Journalism and Communications 新闻传播科学, 2026, 14(11), 44-52
Published Online November 2026 in Hans. https://www.hanspub.org/journal/jc
https://doi.org/10.12677/jc.2026.1411285

Hans汉斯

生成式人工智能标签与新闻信任悖论：信源权威性的调节作用

王　悦

北京印刷学院新闻传播学院，北京

收稿日期：2026年10月2日；录用日期：2026年10月31日；发布日期：2026年11月10日

摘　要

生成式人工智能正深度嵌入新闻生产与传播的全链条，AI内容标识制度随之落地。然而，标识能否修复公众信任，仍是一个有待检验的实证问题。本研究基于预期违背理论，采用2 (内容来源标签：AI生成vs.人类撰写) × 2 (信源权威性：权威媒体vs.自媒体)被试间实验设计，系统考察了内容来源标签与信源权威性对公众新闻信任的交互效应及其心理机制。研究发现：AI标签对新闻可信度存在显著的负面主效应，但这一效应高度依赖于信源权威性，当权威媒体标注AI生成时，受众感知可信度大幅下降；当自媒体标注AI生成时，信任折损并不显著，此即“信任悖论”的实证确认。有调节的中介分析表明，感知信任在“AI标签→信息接受意愿”之间发挥中介作用，且该中介效应仅在权威媒体情境下成立。三重交互分析进一步揭示，既有新闻信任构成第二层边界条件，高信任群体受“信任悖论”的冲击尤为剧烈。本研究为“信任悖论”提供了机制层面的实证解释，拓展了预期违背理论在AI传播情境中的应用边界，并为AI内容标识制度的优化与权威媒体的智能化转型提供了实证依据。

关键词

AI生成内容，新闻信任，信源权威性，预期违背，实验研究

Artificial Intelligence Labels and the News Trust Paradox: The Moderating Role of Source Authority

Yue Wang

School of Journalism and Communication, Beijing Institute of Graphic Communication, Beijing

Received: October 2, 2026; accepted: October 31, 2026; published: November 10, 2026

文章引用: 王悦. 生成式人工智能标签与新闻信任悖论: 信源权威性的调节作用[J]. 新闻传播科学, 2026, 14(11): 44-52. DOI: 10.12677/jc.2026.1411285

Abstract

Generative artificial intelligence (AI) is deeply embedded across the entire chain of news production and dissemination, and AI content labeling systems have been implemented accordingly. Whether labeling can restore public trust, however, remains an empirical question. Drawing on Expectancy Violations Theory, this study uses a 2 (content source label: AI-generated vs. human-written) × 2 (source authority: authoritative media vs. self-media) between-subjects experimental design to systematically examine the interaction effect of content source labels and source authority on public news trust and its psychological mechanism. The study finds that AI labels have a significant negative main effect on news credibility, but this effect is highly contingent on source authority: when authoritative media label content as AI-generated, perceived credibility drops sharply; when self-media do so, the trust penalty is not significant. This empirically confirms the "trust paradox." Moderated mediation analysis shows that perceived trust mediates the path from AI labels to willingness to accept information, and this mediation effect holds only in the authoritative media condition. Three-way interaction analysis further reveals that prior news trust constitutes a second-layer boundary condition, with high-trust groups being particularly severely affected by the "trust paradox." This study provides a mechanism-level empirical explanation of the "trust paradox," extends the applicability boundary of Expectancy Violations Theory in AI communication contexts, and offers empirical evidence for optimizing AI content labeling systems and the intelligent transformation of authoritative media.

Keywords

AI-Generated Content, News Trust, Source Authority, Expectancy Violations, Experimental Study

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

生成式人工智能正以前所未有的速度渗透新闻生产与传播的全链条。从选题策划、信息采集、内容制作到编辑审校，AI 技术已从辅助工具升级为贯穿新闻生产流程的驱动力量。截至 2025 年 6 月，我国生成式人工智能用户规模达 5.15 亿人，较 2024 年 12 月增长 2.66 亿人；生成式人工智能普及率达 36.5%，较 2024 年 12 月提升 18.8 个百分点。[1]。AI 技术在提升新闻生产效率的同时，也在深刻改变着新闻的生产逻辑与传播生态。

