Journal of Security and Safety Technology 安防技术, 2026, 14(2), 90-95
Published Online June 2026 in Hans. https://www.hanspub.org/journal/jsst
https://doi.org/10.12677/jsst.2026.142009

Hans汉斯

人工智能技术在网络信息伪造中的预防及应对策略研究

郑建拥*，王　茜，范　翔

北京跟踪与通信技术研究所，北京

收稿日期：2026年5月22日；录用日期：2026年6月15日；发布日期：2026年6月22日

摘　要

现代网络媒体中越来越多地体现出信息伪造对人类行为和感知的操控，人工智能(AI)的快速发展及其紧密融入个人日常生活和社会内部结构的进程，更是增强和加速了这种物理和数字域的融合。针对人工智能已经融合并裹挟着网络信息、神经、纳米和生物技术等前沿领域，成为操纵社会舆论和民众认知的重要工具等现实问题，研究人工智能技术，重点是生成式人工智能技术(GAI)在网络信息伪造的演进、发展和运用，分析总结其在现代网络信息社会中发展潜力和运行模式，认识和理解这种新兴的信息传播形式，并提出预防和应对建议。

关键词

网络信息，生成式人工智能，虚假信息，应对策略

Research on Prevention and Response Strategies of Artificial Intelligence Technology in Network Information Forgery

Jianyong Zheng*, Qian Wang, Xiang Fan

Beijing Institute of Tracking and Telecommunications Technology, Beijing

Received: May 22, 2026; accepted: June 15, 2026; published: June 22, 2026

Abstract

Modern online media increasingly reflect the manipulation of human behavior and perception through information falsification. The rapid development of artificial intelligence (AI) and its

*通讯作者。

文章引用: 郑建拥, 王茜, 范翔. 人工智能技术在网络信息伪造中的预防及应对策略研究[J]. 安防技术, 2026, 14(2): 90-95. DOI: 10.12677/jsst.2026.142009

seamless integration into personal daily life and the internal structure of society have further enhanced and accelerated this fusion of physical and digital domains. Given the reality that AI has integrated and coerced cutting-edge fields such as network information, neurology, nanotechnology, and biotechnology, becoming an important tool for manipulating social public opinion and public perception, the focus of AI technology research is on generative AI (GAI) in the evolution, development, and application of network information falsification. This involves analyzing and summarizing its development potential and operational modes in the modern network information society, recognizing and understanding this emerging form of information dissemination, and proposing preventive and responsive suggestions.

Keywords

Network Information, GAI, False Information, Response Strategy

1. 引言

当前，随着技术的发展和形势的变化，思想、技术和网络空间的交汇点正在发生急剧的变化。AI 技术在其中发挥着融合和催化的重要作用，尤其是 GAI 与大数据相结合，可以大幅改进、定制、扩大甚至流水线化的生产虚假信息，不仅降低了网络信息伪造的门槛，使其变得更加隐秘，还增强了信息传播和影响的效能。例如，个性化 AI 助手和聊天机器人现在可以让用户参与看似真实的对话，隐秘地灌输根据用户的心理特征和偏好定制的操控性内容，这种复杂的信息运作模式可以深刻影响、甚至控制个人和群体的心理、意志和理念[1]，认识、理解，最终鉴别伪造的信息对于维护社会舆论生态变得越来越重要[2] [3]。

