国际信息环境小组

应对生成式人工智能虚假信息

科学证据荟萃分析的结果

决策者摘要 2026 年 2 月

DOI 编号：10.61452/QXAF2136

A.Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M.Given, S. Lewandowsky, E. M.Navarro-Lopez, P. N.Howard

应对生成式人工智能虚假信息

科学证据荟萃分析结果

决策者摘要2026 年2 月

引用方式：
International Panel on the Information Environment [A.Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M.Given, S. Lewandowsky, E. M.Navarro-López, P. N.Howard (eds.)], “Responding to Generative AI Misinformation:Results from a Meta-Analysis of Scientific Evidence,” Zurich, Switzerland:IPIE, 2026.Summary for Policymakers, SFP2026.2, doi: 10.61452/CYLJ2059.

概要

如今，借助各类易于获取的工具，生成式人工智能 (GenAI) 能够快速生成大量具有误导性的文本、图像、音频和视频。这一问题值得决策者重视，因为虚假或误导性内容可针对特定受众定制产出。此类内容传播迅速，且能模仿真实的交流方式，令公众难以辨别真伪。

本决策者摘要提炼了 IPIE 综合报告 (SR2026.2) 的主要结论，该报告探讨了 GenAI 产生的虚假信息带来的影响，以及最有可能减轻其影响的措施。

评估结论来自一项大规模实验性科学证据荟萃分析，分析数据来源于 2018 至 2025 年间发表的 24 篇同行评审文献，包含 60 组随机对照试验效应量估计值，研究对象共计 33,801 名。

该报告得出了四项主要结论：

1. 目前，文本类 GenAI 虚假信息带来的说服性诱导风险高于视觉类虚假信息。
2. 目前所依据的证据基础并未覆盖世界上的大多数地区；研究主要集中在英语国家和高收入国家，导致我们对相关问题的认知存在显著空白。
3. 最能持续发挥效用的干预措施，是提前向用户推送纠偏信息，使其能够自主判别内容的准确性与可信度。
4. 内容标注在规范统一时方能奏效。内容标注整体上会降低人们对内容的感知信任度，但其具体影响程度因各企业生成与应用标签的方式不同而差异显著。

由此得出的政策启示十分明确。决策者应优先治理文本类 GenAI 虚假信息，并将预提示信息和纠偏信息作为一项根本策略加以支持。内容标注机制需审慎设计，并经充分测试后方可实施。同时，我们必须拓展独立研究范围，突破当前仅聚焦英语国家与高收入地区的局限。

研究人员访问获取平台和模型数据至关重要，既能帮助公众更清晰地认识 GenAI 虚假信息，也有助于检验哪些防护措施在实践中切实有效。

引言

生成式人工智能 (GenAI) 系统能够生成文本、图像、音频和视频，这些内容有时甚至让人难以区分其与人类创作内容的差别。AI 生成内容现已广泛传播，频繁出现在全球各类选举活动中，其发布方往往身份不明或带有恶意；这些内容正持续影响公共卫生、地区冲突、市场等领域的公众认知，并产生显著现实影响 [1]。这些系统包含生成文本的大语言模型、图像与音频合成工具以及视频生成平台。它们还能够产出具有误导性或危害性的内容，例如假新闻、谣言、宣传材料及深度伪造音视频。本文将上述各类现象统称为“虚假信息”。

本决策者摘要重点介绍了一项荟萃分析的主要研究结论，该分析基于实验证据，探究公众接触 GenAI 生成的虚假信息后受到的影响 [2]。此外，本文还评估了两种已获得充分实验证据支撑、可用于政策研讨的应对手段：纠偏信息和内容标注。报告重点采用随机对照试验开展研究，原因在于这类研究能够在不同研究间形成更可靠的对比，且能够更系统地估算平均效应。

相关证据表明，信息环境正处于快速演变之中。报告发现，公众对文本类虚假信息与视觉类虚假信息的反应正呈现分化趋势；另外，现有研究基础在地域覆盖范围与研究方法层面仍存在局限。同时，报告明确了两类可降低 GenAI 虚假信息感知可信度的可行应对方案，但二者的效果稳定性存在差距。