在技术层面，AI 已从辅助工具升级为贯穿新闻的采编发全流程的工具。主流媒体、商业媒体平台已普遍将 AI 工具应用于快讯播报、数据新闻、行业资讯、民生简讯等写作场景，机器写稿、AI 改稿、智能配图、个性化内容生成成为新闻媒体的新常态。新闻真实性是新闻业最根本的职业信条与安身立命之本，也是公众信任能够被建立、被维系的逻辑起点。当新闻的生产者从“人”延展到“机器”，公众的新闻信任也会随之发生改变，信任是新闻业与公众之间关系的核心纽带。从经典传播学视角看，新闻信任构成公众采纳信息、形成判断、参与公共讨论的认知前提。数字传播环境的到来，使新闻信任的结构面临根本性挑战。传统新闻信任建立在“职业权威”和“机构声誉”的基础之上新闻机构通过长期的专业实

[1]https://www.cnnic.net.cn/n4/2025/1020/c326-11389.html

践积累公信力，公众则依据对机构的既有认知形成信任判断。然而，人工智能技术的介入，正在松动这一传统信任模式的根基。当新闻生产不再专属人类记者，当“作者”身份变得模糊甚至不可见，公众用以判断新闻可信度的认知线索面临重构。

技术渗透的广度与信任缺失的深度之间的巨大张力，构成了当下新闻业面临的核心困境。一方面，AI 技术被寄予厚望在提升效率、降低成本、拓展报道边界上；另一方面，AI 生成内容的可靠性、透明度与可信度问题持续引发关注，公众对 AI 新闻的信任赤字日益凸显。

本研究的理论价值在于将预期违背理论引入 AI 新闻信任研究，为“信任悖论”提供机制层面的实证解释；同时将信源权威性从“主效应变量”重新定位为“调节变量”，深化对 AI 新闻信任形成条件的理解。实践价值在于为权威媒体在智能化转型中审慎使用 AI 技术提供决策参考。

2. 文献综述

围绕 AI 生成新闻的可信度问题，已有实证研究呈现出一幅复杂乃至对立的图景。相当数量的研究发现，AI 标签会对新闻可信度产生负面效应。Kim 等人对 31 项研究、41 个效应量进行的元分析表明，AI 生成标签对信源可信度和信息可信度均存在“虽小但统计显著”的惩罚效应。Nanz 等人在德国开展的预注册情境实验进一步发现，受众对 AI 生成的政治新闻信任度显著低于人类记者撰写的新闻，且这种信任折损在政治议题中尤为突出。高芳芳等在国内情境下的在线实验同样证实，当新闻视频被标示为 AI 自动生成时，被试感知到的新闻可信度显著低于被标示为记者采编时[1]。魏晨曦通过内容分析与问卷调查的混合研究也发现，受众对传统新闻的整体可信度评价显著高于 AIGC 新闻，传统新闻在专业性、公正性和报道深度上的优势是这一差异的主要来源[2]。庞华和张欣彤则进一步揭示，AI 来源标签显著提升了用户的偏见感知并降低了感知可信度，即便在高度拟人化的语言风格下，这一负面效应依然存在[3]。

然而，另一条研究线索却得出了不同的结论。一项基于网络随机对照实验的研究发现，AIGC 标签对感知准确性、信息可信度和分享意愿的主效应均不显著——标签的存在并未显著改变用户的信任判断。Gallegos 等人在美国情境下的研究同样发现，将信息标注为“AI 生成”或“人类撰写”并不会显著改变其说服效果。

上述看似矛盾的结论提示我们，AI 标签对新闻信任的影响并非一个“有无”的简单问题，其效应方向与强度可能受到多重因素的调节。单纯检验主效应，既难以解释研究结论的分歧，也无法揭示信任形成的深层机制。研究者开始将目光转向可能的边界条件：新闻主题、媒体类型、语言风格、受众的 AI 接受度与媒介使用习惯等变量被相继引入分析框架。

在诸多边界条件中，信源权威性尤为值得关注。杨娟等的实验发现了一个极具理论张力的“信任悖论”：新闻的生成形式(AI vs.人类)与信源权威性之间存在显著的交互效应——当权威媒体(如新华社)标注 AI 生成时，公众的信任率显著低于人工撰写时，且决策反应时间显著更长；而在非权威信源条件下，AI 生成与人工撰写之间则无显著差异。换言之，权威性的加持非但未能缓冲 AI 标签的负面效应，反而可能放大了信任损失。

