场景的平台,应定期提交风险评估报告,对算法推荐、热榜排序、广告投放和内容审核如何影响网络生态进行说明,做到从结果审查扩展到过程审查。

其次,算法审计也需要重视算法生产过程的可监督性。监管应关注平台内部产品目标、运营反馈、商业激励和政府规制如何共同影响算法生产。可探索建立第三方技术评估、行业标准测试、专家委员会审查和用户代表参与的外部监督机制,在保护商业秘密和数据安全的前提下,提高平台生态责任的可验证性。

最后,建立平台生态风险的常态化透明报告制度。报告内容不仅包括内容处置数量,还应包括推荐系统调整、热榜干预、异常流量识别、未成年人保护、用户申诉、风险复盘和生态质量改善情况。透明报告不仅是规范企业社会责任的文本,还要成为可比较、可核查、可追责

的制度工具。

3. 以数智媒介素养建设牵引认知环的治理,健全公共事实支持与主流舆论塑造机制

认知环的治理目标需要建设能够支撑公共事实、理性讨论和价值认同的信息生态。第一,健全公共事实支持体系。对于重大公共事件、灾害事故、公共卫生、未成年人保护、民族宗教和国际冲突等容易引发情绪扩散的议题,应建立快速、专业、可理解的权威信息发布机制。权威发布不应只给出结论,还要解释证据来源、判断过程和不确定性边界,减少信息空白带来的猜测。事实核查也应与平台传播机制联动,对已证伪信息提供纠错提示、传播链条降噪和相关用户提醒;对尚未查清但风险较高的信息,应通过风险标注、减少算法放大和引导权威来源降低误导性传播。

第二,推动数智媒介素养教育常态化。面向未成年人、老年人、平台创作者、政务人员和普通用户,应开展分层次教育,提升公众理解AI生成内容、深度合成、推荐算法、智能体行为和网络诈骗等基本风险的能力。媒介素养教育一方面需要引导用户识别真假,另一方面需要培养其主动探索多元信息来源、参与协作核验、保持公共理性的能力。

第三,增强主流舆论塑造能力。主流价值供给要适应平台化、视频化、互动化和智能化传播环境,提高叙事能力、互动能力和场景适配能力。可以鼓励主流媒体、政务媒体和专业机构运用人工智能工具生产兼具真实性、思想性和可传播性的内容,使主流价值进入用户日常接触的信息流、短视频、直播、问答和互动社区。由此推动认知环的治理从被动纠偏进一步走向主动建设。

4. 以全链条制度安排推动秩序环的治理,完善分类分级、权责对应制度和法治程序

秩序环治理的关键是使制度安排能够覆盖数据、模型、算法、平台和智能体等全链条。首先,应建立网络生态AI风险分类分级目录,避免对所有场景采用同一强度的治理工具。低风险场景可重在提示、标识和企业自律;中等风险场景应加强备

案、透明报告和用户保护;高风险场景应引入安全评估、人工复核、日志留存和应急预案;涉及国家安全、未成年人保护、公共事件传播和关键基础设施等特殊场景,则应设置更严格的准入和持续监管要求。

其次,应建立权责对应的分层责任体系。模型开发者应承担训练数据合规、模型安全测试和风险提示责任;应用部署者应承担场景适配、用户说明和使用管理责任;平台运营者应承担分发审核、申诉救济、日志保存、风险报告和生态修复责任;商业推广者应承担营销真实性和传播获利责任;第三方检测机构应承担测评独立性和专业性责任;监管部门应承担规则制定、监督执法和公共利益保护责任。责任配置应随风险链条分类分层展开,避免风险发生后责任全部堆积到末端平台或普通用户身上。

再次,应围绕智能体建立专门治理规则。可借鉴“有籍可查、有规可循、有责必问、有界可依”的思路^[59],对智能体实行登记备案、身份标识、权限分级、行为日志、工具调用审查和人工监督,对能够自动发布、自动交易、自动调用接口的智能体,应设置更加严格的任务授权和异常终止机制。智能主体时代的法治程序,既要管内容也要管行动,既要识别输出也要记录过程。

最后,应用法治护航人工智能赋能治理。人工智能可以用于初筛、预警、分类和辅助研判,但涉及表达权利、重大利益、公共事件和账号处置的决定,不宜完全自动化。应建立健全人机共生协作生产的程序链条。对于自动标识、限流、降权、封禁等措施,平台应提供必要说明和便捷申诉渠道;对于错误处置,应建立纠错和恢复机制,使技术效率嵌入法治程序。