在网络信息伪造过程中，AI 的运用建立在其技术和能力的融合之上。基础 AI 模型是可以学习优化大规模数据分析，识别和分类大量数据中的模式，结构和异常的系统，并将这些分析结果转化为表征和预测[3]，提供通用的数据表示的同时，也可针对特定任务进行微调或扩展，这是 AI 应用的基础。这种基础模型可用于分析异质数据集的各种特征和变化，包括一般的图像、文本、语言，甚至更精确的特征，如人类情感和动作[4]。而 GAI 模型则利用基础模型学习的表示和特征来生成新的内容(文本、图像、叙事、视频甚至音乐)，例如，在人脸图像上训练的 GAI 模型可以生成与训练数据集中的人脸非常相似的新的、照片般逼真的人脸。GAI 模型应用的基础在于可以与人类的日常相融合，可以监测、影响人类的生活、行为和心理。在未来，通过逐步学习和模拟人类行为，GAI 系统有望在维持模仿人类对话特征交互的同时，开发出动态内容，并在一定程度上模仿人类的社会关系[5] [6]。而 GAI 的这种自主性地捕捉、模拟和与人类行为互动的新能力[7]将成为网络信息伪造的主要推动力之一。

在如今的数字时代，网络触及的每个人都在产生大量的数字信息，这些日常行为数据可以通过 AI 计算来分析，使得分析、分类、描述以及在某种程度上预测和影响人类行为变得越来越可能，而人类的“生活模式”，包括谈话和情感、生物特征和行为，都可以成为信息伪造的燃料[3]。正如文献[8]中所描述的，“我们已经进入了一个技术时代，我们的私人和集体经历已经成为行为监视的免费素材”。

2. 发展演进

2.1. GAN 推动 GAI 普及

GAI 技术已存在数十年，但生成对抗网络(Generative Adversarial Network, GAN)的出现在 GAI 发展

历程上具有里程碑意义，这种先进的机器学习应用不仅大幅提升了合成媒体的拟真度，使其堪比耗时费力的人工制作内容，更实现了该技术的平民化应用，可以说正是 GAN 开启了 GAI 的大规模运用。而作为 Midjourney、Dall-E 和 Stable Diffusion 等工具基础的多模态 AI 系统取得突破，加之其低廉甚至免费的访问成本，使得用户通过描述性文本指令即可轻松生成图像视频。2019 年上线的 ThisPersonDoesNotExist.com 网站(基于当时流行的 StyleGAN 架构，后升级为 StyleGAN2)在信息伪造中得到广泛使用。正如其名，该平台通过人体图像数据集训练，能生成独一无二的虚拟人物肖像[9]。

2.2. 扩散模型取代 GAN

近年来扩散模型的普及已使 GAN 生成的图像相形见绌，促成了 GAI 技术再次突破，这使得媒体和公民社会倡导的诸多“认知免疫”措施(例如通过普及欺骗性技术手段来增强网民防范意识)逐渐失效。与 GAN 仅能提供现成图像不同，基于扩散模型的现代 GAI 工具赋予用户近乎无限的控制权。文本到图像的生成模式让创作者能精准定义画面主题。以 Midjourney 为代表的扩散模型甚至允许用户调整图像最细微的局部特征——而这些细节特征恰是鉴别 GAN 图像真伪的关键突破口。例如，GAN 生成的人物肖像在处理眼镜、首饰等需要对称的无机物时往往力不从心。但通过 Midjourney，不仅能精确生成对称的眼镜，还可对其进行修改或彻底移除。这种基于指令的创作模式使用户在定义主体时获得了前所未有的控制精度。

多模态模型在扩散技术基础上进一步突破，能结合文本、音视频等多维度数据解析提示词，在图像生成前建立跨模态的逻辑关联。当向 OpenAI 的 ChatGPT-4o 输入相似指令时，其输出图像显著优化：不仅准确使用大西洋理事会的品牌标识、标准字体和配色方案，更生动呈现 DFRLab 标志性的网络安全研究场景。这种提升源于多模态模型独有的“指令后推理”能力——在接收提示词后、生成图像前，系统会自主补充语境信息弥补原始指令的简略不足[10]。

3. 生成式人工智能运用分析

3.1. 核心能力

GAI 的核心在于其强大的内容生成能力。与传统 AI 不同，GAI 不仅能够处理和分析数据，还能够自主生成新的内容，这种能力使得 GAI 在信息伪造中具有巨大的潜力。GAI 通常与经过大量数据训练的基础模型相结合，在识别和理解复杂的数据模式和特征，可以自主学习和优化内容生成的过程。通过调整模型参数和输入数据，还可以生成符合特定需求的内容，如个性化的新闻文章、逼真的图像和视频等[11]。

