第 3 卷◆第 4 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

融合语言学先验知识的 Deepfake 内容治理框架综述

李文静　潘瑶婷　唐钰婷　韦彩虹　林倩如*
广西职业师范学院
DOI:10.32629/acair.v3i4.17892

[摘　要] 生成式人工智能的迅猛发展使Deepfake等深度伪造技术构成的网络舆论风险日益加剧,对社会信任和公共安全构成严峻挑战。当前主流的、纯数据驱动的检测模型因其“黑箱”特性,存在可解释性弱、泛化能力不足的固有局限,难以应对快速演进的伪造技术。本文提出一种创新的跨学科治理范式,即将系统化的语言学先验知识深度融入Deepfake内容治理框架。文章系统剖析了纯技术路径的困境,并构建性地阐述了从微观的声学特征、中观的话语结构到宏观的叙事模式等多个维度的语言学知识如何作为治理的理论基石。在此基础上,本文首创性地提出了一个涵盖“数据层-模型层-治理层”的三层治理框架模型,详细论证了其融合策略与技术路径,旨在构建一个更鲁棒、更可信且具备持续进化能力的人机协同治理体系。最后,文章探讨了该框架面临的技术与伦理挑战,并展望了未来研究方向,旨在为构建跨学科的综合治理体系提供关键的理论参考与方法论支撑,推动治理范式从被动防御向主动免疫的根本性转变。
[关键词] 生成式人工智能；Deepfake；舆论治理；语言学先验知识；多模态检测
中图分类号：TP18　文献标识码：A

A Survey on Deepfake Content Governance Frameworks Incorporating Linguistic Prior Knowledge

Wenjing Li　Yaoting Pan　Yuting Tang　Caihong Wei　Qianru Lin*
Guangxi Vocational Normal University

[Abstract] The rapid advancement of generative artificial intelligence has intensified the risks posed by deepfake technologies, such as Deepfake, to online public opinion, presenting severe challenges to social trust and public security. Current mainstream, purely data-driven detection models suffer from inherent limitations, including weak interpretability and insufficient generalization capabilities, making them ill-equipped to counter rapidly evolving forgery techniques. This paper proposes an innovative interdisciplinary governance paradigm by deeply integrating systematic linguistic prior knowledge into the Deepfake content governance framework. The article systematically analyzes the dilemmas of purely technical approaches and constructively elaborates how linguistic knowledge across multiple dimensions—from microscopic acoustic features to mesoscopic discourse structures and macroscopic narrative patterns—can serve as a theoretical foundation for governance. Building on this, the paper pioneers a three-tier governance framework model encompassing the "data layer-model layer-governance layer," detailing its fusion strategies and technical pathways. The aim is to establish a more robust, credible, and continuously evolving human-machine collaborative governance system. Finally, the paper explores the technical and ethical challenges faced by this framework and outlines future research directions, seeking to provide critical theoretical references and methodological support for constructing an interdisciplinary comprehensive governance system, driving a fundamental shift from passive defense to active immunity in governance paradigms.
[Key words] Generative artificial intelligence; Deepfake; Public opinion governance; Linguistic prior knowledge; Multimodal detection

1 引言：从技术对抗到跨学科治理的范式转变

生成式人工智能的崛起,特别是Deepfake技术的日益成熟与普及,正在深刻重塑网络信息生态。这种技术能够以极高的逼真度生成或篡改人脸、语音与行为,使得“有图有真相”的时代

信条彻底崩塌,从而为虚假信息、舆论操纵和社会不信任打开了潘多拉魔盒。从政治人物的伪造演讲到商业领袖的虚假声明,从色情报复到金融诈骗,Deepfake技术带来的潜在危害已经渗透到社会生活的各个层面。面对这一挑战,学术界与工业界的初始回应主要集中在纯技术层面的对抗,即采用“以AI治AI”的思路,训练复杂的深度学习模型来鉴别伪造内容。然而,这种单一技术路径已经显现出明显的局限性:一方面,生成模型与检测模型陷入了一场无止境的“军备竞赛”,检测方总是滞后于生成方的技术迭代;另一方面,纯粹依赖数据驱动的模型如同一个“黑箱”,其决策过程缺乏透明性,导致结果难以被人类信任,且在遭遇未知伪造手法或经过精心设计的对抗样本时,泛化能力显著下降。

在这种背景下,语言学——这门研究人类语言结构、使用与意义的古老学科——其价值正日益凸显。语言作为人类沟通与信息传递的核心载体,其内在规律为识别虚假内容提供了丰富的、尚未被充分开发的判据。无论是语音中的微妙韵律,还是话语中的逻辑结构,亦或是宏观的叙事框架,真实的人类表达都遵循着某些深层的、难以被生成模型完美复现的模式。本文据此提出核心论点：将系统化的语言学先验知识融入Deepfake内容的治理框架,是实现从被动防御到主动免疫、从“黑箱”判定到白盒分析的关键范式跃迁。本综述将系统梳理这一交叉领域的研究现状,构建一个多层次的理论框架,并深入探讨其实现路径与未来前景,以期为网络舆论风险的防控与治理提供一条创新的学术思路与实践指南。