为了建立证据基础，本综述检索了 Scopus 和 Web of Science 这两个主要文献数据库，采纳了专家建议，合并了检索结果，剔除了重复文献，最终对 6,952 篇文献进行初筛。经资格筛选和全文编码后，有 87 篇文献符合详细审查标准。随后，依照 Cochrane 推荐的方法，进行额外偏倚检验、筛选及方案审查，最终选定 24 篇高度相关的研究纳入荟萃分析。主要《综合报告》采用三层次随机效应荟萃分析方法，并使用《修订版随机试验偏倚风险评估工具》评估偏倚风险。有关完整方法的详细信息，请参阅《综合报告》全文 [2]。

结果 1：具有误导性的 GENAI 文本是首要威胁

关键发现

目前，文本类 GenAI 虚假信息带来的说服性诱导风险高于视觉类虚假信息。

报告发现，公众对文本类和视觉类生成式 AI 内容的可信度感知存在显著差异。多项近期研究表明，由 GenAI 生成的虚假文本，往往会被受众判定为比真实信息更准确、更可信、更具说服力。与此同时，关于生成式 AI 生成的虚假视觉信息（如深度伪造）的研究显示，这类内容普遍被认为比真实信息的可信度更低。2021 年之后开展的视觉类研究汇总估计表明，受众对 AI 伪造视觉内容的感知可信度存在中等程度下降。

这并不意味着视觉类虚假信息是安全的或无关紧要。深度伪造技术仍然存在风险，报告指出，随着视觉系统不断改进，公众的反应可能会再次发生转变。但现有证据表明，文本类虚假信息更需政策层面予以紧迫关注。这一点意义重大，因为以文本生成为核心的 AI 系统成本低廉、可扩展性强、易于个性化定制，并且正日益融入搜索、消息传递和对话界面之中。报告还指出，对话式 AI 生成的内容可能比静态 AI 生成的信息具有更强的说服力，这表明当前的估计可能偏保守。

整体研究结论足以传递一条清晰的政策信号：当前最为紧迫且最具说服力的风险，源自虚假或误导性文本，而不仅是视觉深度伪造内容。

结果 2：全球研究证据存在空白

报告明确指出，纳入审查的文献大多以英语国家和高收入国家为研究对象，包括奥地利、德国、荷兰、韩国、英国及美国等。现有证据仅限于英文全文，另有五篇经筛选的其他语言文献未能达到纳入标准。由此可见，当前关于 GenAI 虚假信息的认知仅集中在少数国家、语言类型与媒介环境。

这是一项突出的局限。这表明，决策者、监管机构、研究人员与各类企业，对于 GenAI 虚假信息在全球多数地区的传播运作机制，认知仍然十分有限。报告

关键发现

目前的证据基础未覆盖全球绝大多数国家和语种。

明确指出，除少数已就此主题发表研究成果的国家与语种外，全球在此领域存在巨大的认知空白。对于多语言社会、跨境监管，以及构建适合不同语言、文化与政治体系的有效防护机制而言，这一信息缺口所带来的影响不容小觑。

此外，由于技术发展日新月异，现有证据也受到一定限制。许多实验仅针对 GPT-2 这类老旧模型或是 FaceSwap 等早期工具开展测试。报告警示，科研进度往往滞后于当下的技术发展水平。由此导致部分政策制定的依据，来自性能远不及当前公开使用产品的旧版系统。因此，开展覆盖更广应用场景的独立研究，是确保监管和治理举措与影响公共言论的前沿技术保持同步的切实必要之举。

结果 3：预防性纠偏措施具有积极作用

预防性纠偏信息包括简短科普警示、提醒或提示，以及在用户接触可能具有误导性的内容前，提前为用户建立辨别意识的干预方式。例如，一些干预措施简要讲解深度伪造技术如何实现逼真的视频篡改，或提醒用户 GenAI 系统可能会产生错误。这些方法旨在提高用户辨别内容是否可信的能力。

关键发现

预防性纠偏信息是最能持续发挥效用的干预手段。

在 2020 年后开展的研究中，综合估计结果表明，预防性纠偏信息会降低人们对 GenAI 虚假信息准确性、可信度、可接受度的感知，效应量处于小幅至中等水平。纳入综述的各项研究结论较为统一：不同研究之间的结果差异较小，且在本证据子集中，在不同受试人群、实验设计下均观测到上述干预效果。