GAI 还具备模拟人类行为和交互的能力。通过分析大量的人类行为数据和交互模式，GAI 可以学习并模仿人类的行为方式。这使得其在通过伪造信息操控认知方面具有得天独厚的优势。它可以生成看似真实的对话和互动，从而在不知不觉中影响人们的决策和行为。

3.2. GAI 在网络信息伪造中的运用潜力

GAI 在网络信息伪造中的应用非常广泛，主要在以下三方面[12]：

一是在社会舆论方面，生成虚假媒体内容。其可以制作高度逼真的虚假信息和媒体内容，通过社交媒体等渠道广泛传播。这些内容往往能够引发公众的恐慌和混乱，破坏社会的稳定和秩序。例如，在冲突地区，敌对势力可以制作关于敌方军队动向的虚假新闻和图像，引发公众的恐慌和逃难行为。

二是在目标定位方面，精准定位目标受众。其能够利用大数据分析技术，精准定位目标受众。通过对社交媒体、搜索引擎等渠道的数据进行挖掘和分析，可以了解受众的兴趣、偏好和行为模式，从而制定个性化的伪造策略，使得效果更加显著。

三是在自动化方面，实现伪造信息的规模化生产。这不仅可以大幅降低成本和难度，提高信息伪造

的效率，还可以降低对人力资源的依赖。

3.3. GAI 在网络信息伪造中的运作流程

运作流程如图 1 所示：

1) 主题标签，是社交媒体平台上用于分类和链接相关内容的关键标记。标签能够跟踪与特定主题或问题有关的讨论、情绪和趋势话题，协助了解叙事并理解不同平台上的讨论动态。