上述文献为本文奠定了重要基础，但也留下了三个有待填补的缺口。其一，机制检验的缺失。杨娟等虽然用“预期违背”来解释“信任悖论”现象，但未对这一解释进行实证中介检验——权威性是否通过增强受众的负向预期违背，进而放大了 AI 标签的信任惩罚效应？这一问题仍有待回答。其二，个体差异的忽视。既有研究多聚焦于刺激层面的变量(标签、信源、语言风格)，但受众的个体差异同样可能影响信任判断。魏晨曦已发现年龄和媒介使用习惯显著调节 AIGC 新闻的可信度评价，Toff 和 Simon 的研究也暗示 AI 标签的负面效应可能主要集中在原本对新闻信任度较高的群体中，这提示“既有新闻信任”可能构成一个重要的个体层面调节变量，但尚无研究将其置于“信源权威性 × AI 标签”的交互框架中考

察。其三，行为后果的延伸不足。多数研究止步于信任态度本身，而对于信任判断如何进一步影响信息接受意愿、分享意愿等行为后果关注有限。那宇翔等虽在健康新闻领域检验了信息接受意愿，但尚未将这一路径置于“权威性 × AI 标签”的交互情境中[4]。

3. 研究设计

3.1. 实验设计

本研究采用 2 (内容来源标签：AI 生成 vs.人类撰写) × 2 (信源权威性：权威媒体 vs.自媒体)被试间实验设计(Between-Subjects Design)。被试将被随机分配到四个实验组别之一，阅读一篇内容完全相同、仅在来源标注与信源身份上存在差异的新闻材料。

选择被试间设计而非被试内设计，主要基于以下考量：其一，避免同一被试接触多种实验条件可能引发的顺序效应(order effects)与需求特征(demand characteristics)；其二，模拟真实的新闻消费场景，在自然阅读情境中，受众通常只看到新闻的单一版本，而非同时接触多个版本进行比较；其三，降低被试因反复接触类似材料而产生的疲劳效应与实验敏感度。

3.2. 变量测量与操作化

3.2.1. 自变量的操纵

(一) 内容来源标签

在“AI 生成”条件下，新闻正文首段以加粗字体标注：“本文由人工智能根据公开信息自动生成”；在“人类撰写”条件下，标注：“本文由本报记者/特约撰稿人撰写”。两类标注在页面位置、字体大小、颜色上保持一致，仅文字内容不同。

(二) 信源权威性

通过新闻发布平台的身份标识进行操纵。在“权威媒体”条件下，在新闻页面顶端显著位置展示“新华社”的品牌标识与名称，并配以“权威发布”字样；在“自媒体”条件下，展示“某热点资讯”的自媒体品牌标识与名称。两类信源标识在页面位置和视觉呈现格式上保持一致，以排除视觉显著性差异对实验结果的干扰。

3.2.2. 中介变量的测量

借鉴 Lankton 和 McKnight 的技术信任信念模型以及那宇翔等的研究，采用 7 点李克特量表(1 = 非常不同意，7 = 非常同意)进行测量。量表包含 4 个题项：① “我认为这篇新闻的内容是可靠的”；② “我认为这篇新闻的内容是准确的”；③ “我认为这篇新闻的内容是客观的”；④ “我认为这篇新闻的内容是值得信赖的”。该量表在既有研究中已被验证具有良好的内部一致性(Cronbach’s $\alpha > 0.80$)。在本研究样本中，该量表的 Cronbach’s α 为 0.87。

3.2.3. 因变量的测量

采用 7 点李克特量表测量被试对新闻信息的接受程度与行为倾向。量表包含 3 个题项：① “我愿意接受这篇新闻中的信息”；② “我愿意根据这篇新闻的信息做出判断”；③ “我愿意将这篇新闻分享给他人”。在本研究样本中，该量表的 Cronbach’s α 为 0.84。

