当AI事实检查出错时：人工智能标识有效性的比较

闫文捷 谭心瑶

[本文提要]面对公众可能受到人工智能误导的风险，为AI生成内容添加标识成为当下广泛采用的预警方式。人工智能应用于事实核查，在多大程度上可以影响人们对新闻真实性的判断和转发行动？为智能核查添加标识又在多大程度上能够起到警示作用，进而减少人们对AI结论的过度信任、实现更准确的新闻判断？本研究尝试通过一项四因素混合设计在线实验（N = 450）对这些问题做出探讨。结果显示，人们对新闻准确度的评价、转发意愿以及真伪识别的准确率均受到人工智能事实核查结论的显著影响。视觉显著、内容明确的AI标识在智能核查判定新闻内容不实的情况下，倾向于进一步强化公众在新闻评价和传播行动中更趋于保守的选择，即便核查判定有误，这一效果仍然存在。本文基于以上研究结果，对人工智能应用于事实核查领域时，为智能核查内容添加标识所面临的挑战展开了讨论。

[关键词]虚假信息 人工智能 事实核查 AI标识
DOI:10.16057/j.cnki.31-1171/g2.2026.04.009

引 言

在虚假内容频发的数字媒介环境下，事实核查作为应对虚假信息的重要手段，日渐被各类媒体平台与公共机构所采纳。已有研究表明，持续接触事实核查能够有效提升公众对信息真伪的识别能力，并矫正其由于接触虚假信息而形成的错误信念（Bowles et al.，2025）。事实核查虽然能够起到一定的纠偏效果，但是由新闻生产者和机构完成的专业核查较为耗时，并且要求核查人员具备必须的职业技能。这些都使得专业事实核查的发展面临规模化难题，难以应对以几何级数倍增的网络虚假信息。

为了提升事实核查的规模和效率，近年来学界与业界开始探索将人工智能技术与事实核查流程相结合的新路径，推进事实核查系统的自动化发展。理想的自动化事实核查工具可实时检测到新出现的事实性宣称，并通过查阅已核验声言的数据库、分析信源可

[作者简介]闫文捷系北京师范大学新闻传播学院教授，谭心瑶系北京师范大学新闻传播学院硕士研究生。本文为国家社科基金一般项目“社交平台虚假信息治理模式比较研究”（项目编号：21BXW065）阶段性成果。

信度等方式，对目标宣称的准确度做出评级（Hassan et al.，2015）。随着技术的发展，当前的人工智能模型已有能力在检验信息真实性的同时提供有关信息内容与社交线索的解释说明，具有实时检测虚假信息并予以纠正的潜力（Gong et al.，2024）。现实中，X等海外社交平台已探索接入基于大语言模型（Large Language Model，LLM）的人工智能机器人，允许用户直接调用AI服务验证平台内信息的真实性，微博的Ai智搜也尝试在综合平台内多个来源信息和相关新闻报道的基础上，为用户提供泛核查类解释文本。有研究者观察到，面对突发性危机事件，那些善于运用人工智能技术识别与纠正谣言的官方机构与社交媒体平台有可能更为及时准确地做出回应，进而减轻公众因此而受到的负面影响（Habas & Abu Alasal，2025）。

在提升核查效率与可及性的同时，利用人工智能进行事实核查也伴随着一些潜在的风险。现有的人工智能大语言模型因受到技术条件的限制，尚不能保证其输出的信息结论准确无误。不仅如此，AI在词汇、语法等方面的规范表现反而可能抬高识别与核实AI生成信息是否存在偏误的门槛（杨奇光，张宇，2025）。不难想象，当普通用户在缺乏专业知识背景和必要引导的情况下使用人工智能自主开展事实核查时，对于人工智能生成信息的盲目信任很可能造成事与愿违的结果，不但无法达成辨伪识真的目的，反而对自身的认知与判断产生误导。面对这样的现实风险，为AI生成的内容添加标识日渐在国际上形成普遍共识，并且成为人工智能立法和监管领域优先级最高的事项之一。比如，联合国大会于2024年3月发布《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》决议草案，鼓励开发和部署有效、可获取、适应性强、具有国际互操作性的人工智能内容认证和来源识别机制。各国人工智能立法亦相继对AI生成合成内容标识进行规定。包括欧盟《人工智能法案》、美国《加州人工智能透明度法案》和英国《人工智能（监管）法案》等在内的法律提案，均对人工智能水印和标识义务做出了明确要求。作为一种警示标识，AI标识不仅提示用户内容由AI自主生成，也建议用户对内容进行自主甄别与复核，以此提示用户AI生成错误信息的风险。按照法规政策的要求，OpenAI、Anthropic、xAI等国际主流大语言模型服务商所提供的交互式页面内均包括相应的警示标识。中国国家互联网信息办公室、工业和信息化部、公安部和国家广播电视总局联合颁布的《人工智能生成合成内容标识办法》已于2025年9月1日起正式施行，同样要求人工智能生成合成与内容传播的服务提供者在生成合成内容或交互场景页面内添加显式标识。