2) 叙事提取，首先收集数据并生成上下文文本/图像，接着通过过滤器删除带有相关标签或关键字但缺乏意义的内容，只保留包含争论或事实信息并符合可读性标准的文档。

3) 情绪分类，将文本作为输入，分析作者对内容的情绪，包括对可提取实体的敌意、支持和仇恨(以及更多种程度)等情绪。

4) 目标受众建模工具，是创建目标受众群体的详细资料，提供对不同受众群体对各种干预措施的反应，指导传播策略的制定。

5) 网络分析，用于理解和分类社交媒体上不断变化的传播的内容。通过实体建模分析其关系，如个人账户或标签作为节点，交互作为连接，形成揭示复杂社区结构的网络。

6) 信息环境理解，通过来自调查、访谈和小组讨论的洞察，结合在线和社交媒体数据，直接映射到信息环境评估元素上，提供对观众的感知、信念、动机和行为的整体理解。

7) 内容生成，基于 GAI 技术创建与传播活动目标相一致的实时内容[3]。

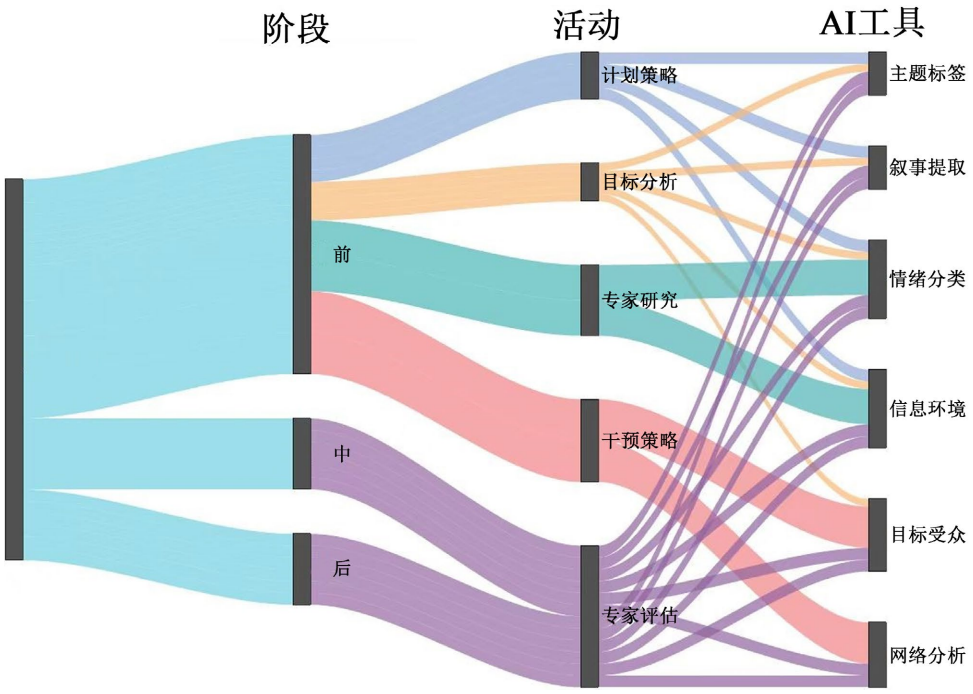

Figure 1. GAI-enabled network information forgery process
图 1. GAI 赋能网络信息伪造流程

3.4. GAI 驱动的网络信息伪造可能的后果

过去虚假信息依赖专业团队和媒体账号传播，如今普通用户输入提示词即可生成专业级虚假内容，传播门槛大幅降低，使得普通用户即可成为虚假信息生产和传播节点，在算法助推下形成“去中心化”

的信息攻击网络，严重危害网络环境。其可能带来的后果复杂且深远，最重大的危害在于对社会信任体系的系统性侵蚀。

4. 预防和应对

AI 正迅速改变网络信息环境的面貌和影响力，使其能够深入影响到平民生活和社会生态。面对这一新兴挑战和威胁，已经有不少团体致力于开发 AI 技术用于信息伪造，同时国家还采取一系列预防和应对措施来维护其社会网络环境安全稳定。

4.1. 鉴伪技术

GAI 驱动的网络信息伪造技术正以前所未有的速度演进，但相应的防御体系也在快速升级。目前最有效的应对方式是构建“主动标识 + 智能检测 + 全链路溯源”的多层次技术防线，从内容生成源头到传播终端形成闭环治理。

主动标识：这是防范 GAI 信息伪造的第一道防线，核心是在 AI 生成内容时自动嵌入可识别的标记。如模型水印，可在 AI 模型训练阶段就植入唯一性数字指纹，任何由该模型生成的内容都会携带对应标识，实现从源头追溯。

智能检测：面对高度仿真的伪造内容，仅靠肉眼已无法分辨，必须利用 AI 去识别 AI，依靠算法级鉴别。如多模态融合检测，可以结合语音语调、唇形同步、背景噪声等多个维度进行综合判断。

全链路溯源：这是当前最前沿的防护技术，目标是为每一条数字内容配备“基因身份证”。如量子密钥分发，基于量子不可克隆原理，确保内容认证密钥的安全传输，防止中间人篡改验证信息。

4.2. 社会防范

一是制定适应性政策。完善法律监管和追责机制，为 AI 工具的使用提供明确界限[13]。

二是鼓励技术研发和创新。加大对网络安全、信息监管和生物安全等领域的研发投入力度，推动相关技术的创新和应用。

三是民众信息安全科普。多国政府已将 AI 伪造信息的科普纳入国家战略，通过政策引导、技术监管与公众教育相结合的方式，向民众科普 AI 信息伪造的危害，强调提升公民的媒介识别能力，旨在提升全民数字素养，防范深度伪造等技术对社会信任体系的冲击。