3.2.4. 调节变量的测量

在实验刺激呈现之前，测量被试对主流媒体的总体信任基线。采用 7 点李克特量表(1 = 非常不同意，7 = 非常同意)，包含 3 个题项：① “总的来说，我相信主流媒体发布的新闻”；② “我相信新闻媒体会如实报道事实”；③ “我对新闻报道的整体质量有信心”。该量表得分将作为个体层面的调节变量，

纳入后续的调节效应分析。在本研究样本中，该量表的 Cronbach's α 为 0.81。

4. 数据分析

通过双因素方差分析、有调节的中介模型和三重交互模型，系统检验了四项研究假设。数据分析结果为理解 AI 标签对新闻信任的影响机制提供了较为完整的实证图景。

第一，AI 标签对新闻可信度存在显著的负面主效应，但信源权威性是关键的边界条件。双因素方差分析首先确认了内容来源标签的主效应，AI 生成新闻的感知可信度显著低于人类撰写新闻，这一发现与既有元分析的结论一致，再次验证了 AI 标签本身确实会对新闻可信度产生负面冲击。然而，主效应的显著性并不能完整揭示 AI 标签信任效应的复杂性。进一步的交互效应分析表明，内容来源标签与信源权威性之间存在显著的交互作用，且这一交互模式呈现出不对称的特征，当新闻来自权威媒体时，AI 标签导致感知可信度大幅下降；当新闻来自自媒体时，AI 生成新闻与人类撰写新闻之间无显著差异。这一模式表明，AI 标签的信任惩罚效应不是普适性的，而是高度情境化的，它仅在权威信源情境下显著存在，在自媒体情境下则几乎消失。这一“信任悖论”的实证确认，构成了本研究最核心的经验发现：权威性的加持非但未能缓冲 AI 标签的负面效应，反而放大了信任损失。这一结论的适用范围限于本实验所采用的政策类新闻材料与在线样本，不宜直接推广至所有新闻类型与所有受众群体；其理论意涵在于揭示信源情境对标签效应的调节，而非否定标签制度在一般意义上的透明化价值。

第二，感知信任在“AI 标签与信息接受意愿”之间发挥中介作用，且这种中介效应仅在权威媒体情境下成立。有调节的中介模型揭示了一个清晰的机制路径。在权威媒体条件下，感知信任的中介效应显著，AI 标签通过显著降低受众对新闻内容的感知信任，进而削弱了其信息接受意愿。在自媒体条件下，这一中介路径不成立，AI 标签既未能显著改变感知信任，也未能通过感知信任影响信息接受意愿。有调节的中介指数进一步证实了信源权威性对这一中介路径的调节效应。这一发现揭示了“信任悖论”的深层心理机制：当权威媒体使用 AI 生成内容时，AI 标签首先触发了受众对新闻内容可靠性、准确性和客观性的质疑，即感知信任的侵蚀。这一心理层面的变化进一步传导至行为层面，降低了受众对该新闻的接受意愿和分享意愿。换言之，“信任悖论”的根源在于权威媒体的身份放大了 AI 标签对感知信任的侵蚀效应，而这一侵蚀最终体现为受众信息接受行为的改变。在自媒体情境下，这一机制链条断裂，受众对自媒体本身的信任预期本就不高，AI 标签未能产生额外的信任侵蚀效应。

第三，既有新闻信任构成第二层边界条件。三重交互模型的结果表明，内容来源标签、信源权威性与既有新闻信任三者之间的三重交互效应显著。简单斜率分析进一步揭示了一个具有理论张力的差异化模式：对于高既有新闻信任的受众，权威媒体与 AI 标签的交互效应极为显著，AI 标签导致感知可信度大幅下降；对于低既有新闻信任的受众，虽然交互效应仍然存在但效应强度大幅减弱。换言之，“信任悖论”对高信任群体的冲击尤为剧烈，对低信任群体的影响相对有限。这一发现呈现出一个深刻的悖论：信任越深，崩塌越重。对权威媒体抱有更高信任期待的受众，在面对 AI 技术介入时，其预期违背更为剧烈，信任折损也更为彻底。这一结果将“信任悖论”的适用边界从刺激层面(信源特征)拓展到了个体层面(受众特征)，揭示了 AI 标签信任效应的差异化分布，它并非对所有受众均等存在，而是高度依赖于受众既有的新闻信任基线。