5. 以长效机制促进生态环的标本兼治,形成前端预警、受害者救济、制度复盘与网络免疫的闭环

生态修复的目标,是把专项治理成果转化为长期能力。首先,应强化前端预警防范,把安全、价值和责任要求嵌入产品设计、模型上线、平台规

则、商业投放和内容推荐环节。借鉴前端安全设计理念,把风险预判、伤害检测、用户保护和平台自查融入产品生命周期,减少把风险治理完全后置于投诉举报和行政处置环节。

其次,应完善生态修复和受害者救济机制。对深度伪造、网络谣言、网络暴力、恶意营销和虚假动员造成的损害,除了运用删除内容和处罚账号等政策工具外,还应推动事实恢复、名誉修复、心理支持、权利救济和平台纠错等策略的运用。对于影响较大的虚假信息,平台应在原传播链条上提供澄清提示,避免辟谣信息只停留在另一个信息空间。要建立受害者友好的救济机制。对于未成年人、老年人和数字弱势群体,应设置更加便捷的保护通道。救济机制除了要求用户自行举证外,也应利用平台技术能力和监管协同机制帮助识别侵害来源。

再次,应强化制度复盘能力。重大网络生态事件处理之后,应围绕“风险如何生成、平台机制是否放大、模型工具是否被滥用、规则是否存在空白、处置是否及时、公众沟通是否有效”等问题进行复盘。复盘结果要转化为规则修订、技术补强、平台整改和行业标准更新的依据。通过建立典型案例库和风险知识库,在合规前提下供监管部门、平台、研究机构和行业组织共享,推动经验沉淀为制度资产。

最后,从长远看,生态修复还需要形成网络免疫能力。网络免疫意味网络生态在遭遇虚假操纵、情绪极化、算法滥用和智能体失控等冲击时,能够及时发现、稳住秩序、修复损害并获得学习。只有把预防、吸收、处置、修复和学习连接起来,网络生态治理才能从清除风险走向增强韧性。

6. 以命运共同体理念牵引国际合作,提升中国经验的规则表达能力

人工智能网络生态风险具有跨国性。开源模型、跨境平台、跨国诈骗、深度伪造、数据流动和算法服务很难完全在一国内运行。我国应积极参与 AI 生成内容标识、深度合成检测、模型安全测评、

智能体治理、未成年人在线保护、跨境网络犯罪打击和数据安全等领域的国际规则协调。在参与国际合作时,应坚持网络主权、数据安全、发展权和公共利益并重原则。

同时,要把中国网络生态治理经验转化为可解释、可交流、可比较的知识表达。中国路径的特点在于国家统筹、平台尽责、技术可信、社会参与、法治护航和生态修复相结合。未来可围绕五环分析框架形成中国自主知识体系的表达:以介质治理回应生成源头问题,以结构治理回应平台放大问题,以认知治理回应公共事实和价值秩序问题,以秩序治理回应智能主体和制度衔接问题,以生态修复回应长效机制和数字文明建设问题,推动中国经验有力服务智能经济与智能社会建设。

总体而言,“人工智能+”网络生态治理体系既要充分利用人工智能提升网络生态治理能力,又要避免把治理责任推给技术系统本身;既要借鉴国际经验中的制度工具,又要把握中国网络生态的规模、价值目标和制度条件;既要重视风险处置速度,又要重视事实恢复、权益救济和生态修复。只有把源头信息建设、平台生态责任、公共事实支持、全链条秩序和长效机制贯通起来,才能推动网络生态治理走向成熟稳定的体系化治理。

参考文献:

- [1] 习近平. 习近平在中共中央政治局第二十三次集体学习时强调健全网络生态治理长效机制持续营造风清气正的网络空间[N]. 人民日报,2025-11-30(1).
- [2] 国务院. 国务院关于深入实施“人工智能+”行动的意见[EB/OL]. (2025-08-21) [2026-05-01]. https://www.gov.cn/gongbao/2025/issue_12266/202509/content_7039598.html.
- [3] 苏竣. 科技进步与智能社会治理[J]. 数字经济发展与治理,2023(1):143-146.
- [4] 张洪忠,卜嘉敏,张思滢. AI 大模型驱动的全球互联网新生态:基于信息流的空间格局、用户版图与内容生产分析[J]. 新闻与写作,2025(12):5-17.
- [5] 王志刚. 资本逻辑与社会秩序异化[J]. 长白学刊,2006(5):55-57.
- [6] 王江蓬. 中国互联网生态治理的演进脉络、内在逻辑