这些研究结果表明，尽管目前可获得的大部分证据基于对早期 GenAI 系统的研究，但这类干预措施在短期内大概率仍能发挥效用。

《综合综述》还强调了一项重要注意事项：仅当要求人们判别内容的准确性与可信度时，预防性纠偏信息才能显现正向作用。当研究仅关注识别情况时，证据的一致性较差。此前关于深度伪造的研究发现，

警示信息有时反而会加剧人们对真实视频和经篡改视频一并产生不信任。这一点与虚假信息研究及政策层面普遍存在的顾虑相吻合：虚假叙事和宣传往往旨在让个人对所有公共信息丧失信任。因此，各类干预措施的设计应当以提升公众区分可靠信息与误导性信息的能力为目标。

预防性纠偏信息是经过最多试验验证的措施之一，可以在线上相对快速地部署，尤其是在社交媒体平台上。示例包括：提示 GenAI 可能产生错误的标注说明、讲解深度伪造技术如何制作逼真篡改视频的科普内容，以及介绍 AI 幻觉等现象的简易宣教材料。根据所属学科领域的不同，此类措施也被称为建议、认知免疫、事前告知、预先铺垫或警示标签。

结果 4：统一规范的内容标注能够发挥作用

关键发现

内容标注仍具备良好应用前景，但前提是标注清晰、标准统一。

内容标注是指在信息上附加简短的视觉、文本或多模态标签。这些标签用于提示相关内容可能由 GenAI 生成、经过篡改或具有误导性。在汇总分析中，标签与误导性 GenAI 信息在受众心中的感知可信度降低这两者之间存在虽小但具有统计显著性的关联。该结果虽然具有积极意义，但并非在所有场景下都有效。报告指出，标注手段在不同研究中的效果差异程度，远高于预防性纠偏信息。

由于证据有限，各类手段产生的效果差异极大，部分研究甚至未观测到任何作用。在某些情况下，内容标注可能会降低可信度，但另一些情况下则无此作用。本评估着重指出了可能存在的调节变量，包括标签设计、文字措辞、呈现方式、内容类型、实验环境以及标签来源。换言之，标签的设计和实施方式，其重要性不亚于是否使用标签。

关于内容标注的证据基础同样有限。这批研究样本规模相对较小，且参与对象几乎全部来自美国，这导致我们难以判断标签在不同语言、文化与法律体系下能否发挥同等作用。因此，正确结论并非内容标注毫无成效。相反，内容标注可以起到作用，但前提是企业审慎设计标签形式，并在多元场景下开展测试，而非想当然地认定该手段在所有环境中都有效。

结论

本决策者摘要指出风险态势正在发生转变。 文本类 GenAI 生成的虚假信息产生的说服诱导威胁高于视觉类 GenAI 内容。但这并不意味着深度伪造及其他合成媒体不存在危害。这一结论表明，决策者不应因视觉伪造手段备受关注，就忽视虚假 AI 生成文本日益增强的说服诱导力。

本报告重申 IPIE 此前评估 [3]、[4] 的结果，并提出了两项切实可行的应对措施。其一，预防性纠偏信息在现有证据基础中表现最为稳定、研究支撑最为充分，尤其在内容接触前提供并有助于提升个体评判能力时效果最佳。其二，内容标注同样具备一定作用，但其效果不稳定，因场景条件而异。标签须经审慎设计，并以明确、统一的方式使用，方能发挥最大效用。

最后一点关乎研究证据本身。当前研究范围仍未能涵盖全球大多数地区。这一现状使得决策者、平台以及通用 AI 模型研发者难以充分掌握相关风险在不同语言、文化与媒介体系中的传播方式。因此，开展独立研究、持续整合证据以及放开研究者调取平台和模型数据的权限至关重要。如若缺少上述举措，在快速演变的信息环境下，公共政策将始终只能被动应对、存在疏漏，且视角过于狭隘。

参考资料

[1] International Panel on the Information Environment [I.Trauthig, P. N.Howard, S. Valenzuela (eds.)], “The Role of Generative AI Use in 2024 Elections Worldwide,” Zurich, Switzerland:IPIE, 2025.Technical Paper, TP2025.2, doi:10.61452/HZUE9853.