5. 讨论

本研究通过一项 2 × 2 被试间实验，系统检验了内容来源标签与信源权威性对公众新闻信任的交互效应，并进一步探究了感知信任的中介作用与既有新闻信任的调节作用。四项研究假设均获得实证支持。以下围绕核心发现、理论贡献、实践启示与研究局限展开讨论。

(一) 核心发现：信任悖论的实证确认与机制解释

本研究的首要发现是：AI 标签对新闻可信度的负面效应并非普遍存在，而是高度依赖于信源的权威性。当新闻来自权威媒体(新华社)时，AI 标签导致感知可信度显著下降；而当新闻来自自媒体时，AI 标签对感知可信度无显著影响。这一发现构成了对杨娟“信任悖论”的直接复制与扩展，不仅验证了交互效应的存在，还将其延伸至信息接受意愿这一行为层面。

个体在互动中对他人行为持有特定的预期，当实际行为偏离这一预期时，会产生认知唤醒和态度调整；偏离的方向(正向或负向)与程度决定了态度变化的方向与强度。将这一逻辑迁移至 AI 新闻信任场景：受众对“权威媒体”与“人类记者”之间建立了牢固的认知联结，权威媒体的公信力建立在专业记者长期的事实核查、深度调查与伦理自律之上。因此，当一篇新闻同时呈现“新华社”的品牌标识和“本文由人工智能生成”的标签时，受众的既有预期(“权威媒体的新闻应由人类记者生产”)与实际情况(“AI 生成”)之间形成了强烈的负向偏离。这一偏离触发了认知冲突，受众需要调用更多的认知资源来重新评估信息的可信度，这正是“决策反应时显著延长”的认知机制。

相比之下，在自媒体情境中，受众本身对平台的信任预期较低，AI 标签造成的“预期偏离”幅度较小，甚至可能被理解为“至少比来路不明的自媒体内容更客观”因此未产生显著的信任折损。

本研究的中介分析进一步验证了这一机制：在权威媒体条件下，感知信任完全中介了“AI 标签→信息接受意愿”的关系；而在自媒体条件下，这一中介路径不成立。换言之，当新华社标注“AI 生成”时，受众首先感知到的是“这篇新闻不那么可靠/准确/客观了”。这一感知信任的下降，直接导致了信息接受意愿的削弱。AI 标签的负面效应，正是通过侵蚀感知信任这一心理路径传导至行为层面的。

需要强调的是，上述解释并非排斥其他理论机制的可能性。至少还存在两条值得认真对待的竞争性解释路径。其一，启发式–系统化加工模型(HSM)提示，AI 标签可能同时激活两条加工路径：对于权威媒体，受众更可能启动系统化加工，对“AI 生成”所隐含的准确性风险进行深度审视，从而放大负面判断；对于自媒体，受众则更可能依赖“自媒体本就不太可信”的启发式线索，标签的边际信息价值有限。其二，归因理论提示，受众对 AI 介入的“动机归因”可能构成另一条机制：当权威媒体使用 AI 时，受众可能将此举归因为“压缩成本”“追求效率”等组织动机，进而产生“权威媒体在偷懒”的负面推断；当自媒体使用 AI 时，同样的行为则可能被归因为“资源有限”而获得更多谅解。这意味着，“信任悖论”的心理机制可能并非单一的感知信任路径，而是“预期违背–感知努力–动机归因”多机制并行的结果。本研究受测量工具所限，未能对这些竞争性机制进行直接检验，这是后续研究需要重点推进的方向。

(二) 边界条件：既有新闻信任的放大效应

本研究的另一重要发现是：“信任悖论”并非对所有受众均等成立，其效应强度受到个体既有新闻信任水平的调节。三重交互分析表明，越是信任新闻媒体的受众，当看到权威媒体使用 AI 生成内容时，其信任受损越严重；而对新闻媒体本就缺乏信任的受众，这一效应大幅减弱。

这一发现揭示了一个深刻的悖论：信任越深，崩塌越重。对权威媒体抱有更高信任期待的受众，在面对 AI 技术介入时，其“预期违背”更为剧烈。这也解释了为什么 AI 标签的负面效应“主要集中在原本对新闻信任度较高的人群中”，高度信任本身构成了脆弱性的来源。