- 及基本经验[J]. 中国编辑,2022(11):61-67.
- [7] EMERSON K, NABATCHI T, BALOGH S. An integrative framework for collaborative governance[J]. *Journal of public administration research and theory*, 2012, 22(1):1-29.
- [8] 汪旭晖, 曹学义. 从单向规制到多元适配:互联网平台企业现代治理的多维政策逻辑与实施路径[J]. *中国软科学*, 2025(12):118-130.
- [9] GORWA R. What is platform governance? [J]. *Information, communication & society*, 2019, 22(6): 854-871.
- [10] VAN DIJCK J, POELL T, DE WAAL M. The platform society: public values in a connective world[M]. Oxford: Oxford University Press, 2018:45.
- [11] FLEW T, MARTIN F, SUZOR N. Internet regulation as media policy: rethinking the question of digital communication platform governance[J]. *Journal of digital media & policy*, 2019, 10(1):33-50.
- [12] 韩炜, 彭波. 平台如何推动治理机制的演化:基于四家平台企业的多案例研究[J]. *中国软科学*, 2026(2): 161-173.
- [13] GORWA R, BINNS R, KATZENBACH C. Algorithmic content moderation: technical and political challenges in the automation of platform governance[J]. *Big data & society*, 2020, 7(1):1-15.
- [14] MITTELSTADT B D, ALLO P, TADDEO M, et al. The ethics of algorithms: mapping the debate[J]. *Big data & society*, 2016, 3(2):1-21.
- [15] 肖红军. 算法责任:理论证成、全景画像与治理范式[J]. *管理世界*, 2022, 38(4):200-226.
- [16] 薛澜. 新兴科技发展中的人工智能治理[J]. *中国网信*, 2022(1):48-51.
- [17] CRAWFORD K. The atlas of AI: power, politics, and the planetary costs of artificial intelligence[M]. New Haven & London: Yale University Press, 2021:12.
- [18] CATH C, WACHTER S, MITTELSTADT B, et al. Artificial intelligence and the ‘good society’: the US, EU, and UK approach[J]. *Science and engineering ethics*, 2018, 24(2):505-528.
- [19] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online [J]. *Science*, 2018, 359 (6380): 1146-1151.
- [20] DEL VICARIO M, BESSI A, ZOLLO F, et al. The spreading of misinformation online [J]. *Proceedings of the national academy of sciences*, 2016, 113(3):554-559.
- [21] MORAVEC P, MINAS R, DENNIS A R. Fake news on social media: people believe what they want to believe when it makes no sense at all [J]. *MIS quarterly*, 2019, 43 (4): 1343-1360.
- [22] McLOUGHLIN K L, BRADY W J, GOOLSBEE A, et al. Misinformation exploits outrage to spread online [J]. *Science*, 2024, 386(6725):991-996.
- [23] 李冲, 李霞. 人工智能风险的试探性治理:概念框架与案例解析[J]. *中国软科学*, 2024(4):91-101.
- [24] GIL-GARCIA J R, DAWES S S, PARDO T A. Digital government and public management research: finding the crossroads [J]. *Public management review*, 2018, 20(5):633-646.
- [25] COMFORT L K. Crisis management in hindsight: cognition, communication, coordination, and control [J]. *Public administration review*, 2007, 67:189-197.
- [26] ZUIDERWIJK A, CHEN Y C, SALEM F. Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda [J]. *Government information quarterly*, 2021, 38(3):101577.
- [27] OSTROM E. Governing the commons: the evolution of institutions for collective action[M]. Cambridge: Cambridge University Press, 1990:30.
- [28] OSTROM E. A general framework for analyzing sustainability of social-ecological systems [J]. *Science*, 2009, 325 (5939):419-422.
- [29] DENARDIS L. The internet in everything: freedom and security in a world with no off switch [M]. New Haven & London: Yale University Press, 2020:178.
- [30] MUELLER M L. Against sovereignty in cyberspace[J]. *International studies review*, 2020, 22(4):779-801.
- [31] LESSIG L. The law of the horse: what cyberlaw might teach[J]. *Harvard law review*, 1999, 113(2): 501-549.
- [32] DUTTON W H. The fifth estate emerging through the network of networks[J]. *Prometheus*, 2009, 27(1):1-15.
- [33] GILLESPIE T. The politics of ‘platforms’ [J]. *New media & society*, 2010, 12(3):347-364.
- [34] KLONICK K. The new governors: the people, rules, and processes governing online speech[J]. *Harvard law review*, 2018, 131: 1598-1670.
- [35] GILLESPIE T. Platforms intervene[J]. *Social media + society*, 2015, 1(1): 1-2.
- [36] JUST N, LATZER M. Governance by algorithms: reality