[2] International Panel on the Information Environment [A.Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M.Given, S. Lewandowsky, E. M.Navarro-López, P. N.Howard (eds.)], “Effects of Misinformation Produced with Generative AI and its Countermeasures:A Meta-Analysis of Experimental Scientific Evidence,” Zurich, Switzerland:IPIE, 2026.Synthesis Report, SR2026.2, doi:10.61452/UGTR3022.

[3] International Panel on the Information Environment, “Countermeasures for Mitigating Digital Misinformation:A Systematic Review,” IPIE, Zurich, Switzerland, SR2023.1, July 2023. [Online].Available: https://www.ipie.info/research/sr2023-1

[4] International Panel on the Information Environment, “Platform Responses to Misinformation:A Meta-Analysis of Data,” IPIE, Zurich, Switzerland, SR2023.2, July 2023. [Online].Available: https://www.ipie.info/research/sr2023-2

鸣谢

贡献者

起草作者：Aliaksandr Herasimenka（咨询科学家，英国）、Sebastián Valenzuela（IPIE 首席科学官、科学与方法委员会主席，智利）、Shelley Boulianne（IPIE 科学与方法委员会委员，加拿大）、Frank Esser（IPIE 科学与方法委员会委员，瑞士）、Lisa M. Given（IPIE 科学与方法委员会委员，加拿大/澳大利亚）、Stephan Lewandowsky（IPIE 科学与方法委员会委员，澳大利亚/英国）、Eva M. Navarro-López（IPIE 科学与方法委员会委员，西班牙/英国/墨西哥）、Philip Howard（IPIE 总裁兼首席执行官，加拿大/英国）。研究助理：Anna George 和 Xianlingchen Wang。独立一般审查：George Georgarakis 和 Mathias Harrer。设计：Domenico Di Donna。审稿：Beverley Sykes。也衷心感谢 IPIE 秘书处的大力支持：Lola Gimferrer、Jessica Gold、Wiktoria Schulz、Donna Seymour、Anna Staender 及 Alex Young。

资助方

国际信息环境专家小组 (IPIE) 衷心感谢各资助方的支持。如需查看资助合作伙伴的完整名单，请访问 www.ipie.info。本报告中表达的任何意见、发现、结论或建议均为 IPIE 的意见，并不一定反映资助者的观点。

利益声明

IPIE 的报告由组成重点科学小组和贡献者团队的全球附属研究机构和咨询科学家网络编写和审查。所有贡献者和审查人都已完成利益声明，并在适当工作阶段经 IPIE 审查。

引用规范

IPIE 决策者摘要是面向广泛读者、对研究现状进行高度概括的文件。IPIE 综合报告利用科学的荟萃分析技术、系统评价和其他工具来收集证据、概括知识和建立科学共识，专为专家读者编写。IPIE 技术论文聚焦特定方法论相关问题或针对具体监管议题开展政策分析。所有报告均可在 IPIE 网站 (www.IPIE.info) 获取。

本文件引用格式：

International Panel on the Information Environment [A.Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M.Given, S. Lewandowsky, E. M.Navarro-López, P. N.Howard (eds.)], “Responding to Generative AI Misinformation:Results from a Scientific Meta-Analysis,” Zurich, Switzerland:IPIE, 2026.Summary for Policymakers, SFP2026.2, doi: 10.61452/CYLJ2059.

版权信息

关于 IPIE

国际信息环境小组（International Panel on the Information Environment，简称 IPIE）是一个独立的全球性科学组织，致力于提供有关世界信息环境威胁的最具可操作性的科学知识。IPIE 总部设于瑞士，其使命是通过组织、评估和提升研究，为决策者、行业和民间社会提供关于全球信息环境的独立科学评估，以改善全球信息环境为广泛目标。来自世界各地的数百名研究人员为 IPIE 的报告做出了贡献。

如需更多信息，请联系国际信息环境小组 (IPIE)：secretariat@IPIE.info。地址：Seefeldstrasse 123, P.O.Box, 8034 Zurich, Switzerland。

国际信息环境小组

地址：Seefeldstrasse 123,
P.O. Box 8034 Zurich
Switzerland