从理论角度看，这一发现将“信任悖论”的边界条件从刺激层面(信源特征)拓展到了个体层面(受众特征)，构建了一个更加完整的条件性模型：AI 标签的信任惩罚效应，不仅取决于“谁贴了标签”(权威媒体 vs.自媒体)，还取决于“谁在看标签”(高信任 vs.低信任受众)。两者的交互共同决定了信任折损的最终幅度。

(三) 理论贡献

本研究的理论贡献主要体现在以下两个方面：

第一，拓展了预期违背理论在 AI 传播情境中的应用边界。已有研究初步验证了负向预期违背在健康新闻情境中的中介作用，但尚未将其置于“信源权威性 ×AI 标签”的交互框架中。本研究将预期违背理论与信源权威性这一结构性变量相结合，考察了权威性如何通过增强预期违背进而放大信任损失，从而拓展了该理论在 AI 新闻信任研究中的解释力。

第二，深化了对双重加工理论中“启发式警觉”机制的理解。庞华和张欣彤提出的“认知遮蔽效应”表明，AI 标签作为一种元来源线索，能够以高效、低耗的方式主导用户的判断框架(标签遮蔽了拟人化)。本研究的发现与之形成了互补：标签的认知主导力并非铁板一块——“新华社”这一信源标识同样具有高诊断性，它与 AI 标签形成了“双线索竞争”，并最终表现为“权威媒体的 AI 标签比自媒体的 AI 标签造成更大的信任损失”。这揭示了 AI 新闻信任判断中更为复杂的线索竞争图景。

第三，为启发式－系统化加工模型与归因理论在 AI 新闻信任研究中的整合提供了可检验的分析入口。本研究的结果模式提示，未来研究至少可以在两个方向上推进理论整合：一是在 HSM 框架下区分“标签驱动的启发式警觉”与“信源驱动的系统化审视”，考察二者在不同信源条件下的相对权重；二是在归因理论框架下引入“感知努力”(受众对新闻生产投入程度的感知)与“动机归因”(受众对媒体使用 AI 之组织动机的推断)作为并行中介或调节变量。具体而言，可在后续实验中增设“你认为这篇新闻的生产投入了多少人力与时间”“你认为该媒体使用 AI 主要是为了提升效率还是降低成本”等测量题项，并检验其与感知信任的并列中介效应，或检验其是否调节“信源权威性 ×AI 标签”的交互效应。这将有助于把“信任悖论”从单一机制解释推进为多机制竞争性检验，从而构建一个更全面、更具理论张力的解释模型。

(四) 实践启示

1) 对 AI 内容标识制度的启示

制度逻辑预设了“透明即可信”，只要让公众“知道”内容由 AI 生成，就能触发审慎判断。但本研究的发现提示：标识的效果高度依赖于信源身份。对于权威媒体而言，“AI 生成”标签非但未能触发审慎的理性审视(如制度设计所期待的)，反而可能因其对受众预期的强烈违背而加剧信任流失。这并不意味着标识制度本身无效，而是表明标识的认知效应远比制度设计者想象的复杂，它不是一个简单的“标签→识别→审慎”的线性链条，而是一个受多重因素调节的认知过程。本研究的出发点正是为了提升透明度机制的有效性——问题不在于“是否应披露”，而在于“如何披露才能在不同信源情境与受众群体中产生预期的审慎效果”。就此而言，本研究的政策意涵是“透明度的精细化”，而非“透明度的退场”。在制度优化层面，本研究的发现提示，“精细化披露”不应仅被理解为披露形式的丰富化，而应被视为对“预期违背”这一认知机制的针对性回应。据此，后续研究可从以下三个层次展开系统探讨。

第一，分级标签的认知逻辑与效果边界。本研究的核心机制表明，信任折损的根源在于 AI 标签与受众对权威媒体的规范性预期之间形成了强烈偏离。由此推论，“一刀切”式的“本文由 AI 生成”标签，实际上将“AI 仅参与配图”“AI 辅助数据整理”“AI 独立完成全文”等参与程度截然不同的生产形态压缩为同一认知信号，反而可能放大预期违背的幅度。分级标签的意义正在于通过披露 AI 参与的具体环节与程度，使受众的预期调整更为精准——当受众得知 AI 仅承担辅助性工作而人类记者仍主导事实核查与价值判断时，其感知信任的受损程度可能显著降低。但这一推论仍需审慎对待：分级本身可能带来新的认知负担，受众是否具备理解不同分级含义的能力、分级标签是否会被平台选择性使用以规避严格披露，均是实践中不可忽视的限度。后续研究可通过焦点小组访谈与问卷实验相结合的方式，比较“完全 AI 生成”“AI 辅助写作”“AI 仅参与非核心环节”等分级标签在缓冲预期违背上的相对效力，并检验受众的 AI 素养是否调节分级标签的效果。

