项目名称：淄博市社科规划项目研究成果（项目批准号：25ZBSK031）

生成式人工智能对网络意识形态安全的影响及应对策略探究

张丹　赵象孟　王兆宇　李 瑾　庞奇隆

淄博市互联网服务保障中心　山东　淄博　255000

摘要：在生成式人工智能背景下，网络意识形态安全呈现政治性、隐蔽性、渗透性等特征。生成式人工智能面临意识形态渗透、算法偏差等风险隐患。应从技术治理、制度规范与话语创新三维度构建协同应对体系，实现从被动防御向主动治理的转变。

关键词：生成式人工智能；网络意识形态安全；协同治理

Research on the Risks,Causes and Governance Paths of Network Ideological Security in the Context of Generative Artificial Intelligence

Abstract: Under the background of generative artificial intelligence,the security of online ideology presents characteristics such as political nature,concealment, and infiltration.Generative artificial intelligence is confronted with risks and hidden dangers such as ideological infiltration and algorithmic bias.A collaborative response system should be constructed from three dimensions:technical governance,institutional norms,and discourse innovation,to achieve a transformation from passive defense to active governance.

Key words: Generative artificial intelligence; Network ideological security; Collaborative governance

近年来，生成式人工智能的快速发展引发了学界对网络意识形态安全的广泛关注，逐渐形成多维度研究格局。当前网络意识形态安全面临生成式人工智能隐性价值导向、信息数据污染、信息茧房等风险挑战。应建立算法治理、价值引导、制度规范等多元协同治理路径，以维护我国网络意识形态安全。

1.生成式人工智能背景下网络意识形态安全的特征及存在不足

1.1 网络意识形态安全缺失的特征。政治性体现为技术内嵌的价值偏向，主流模型训练数据主要源于西方平台，导致模型在处理敏感议题时存在系统性意识形态偏差，成为西方意识形态渗透的技术工具；隐蔽性表现为润物无声的价值渗透，生成式人工智能输出高度拟人化和决策过程的“黑箱”特性，用户难以追溯其价值逻辑；渗透性体现为跨域传播的影响，英语语料的绝对优势造成跨文化强势渗透，在多个领域形成广泛影响。

1.2 当前研究存在的不足。系统性理论框架尚待构建，缺乏整体性把握与统一的理论阐释体系；技术特性与风险性的内在关联研究不深，缺少对底层架构的技术分析；跨文化传播的实证研究薄弱，针对意识形态渗透的具体机制与应对策略研究不足；治理路径可操作性有待加强，缺乏可量化的具体措施及协同机制。

2.生成式人工智能引发网络意识形态风险的原因分析

2.1 对网络意识形态的渗透风险。在全球数字化竞争格局下，生成式人工智能成为意识形态渗透的重要工具，加大网络意识形态安全压力。部分西方国家凭借技术先发优势，依托生成式人工智能技术，通过批量生成虚假信息、片面解读内容，借助技术输出、产品推广等渠道，强行输出自身意识形态与价值观。同时，其通过掌控国际技术标准、垄断关键技术资源巩固霸权，进一步扩大渗透影响，对我国主流意识形态传播构成严峻挑战。

2.2 算法设计与价值观导向偏差。算法作为生成式人工智能的核心，其设计逻辑直接影响意识形态传播走向。算法开发者的价值观和利益诉求会渗透到设计过程，若未坚守公平客观原则，或将虚假、偏激语料注入训练数据，会导致模型生成与主流价值观相悖的内容。此外，算法个性化推荐功能容易让用户陷入信息茧房，强化认知偏见，进一步放大风险。

2.3 数据质量与隐私保护不足。数据是生成式人工智能的核心基础，数据质量不佳与隐私保护缺失，是

引发风险的重要诱因。数据来源广泛复杂，常包含错误、虚假或带有偏见的内容，会导致模型输出偏差性、有害性内容。同时，用户数据保护措施不完善，泄露风险突出，一旦被恶意利用，放大网络意识形态安全隐患。

3.生成式人工智能威胁网络意识形态安全的应对措施

3.1 筑牢技术治理防线，从源头防范意识形态风险。技术治理是防范生成式人工智能意识形态风险的关键防线。其一，积极推进人工智能技术解释化应用，使内容生成与算法推荐的运行逻辑全程公开透明、可持续追踪核查，为有效监督提供路径化支撑。其二，搭建以“智能技术治理人工智能”为核心的智能化防护体系，运用人工智能手段对生成内容开展实时监测、迅速研判、风险预警，将技术手段转化为主动防御屏障。其三，建立信息推送优化机制，定期推送权威客观、理性中立、多元包容的网络信息，打破信息茧房。

3.2 完善制度规范体系，以法治思维织密安全网络。制度规范是防范生成式人工智能意识形态风险的根本保障。其一，推进《人工智能法》等法律立法工作，明确规定对训练数据进行合规审查、算法输出与主流价值对齐等强制性要求。其二，全面落实内容标识制度，解决“哪些是生成的”“谁生成的”“从哪里生成的”等关键问题，形成全流程安全标识。其三，压实网络平台的主体责任，推动平台从流量导向的运营逻辑转向安全至上、流量次之逻辑，使其成为关键“把关人”。

3.3 创新主流话语体系，以技术赋能增强意识形态引领力。话语创新是变被动应对为主动引领的关键路径。其一，以马克思主义世界观与方法论为指引，坚守生成式人工智能“技术中立”话语背后的意识形态底色，在文化层面强化公众对主流价值的辨识能力与认同根基。其二，积极构建“AI+媒体”全链条赋能体系，依托大模型推动“选题策划—内容生成—多端分发”的新媒体供给升级，提升主流内容的触达率和影响力。其三，围绕核心议题构建数据库，将中国的理论话语转化为具有传播力的数字知识产品，使之成为我国主流意识形态传播的“助推器”。

4.结语

生成式人工智能以其技术逻辑深度重构了网络意识形态的生成与传播机制。今后，需要建立“技术治理+制度规范+话语创新”等多位一体的协同治理模式，推动从被动防御向主动治理的战略转变，最终形成多元协同共治的网络意识形态安全新格局。

参考文献：

[1]陈京春,杨历霖.生成式人工智能的意识形态风险及其法律因应[J].西安交通大学学报(社会科学版),2025,(4):10-19.

[2]代金平,覃杨杨.ChatGPT 等生成式人工智能的意识形态风险及其应对[J].重庆大学学报(社会科学版),2023,29(05):101-110.