2 现状审视：纯技术治理路径的固有局限

2.1技术路径的发展现状

当前,Deepfake内容的治理主要依赖于基于深度学习的检测模型。这些模型通过在海量真实与伪造数据上进行训练,学习区分二者在像素或声谱层面的细微差异。从早期的基于卷积神经网络的检测器,到近年来利用时序一致性分析的方法,再到基于Transformer架构的大规模检测模型,技术路线不断演进。在理想实验环境下,这些模型在特定数据集上的检测准确率可以达到相当高的水平。例如,在某些公开基准测试中,顶尖模型的检测准确率甚至超过百分之九十五。此外,一些研究开始探索利用生物信号特征,如心跳引起的皮肤微颜色变化等生理特征,作为鉴别真伪的依据。这些技术手段在一定程度上为Deepfake内容的识别提供了有效工具。

2.2现有技术的根本局限性

尽管现有技术取得了一定进展,但其局限性是根本性的。首先是泛化能力不足的问题。一个在某个Deepfake生成模型数据集上训练得到的检测器,往往在面对由新的、未知的生成模型制作的伪造内容时,性能会急剧下降。这种“猫鼠游戏”使得治理行动永远处于被动追赶的状态。例如,在某项研究中,针对特定生成模型训练的检测器,在面对其他类型的生成模型时,检测准确率下降了超过三十个百分点。其次是可解释性缺失的困境。大多数检测模型仅能给出一个“真”或“伪”的概率分数,却无法提供令人信服的理由。当一个关键的政治人物视频被指控为伪造时,仅仅依靠一个“黑箱“模型的判定结果,难以用于公众沟通、司法取证或平台裁决。最后,随着多模态生成技术的发展,伪造内容在视觉、听觉和文本模态上的融合将愈发天衣无缝,单纯依赖底层信号分析而忽略高层语义一致性的方法,将越来越力不从心。

3 理论基石：可作为治理利器的语言学先验知识维度

3.1微观层面的声学与语音特征

在微观层面的声学与语音特征研究中,需要特别关注副语言特征的分析。这些特征包括语速变化、音量波动、发声质量等参数,能够有效反映说话者的情绪状态和认知负荷。真实语音中的这些特征呈现出符合生理规律的协调变化,而合成语音往往在这些细微特征上表现出不自然的模式。在中观层面的话语结构分析中,可以引入话语连贯性指标,如指代一致性、话题延续性等语言学指标。通过计算词汇密度、句法复杂度等参数,可以构建说话者的语言指纹,为身份验证提供依据。在宏观层面的叙事分析中,需要建立叙事结构模板库,对不同类型的虚假信息叙事模式进行归类。例如,危机叙事往往遵循特定的情节发展逻辑,而阴谋论叙事则表现出特定的论证结构特征。这些不同层面的语言学特征构成了一个多维度的检测体系,能够从不同粒度揭示伪造内容的异常特征,为构建鲁棒的检测系统提供丰富的特征支持。

3.2中观层面的话语内容与结构特征

在中观的话语内容与结构层面,语言学提供了更为丰富的分析工具。多模态一致性是核心判据之一,包括口型与语音的精确同步、面部表情与语音情感色彩的匹配、以及手势与言语强调点的协调。任何在这些方面的失调都可能暴露伪造的痕迹。逻辑连贯性则关注话语内在的逻辑,真实个体的发言通常具有明确的话题推进线索和合理的因果链条,而某些由AI生成的对话可能会出现话题跳跃突兀或违背常识的陈述。此外,身份一致性也至关重要,通过话语分析,我们可以评估发言者的词汇选择、句法复杂程度、乃至惯用口语,是否与其一贯的公众形象、教育背景和职业特征相符。

3.3宏观层面的叙事与修辞特征

在宏观的叙事与修辞层面,叙事学与修辞学为我们提供了深刻的洞察力。大量的研究表明,虚假信息往往依赖于特定的叙事模板,如“阴谋论”或“悲情动员”,并频繁使用煽动性的修辞手法,如非黑即白的二元对立、贴标签或人身攻击。通过构建此类有害的叙事与修辞模式库,我们可以对生成内容进行高层语义筛查。例如,某些特定的论证结构或情感诉求模式,与已知的虚假信息传播策略具有高度相关性。即使伪造内容的底层视听质量极高,若其叙事模式与已知的虚假信息框架高度吻合,也应触发治理系统的警报。

4 数据、模型与治理的三层融合模型

4.1数据层：语料库建设与知识表示

数据层是整个治理框架的基础,其核心任务是构建具有丰富语言学标注的Deepfake语料库。这一过程需要采集来自不同来源、涵盖多种场景的真实与伪造内容样本。每个样本都需要进行精细的多层次标注:除了基本的真伪标签外,还应包括韵律特征标注、情感标签、逻辑关系标注、叙事结构标注等。这些标注工作需要语言学专家的深度参与,以确保标注质量。在知识表示方面,需要将抽象的语言学概念转化为机器可处理的特征表示。例如,可以将“逻辑连贯性”量化为话题转移的平滑度指标,将“情感一致性”表示为多模态情感特征的协方差矩阵。