第二，交互式解释的效能与限度。静态文字标签提供的是“来源告知”，而交互式解释(如用户点击

标签后可查看 AI 参与的具体环节、人工核查流程与责任主体)提供的是“过程可见”。从预期违背理论的视角看，后者可能通过两条路径发挥作用：其一，以更充分的过程信息帮助受众重建对新闻生产可控性的感知，从而降低预期偏离的强度；其二，通过展示人工核查与责任归属，向受众传递“人类仍在把关”的信号，缓解其对权威媒体“放弃专业责任”的担忧。然而，交互式解释的效能高度依赖于受众的主动点击意愿——在碎片化的新闻消费场景中，多数用户可能根本不会触发交互，这使得其实际覆盖面和干预效果存在显著上限。此外，过度细化的技术性解释也可能超出普通受众的理解能力，甚至引发新的不信任。因此，后续研究应重点检验交互式解释的“触发率–理解度–信任效应”链条，而非仅假设其必然有效。

第三，披露方式的受众差异化效应。本研究的三重交互结果提示，任何精细化披露方案都不可能是“一刀切”的普适解。高既有新闻信任群体对 AI 介入更为敏感，其预期违背更为剧烈，因而可能更需要过程性、解释性的披露信息来缓冲冲击；而低信任群体对标签本身的敏感度有限，过度披露反而可能被视为“公关话术”而适得其反。这意味着，制度设计需要在“充分披露”与“避免信息过载”之间寻求情境化的平衡，而非追求单一的最优披露模板。后续研究可依托大规模问卷调查，考察不同披露方式在高、低信任群体中的差异化效果，为分类施策提供经验依据。

需要承认的是，上述讨论仍属理论推演层面，其具体效果有待专门的实证检验。但恰恰是这些“如何有效落地”的问题，构成了从“发现信任悖论”走向“治理信任悖论”的关键环节，值得作为后续研究的专门议题。

2) 对权威媒体 AI 策略的启示

本研究的发现为杨娟等的警示提供了进一步的证据支持：“权威媒体盲目引入 AI 技术可能对其公信力造成损害，使其在智能化转型过程中面临信任危机。”权威媒体的核心资产是公众长期积累的信任，这一资产既珍贵又脆弱。当 AI 技术被引入新闻生产时，技术赋能与信任维护之间可能存在张力。建议权威媒体采取渐进式、透明化的人机协同策略：在 AI 应用初期，明确披露 AI 参与的具体环节(如“本文由 AI 辅助数据整理，记者独立撰写”)，而非简单标注“AI 生成”；在涉及重大公共利益或价值判断的报道中，审慎评估 AI 介入对受众信任的潜在影响。

3) 对受众 AI 素养建设的启示

研究发现，高新闻信任群体对 AI 标签更为敏感，这一群体通常是社会中的信息敏感群体和公共讨论的积极参与者。当其信任因 AI 标签受损时，不仅影响个体层面的信息接受，还可能对更广泛的信息生态产生连锁效应。因此，面向公众的 AI 标识素养教育应强化“来源意识”，让受众理解 AI 标识的含义、AI 在新闻生产中的实际作用边界，以及如何对 AI 生成内容保持适度的批判性审视。

(五) 研究局限与未来方向

本研究存在以下局限，有待未来研究加以改进。

第一，实验情境的外部效度。本研究采用模拟新闻阅读任务，虽严格控制了内容变量，但真实媒体环境中用户的注意力分散、多信源交叉验证以及既有立场偏见等因素可能调节效应强度。未来研究可采用日记法或追踪调查，在自然阅读情境中检验“信任悖论”的生态效度。