- construction by algorithmic selection on the internet [J]. Media, culture & society, 2017,39(2):238-258.
- [37] EUBANKS V. Automating inequality: how high-tech tools profile, police, and punish the poor [M]. New York: St. Martin's Press, 2018:181.
- [38] JOBIN A, IENCA M, VAYENA E. The global landscape of AI ethics guidelines [J]. Nature machine intelligence, 2019,1(9):389-399.
- [39] MITTELSTADT B. Principles alone cannot guarantee ethical AI [J]. Nature machine intelligence, 2019,1(11):501-507.
- [40] GASSER U, ALMEIDA V A F. A layered model for AI governance [J]. IEEE internet computing, 2017,21(6):58-62.
- [41] TAEIHAGH A. Governance of artificial intelligence [J]. Policy and society, 2021,40(2):137-157.
- [42] TABASSI E. Artificial intelligence risk management framework (AI RMF 1.0) [R]. Gaithersburg, MD: National Institute of Standards and Technology, 2023:8.
- [43] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2024/1689 of the European parliament and of the council [Z]. Luxembourg: Official Journal of the European Union, 2024.
- [44] G7. Hiroshima Process international guiding principles for organizations developing advanced AI systems [R/OL]. (2023-10-30) [2026-09-02]. <https://www.mofa.go.jp/files/100573471>.
- [45] 吴建南, 马太平, 周磊. 人工智能安全治理的规则体系:国际比较与协同进路 [J]. 中国行政管理, 2024,40(12):6-14.
- [46] AÏMEUR E, AMRI S, BRASSARD G. Fake news, disinformation and misinformation in social media: a review [J]. Social network analysis and mining, 2023,13:30.
- [47] ECKER U K H, LEWANDOWSKY S, COOK J, et al. The psychological drivers of misinformation belief and its resistance to correction [J]. Nature reviews psychology, 2022,1(1):13-29.
- [48] RENN O. Risk governance: coping with uncertainty in a complex world [M]. London: Earthscan, 2008:48.
- [49] BANNISTER F, CONNOLLY R. ICT, Public values and transformative government [J]. Government information quarterly, 2014,31(1):119-128.
- [50] 央视网. “五角大楼附近爆炸图”疯传! 美国防部发言人辟谣! AI 图惹的祸? [EB/OL]. (2023-05-23) [2026-05-01]. https://news.cnr.cn/native/gd/20230523/t20230523_526261037.shtml.
- [51] PARISER E. The filter bubble: what the internet is hiding from you [M]. New York: The Penguin Press, 2011:21.
- [52] PENNYCOOK G, RAND D G. The psychology of fake news [J]. Trends in cognitive sciences, 2021,25(5):388-402.
- [53] THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2022/2065 of the European parliament and of the council of 19 October 2022 on a single market for digital services and amending directive 2000/31/EC (digital services act) [Z]. Luxembourg: official journal of the European Union, 2022.
- [54] ESAFETY COMMISSIONER. Safety by design [EB/OL]. (2026-08-01) [2026-08-30]. <https://www.esafety.gov.au/industry/safety-by-design>.
- [55] 谢新洲, 张静怡. 智能传播时代的网络内容风险与治理路径 [J]. 中国网信, 2026(1):48-51.
- [56] 何晶, 李瑛琦. 算法何以生成? 中国互联网平台企业算法生产实证研究 [J]. 新闻与传播研究, 2024,31(2):67-81.
- [57] 胡泳. 信息蜂房:互联网信息生态的范式转变 [J]. 南京社会科学, 2026(1):107-116.
- [58] 肖珺. 人机社会网络生态治理的难点及对策 [J]. 人民论坛, 2026(7):53-57.
- [59] 韩海庭, 毕然. 迈向智能主体时代的网络空间治理体系创新 [J]. 中国网信, 2026(3):72-75.
- [60] 彭小宝, 侯思涵, 程慧文. 数智治理的理论模型框架与链式发展路径:基于 B 区治理模式创新的探索性案例研究 [J]. 管理世界, 2025,41(8):143-165.
- [61] 孙采芳, 姬志闯. 网络生态治理长效机制的健全与完善 [J]. 理论视野, 2026(4):53-59.
- [62] OFCOM. Illegal content duties under the Online Safety Act [EB/OL]. (2026-06-25) [2026-08-30]. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/illegal-content-duties-under-the-online-safety-act>.
- [63] INFOCOMM MEDIA DEVELOPMENT AUTHORITY (IMDA). Online safety assessment report 2024: designated social media services [R]. Singapore: IMDA, 2025:4.
- [64] OECD. Recommendation of the council on artificial intelligence, OECD/Legal/0449 [Z]. Paris: OECD, 2024.
- [65] UNITED NATIONS. Global digital compact [Z]. New York: United Nations, 2024.