然而，目前通行的标识管理办法并未对AI标识的呈现形式和内容进行具体规定。对于医疗健康信息等涉及用户生命安全的高风险信息，DeepSeek等大语言模型会在输出内容的末尾添加醒目的警示标识，提示用户仔细甄别，但在大多数文本内容场景下，更多的大语言模型服务商通常设置同一种标识呈现方式，即在用户交互页面底部以暗色字符对“内容由AI生成，用户需自主甄别”做出声明。而在标识内容方面，目前的大语言模型

服务商倾向于采取类似以上表述的概要式标注，在页面内对人工智能生成内容的潜在谬误进行一般性提示，而并未指出AI出错可能造成的具体风险，也未对如何规避风险做出进一步说明。仅Anthropic为其交互页面中的标识添加了超链接，用户可以通过点击标识访问更加详细的说明页面，从而进一步了解AI错误的可能来源与对AI信息进行核查的手段（Anthropic，2025）。

那么，在大语言模型等人工智能技术被不断深入地应用于事实核查领域之时，它在多大程度上可能发挥协助公众识别虚假信息的实际作用？为了防范由于人工智能事实核查自身出现错误而在基于虚假信息形成的误信误判之外对使用者造成双重误导，以何种方式、在怎样的呈现形式与提示内容的组合条件之下为AI生成的核查内容添加标识便成为研究者与实践者需要直面的问题。面对这些新问题，已有研究刚刚起步，尚未提供明确的经验证据。本文希望通过一项基于中国网民样本的调查实验，围绕这些问题做出探索与回应。

一、基于人工智能的事实核查及其效果

人工智能大语言模型（LLM）作为信息生产和传播的新型参与者，为人工智能在事实核查场景中的应用带来新的可能。这类模型能够处理自然语言输入，并根据需求快速生成个性化的回应，完成跨领域的任务，因此成为普通公众求证信息真实性时一个可资利用的潜在工具。面向ChatGPT用户的调查显示，人们使用LLM获取信息的同时，通常对其提供的内容报以普遍的信任（Sun et al.，2024）。具体到事实核查领域，LLM提供的事实核查不仅能够有效改变人们基于原始新闻形成的错误观念，还因其表现获得了比搜索引擎或专业人士等传统信源更高的公众信任与依赖（Si et al.，2023）。事实核查对新闻真实性的判定作为一种可信度指标进一步推动或遏制着人们的新闻分享行动（Clayton et al.，2020）。据此，本文提出第一组研究假设：

H1：当人工智能提供的事实核查判定新闻为真时，人们阅读事实核查后对新闻的（a）事实准确度评价和（b）分享意愿较之于阅读前显著提升。

H2：当人工智能提供的事实核查判定新闻为假时，人们阅读事实核查后对新闻的（a）事实准确度评价和（b）分享意愿较之于阅读前显著降低。

随着人工智能LLM在各类任务场景中得到广泛应用，其局限性与潜在风险也日益受到关注。研究者观察到LLM在关键信息提取、对话生成、数值推理与图像处理等不同任务情境中均存在可能生成无意义或误导性信息的问题，这一现象被称为AI幻觉（hallucinations）。AI幻觉同样可能出现在事实核查领域。与专业核查不同，人工智能驱动的事实核查可能提供错误的信息与结论（Quelle & Bovet，2024）。最新的实证研究结果表明，GPT-4o、Claude 3.5 Sonnet和Qwen2.5-72B等人工智能LLM为核查结论提供的理

由中均可能存在事实性错误（Lin et al., 2025）。换言之，在当前的技术条件下，人工智能提供的核查论据和结论准确与否尚存较大的不确定性。考虑到人工智能大规模生成信息的能力，由不确定性带来的潜在风险也将随其使用规模一同被成倍放大，因而这种不确定性构成了人工智能事实核查投入实践时不可忽视的局限。

目前，人们在辨别LLM生成的错误信息方面能力有限。幻觉信息通常具备与真实信息相当的流畅性，加之LLM采用对话形式，向其用户呈现风格上看似自信的、经过深思熟虑的文本，这些都令人们难以从形式上直观地辨别AI生成的信息真实与否（Garry et al., 2024）。特别是当LLM提供了充分的解释性信息时，后者可能成为启发式线索，影响用户对于AI智能水平的感知，进而增加用户对AI生成信息的信任；即使用户并不能真正理解解释性信息如何推导出AI提供的结论，这一影响仍然存在（Ehsan & Riedl, 2024）。对于事实核查任务而言亦同样如此，无论AI给出的结论是否准确，用户都可能不加鉴别地接受AI所提供的核查结论（Pareek et al., 2024），从而改变个人信念。据此，我们推断人们对于新闻真实性的判断准确与否直接受制于AI核查结论的准确程度。具体来说，本文提出第二组研究假设：