5. 结语

人工智能及其能力正在对社会的各个方面产生深远的影响，这种影响在目前社会发展中极其重要。人工智能驱动的虚假信息和恶意信息活动将对社会安全稳定构成重大挑战，特别是先进语言模型的发展，以前所未有的规模和复杂程度扩展了可能的虚假信息活动的范围和效果。

然而，挑战与机遇并存。GAI 的潜力并非全然负面，其技术本身可被用于反诈教育、内容审核和数字身份保护等领域。关键在于如何构建多维度的防御体系：政府需完善法规，强制标注 AI 生成内容并建立跨境协作机制；技术开发者应研发更先进的检测工具，提升虚假信息识别能力；公众则需通过教育提升数字素养，增强对深度伪造的免疫力。未来，我们既需要警惕信息伪造的“暗流”，规避其带来的系统性风险，确保技术发展服务于社会福祉而非破坏，也需以开放心态推动技术向善向好发展，让 GAI 成为构建可信数字生态的助力，而非隐患。

参考文献

[1] Wright, D. (2019) AI and Information Warfare in 2025. 2019 *IEEE SmartWorld, Ubiquitous Intelligence & Computing*,

Advanced & Trusted Computing, *Scalable Computing & Communications*, *Cloud & Big Data Computing*, *Internet of People and Smart City Innovation* (*SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*), Leicester, 19-23 August 2019, 317-322. https://doi.org/10.1109/smartworld-uic-atc-scalcom-iop-sci.2019.00098

[2] Hutchinson, W. (2006) Information Warfare and Deception. *Informing Science*: *The International Journal of an Emerging Transdiscipline*, **9**, 213-223. https://doi.org/10.28945/480

[3] Moy, W.R. and Gradon, K.T. (2023) Artificial Intelligence in Hybrid and Information Warfare: A Double-Edged Sword. In: Cristiano, F., Broeders, D., Delerue, F., Douzet, F. and Géry, A., Eds., *Artificial Intelligence and International Conflict in Cyberspace*, Routledge, 47-74. https://doi.org/10.4324/9781003284093-4

[4] Khan, S., Paul, D., Momtahan, P. and Aloqaily, M. (2018). Artificial Intelligence Framework for Smart City Microgrids: State of the Art, Challenges, and Opportunities. 2018 *Third International Conference on Fog and Mobile Edge Computing* (*FMEC*), Barcelona, 23-26 April 2018, 283-288. https://doi.org/10.1109/fmec.2018.8364080

[5] Feldstein, S. (2023) The Consequences of Generative AI for Democracy, Governance and War. *Survival*, **65**, 117-142. https://doi.org/10.1080/00396338.2023.2261260

[6] Ferguson, A.G. (2025) Generative Suspicion and the Risks of Ai-Assisted Police Reports. *Northwestern University Law Review*, **120**, No. 2.

[7] Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Chao, L.N. and Smith, J. (2023) The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA2679-1.html

[8] Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books.

[9] Ploumis, M. (2021) Comprehending and Countering Hybrid Warfare Strategies by Utilizing the Principles of Sun Tzu. *Journal of Balkan and Near Eastern Studies*, **24**, 344-364. https://doi.org/10.1080/19448953.2021.2006005

[10] Bingle, M. (2023) What Is Information Warfare? The Henry M. Jackson School of International Studies, University of Washington. https://jsis.washington.edu/news/what-is-information-warfare/

[11] Caliskan, M. (2019) Hybrid Warfare through the Lens of Strategic Theory. *Defense & Security Analysis*, **35**, 40-58. https://doi.org/10.1080/14751798.2019.1565364

[12] Paulraj, S. and Vairavasundaram, S. (2025) Transformer-enabled Weakly Supervised Abnormal Event Detection in Intelligent Video Surveillance Systems. *Engineering Applications of Artificial Intelligence*, **139**, Article ID: 109496. https://doi.org/10.1016/j.engappai.2024.109496

[13] Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S. and Parkinson, S. (2023) A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*, **11**, 144497-144529. https://doi.org/10.1109/access.2023.3344653