第二，样本的代表性。本研究的样本以 Credamo 见数平台用户为主，偏年轻化、高教育水平，其数字素养与技术接受度可能高于一般公众，可能高估或低估标签效应的强度。未来研究应拓展至中老年群体、低媒介素养人群及农村地区受众，以检验结论的普适性。

第三，新闻类型的局限性。本研究仅使用了一篇政策类新闻作为实验材料，结论能否推广至其他新闻类型(如突发新闻、调查报道、软新闻)仍有待检验。已有研究表明新闻主题可能调节受众对 AI 新闻的感知(蒋忠波等，2022；Liu & Wei，2019)，未来研究可在不同主题和风险等级的新闻中检验“信任悖论”

的稳健性。

第四，因果推断的单向性。本研究聚焦于 AI 标签对受众信任的单向影响，但在实际的人机互动中，受众对 AI 的态度也可能反向影响其对标签的注意与解读。未来研究可采用纵向设计或深度访谈，探究 AI 新闻信任形成的动态过程。

第五，竞争性机制的测量缺失。本研究未直接测量“感知努力”“动机归因”等可能构成竞争性解释的变量，也未系统比较启发式加工与系统化加工在不同信源条件下的相对权重。未来研究可在实验或问卷中增设相应测量题项，构建“预期违背–感知努力–动机归因–感知信任”的多重中介模型，或检验这些变量对“信源权威性 × AI 标签”交互效应的调节作用。此外，未来研究还可通过焦点小组访谈与问卷调查相结合的方式，专门考察分级标签、交互式解释等精细化披露方式的受众接受度与实际效果，从而为 AI 内容标识制度的优化提供更具操作性的证据。

6. 结语

生成式人工智能正在深刻重塑新闻业的生产逻辑与信任基础。本研究通过一项 2 × 2 被试间实验，实证检验了内容来源标签与信源权威性对公众新闻信任的交互效应。研究发现：AI 标签的信任惩罚效应并非普遍存在，而是高度依赖于信源身份，当权威媒体标注 AI 生成时，信任折损最大；当自媒体标注 AI 生成时，则无显著影响。这一“信任悖论”的背后机制在于：权威媒体的高公信力建构了受众对“人类记者生产”的强烈预期，AI 标签的出现造成了剧烈的“预期违背”，进而侵蚀感知信任，削弱信息接受意愿。更重要的是，这一效应在原本对新闻高度信任的群体中更为剧烈，“信任越深，崩塌越重”。

本研究的核心政策意涵在于：AI 内容标识制度不能停留于“打标”这一技术动作，而应纳入对信源特征、受众差异和认知机制的精细化考量。对于权威媒体而言，技术赋能必须以维护公信力为前提，在 AI 时代的新闻实践中，透明不仅意味着“告知来源”，更意味着“管理预期”。

当然，本研究也存在样本代表性、实验情境外部效度等方面的局限，未来研究可通过多群体、多情境的拓展来检验结论的普适性。随着生成式 AI 技术的持续迭代和公众认知的不断演变，AI 新闻信任问题将是一个需要持续追踪、动态回应的学术命题。同时需要重申，本研究的旨趣并非反对 AI 透明度原则，而是主张透明度机制应走向“情境敏感”与“受众敏感”的精细化设计；其结论的适用范围也应限定于本研究所考察的政策类新闻、在线样本与特定信源情境，后续研究需通过多群体、多情境、多机制的拓展加以检验。

参考文献

[1] 高芳芳, 林心婕, 赵睿晗. 受众视角下 AI 新闻生产和播报对新闻可信度的影响研究[J]. 全球传媒学刊, 2025, 12(3): 153-170.

[2] 魏晨曦, 刘亚楠. AIGC 与传统新闻可信度比较研究[J]. 国际公关, 2026(8): 137-139.

[3] 庞华, 张欣彤. 超越拟人化: 来源标签在 AI 新闻中的信任建构与认知遮蔽效应[J/OL]. 全球传媒学刊, 1-19. https://link.cnki.net/urlid/10.1270.G2.20260817.1609.007, 2026-08-24.

[4] 那宇翔, 刘映萱, 赖凯声. 公众能接受 AI 健康新闻吗? 基于信任中介与负向预期违背调节的实验证据[J]. 心理科学, 2025, 48(4): 972-984.