4.2模型层：多层次融合的技术路径

在模型层,需要设计创新的架构来实现语言学先验知识与深度学习模型的有机融合。特征级融合是最直接的途径,将量化后的语言学特征与传统的声音、图像特征向量进行拼接,共同输入到分类器中。多任务学习框架则更加深入,通过让模型同时学习真伪分类和语言学属性预测等多个相关任务,促使模型学习到更具泛化能力的表征。对于可解释性要求极高的应用场景,可以引入符号主义人工智能的方法,将部分语言学规则直接编码为可解释的推理逻辑。此外,知识图谱的嵌入能够为内容一致性检验提供外部知识支持,通过构建包含人物背景、历史言论、事件事实的知识图谱,实现对生成内容的事实性验证。

4.3治理层：人机协同的闭环系统

治理层关注如何将技术能力转化为实际治理效能,其核心是构建一个人机协同的闭环系统。在这个系统中,AI系统承担初步筛查和分析的任务,不仅输出真伪判断,还生成详细的检测报告,明确指出可能存在的语言学异常特征。人类专家则负责对高风险案例进行最终研判,并不断优化系统规则。系统还需要建立完善的反馈机制,将专家的研判结果和新的语言学发现持续反馈至数据层和模型层,形成持续改进的循环。此外,治理层还需要考虑与其他治理手段的协同,如区块链存证、数字水印等技术,形成多维度的治理合力。

5 挑战与展望

5.1技术实现层面的挑战

在技术实现层面,语言学特征的量化与标准化是首要难题。如何将“逻辑连贯性”这类抽象概念转化为稳定、可计算的特征,需要语言学与计算机科学的深度合作。多模态信息的高精度对齐也对算法提出了极高要求,特别是在处理不同来源、不同质量的媒体内容时。计算效率是另一个重要考量,引入复杂的语言学分析可能会显著增加系统的计算开销,这在实际部署中需要精心优化。此外,不同语言和文化背景下的语言学特征存在显著差异,这就要求治理系统必须具备足够的文化敏感性,不能简单地将基于一种语言文化构建的模型直接迁移到其他语境中。

5.2未来研究方向展望

展望未来,有几个充满潜力的研究方向值得关注。首先是面向低资源语言的治理框架研究,当前的技术资源过多集中于英语等主流语言,亟需加强对小语种和方言的支持。其次是需要探索更细粒度的语言学特征,如对隐喻、反讽等高级语言现象的识别与理解。跨文化视角的引入也至关重要,需要研究不同文化背景下语言使用规律的差异对检测效果的影响。最后,构建动态演化的“语言学先验知识库”是一个值得探索的方向,通过持续学习机制使系统能够适应语言使用的时代变迁。

6 结论

Deepfake技术带来的舆论风险治理是一个复杂系统工程,单纯依赖数据驱动的技术路径已无法有效应对日益严峻的挑战。本文系统论证了将语言学先验知识体系化融入Deepfake内容治理框架的必要性与可行性,提出了一个涵盖“数据层-模型层-治理层”的完整治理框架。这一框架的创新之处在于将语言学的深层认知与人工智能的技术能力有机结合,为提升治理体系的鲁棒性、可解释性和可持续性提供了新的思路。未来的研究应当致力于推动技术、语言学、伦理学与社会学等领域的更深度融合,通过跨学科协作共同构筑一道抵御深度伪造风险的坚固防线。同时,需要建立包括技术标准、法律法规、行业自律在内的综合治理体系,才能有效应对生成式人工智能技术发展带来的各种挑战,维护数字时代的信任基础。

[项目信息]

本项目由国家级大学生创新创业训练计划项目资助,项目名称：广西职业师范学院2025年大学生创新创业训练计划项目《生成式人工智能Deepfake技术的网络舆论风险防控与治理研究》,项目级别：国家级,项目类别：一般项目,项目编号：202514684005。

[参考文献]

[1]刘桢宇.数智时代人工智能深度伪造的风险分析与规制研究[J].智能物联技术,2025,57(05):6-16.

[2]刘伟东,马晓飞,刘硕.深度伪造技术洞察及风险治理[J].科技智囊,2025,(08):29-36.

[3]汪琦.深度伪造的国家安全风险与融贯治理路径[J].中国科技论坛,2025,(07):105-115.

[4]刘晓龙,刘欢,赵耀,等.AIGC伪造内容被动检测与主动防御技术综述[J].中国科学:信息科学,2025,55(09):2250-2288.

作者简介：

李文静(2004--),女,汉族,广西百色人,本科,单位：广西职业师范学院,研究方向：商务英语。

潘瑶婷(2004--),女,壮族,广西贵港人,本科,单位：广西职业师范学院,研究方向：人工智能。

唐钰婷(2003--),女,汉族,广西容县人,本科,单位：广西职业师范学院,研究方向：人工智能。

韦彩虹(2002--),女,壮族,广西都安瑶族自治县人,本科,广西职业师范学院,研究方向：人工智能。

*通讯作者：

林倩如(1997--),女,汉族,广西南宁人,研究生,单位：广西职业师范学院,研究方向：外国语言学、人工智能。