H3a：当人工智能提供的事实核查结论正确时，人们阅读事实核查后判断新闻真伪的准确率随之提升；

H3b：当人工智能提供的事实核查结论错误时，人们阅读事实核查后判断新闻真伪的准确率随之降低。

二、警示标识及其在人工智能事实核查中的作用

可想而知，当人工智能提供错误的核查结论时，人们对信息真伪的判断会因此受到误导，甚至可能由于AI未能正确识别而与他人分享虚假信息，造成错误内容的进一步传播与扩散（DeVerna et al., 2024）。基于这一认知，目前有关自动化事实核查的主流观点仍然强调人工智能之于事实核查的主要价值仍在于辅助专业核查，专业核查员需要批判地检验人工智能对核查结论做出的解释以及核查过程所涉及的信息来源，并在此基础上对信息真实性做出可靠决策（Quelle & Bovet, 2024）。然而，正如X等平台的实践所显示的，普通使用者已日益倾向于将人工智能LLM视作能够用来求证信息真实性的权威工具。这些分散的用户实践并不直接受制于专业事实核查机构，也难以在专业核查人员的监督之下规范地展开。这意味着，如果只从降低AI幻觉的路径出发，仅仅聚焦人工智能应用于事实核查时的技术设计和流程设置，并不足以化解智能核查实践中的真实风险。

面对普通公众可能受到人工智能提供的事实核查信息误导的风险，为人工智能生成信息添加标识成为当下广泛采用的预警方式。标识制度被认为有助于同时满足个体辨别

生成内容和信息社会优化内容质量的共同需求（张凌寒，贾斯瑶，2024）。就个人而言，由于标识中包含针对信息准确性的简要提醒，它作为一种“助推”（nudge）手段能够提升人们辨别新闻真伪的能力，也有助于降低用户对虚假信息的信任和传播意愿（Pennycook et al.，2020）。面对大语言模型生成的信息，警示标识同样有助于用户辨别其中的幻觉成分，进而减少误导性内容的影响（Nahar et al.，2024）。从社会需求来看，采取标识制度亦有其现实考量：面对人工智能生成的海量信息，传统人工审核所需的时间与人力成本过高，并且任何审核环节都将影响LLM对话的即时性。相比之下，通过嵌入标识提升用户警觉，乃至促进必要的自主甄别则提供了一种更为可行的解决方案。

根据沟通–人类信息处理（Communication–Human Information Processing，C–HIP）模型，标识在有效吸引用户注意、内容可被理解的前提下，有望增进人们对风险的理解（Wogalter，2018）。因此，标识的视觉特征和具体内容均可影响其实际效果。譬如，颜色明亮、文字较大、位置醒目的标识因其显著度（saliency）更高而能够更加有效地提升人们对风险的感知（Carbrera et al.，2017）。就标识内容而言，由于频繁接触相同提示而导致的惯性效应反而可能削弱标识的警示效果，人们更乐于将持续出现的重复提示视作“熟悉的壁纸”，而非值得关注的重要信息（Anderson et al.，2016）。研究者尤为强调提示内容的明确程度（explicitness）对人们理解力的增进作用。相比笼统的提示，那些明确指出特定风险的警示标识能够更加有效地提升人们对潜在风险的感知与理解，进而促使其在风险面前更加谨慎（Laughery et al.，1993）；而包含了进一步实践指导与建议的警示内容则有助于人们重新评估自身的行为，进而根据标识的建议，做出长期的行为改变（Miller et al.，2016）。

我们运用C–HIP模型的概念框架作为提炼本文中标识特征的论证依据，具体而言，从颜色和位置上的差异来展现标识的视觉特征，同时从标识内容是否对风险来源进行说明（例如AI幻觉）并提出具有针对性的行动建议来衡量其明确程度。如前所述，现实中针对人工智能生成合成内容的警示标识在视觉呈现和具体内容等方面的确存在着不同表现形式。随着标识日渐成为一种AI界面下普遍采用的预警方式，事实核查的研究者有必要考察其不同设计方案如何影响智能核查的使用者对于新闻真伪的判断。基于上述已有研究的论证，我们推论，在人工智能事实核查场景下，AI生成信息标识的警示效果与其自身的视觉呈现和内容有关。标识的显著度和明确性均有助于在更大程度上提升人们对于人工智能生成核查结论的警惕，进而对原始新闻的真实性做出更加审慎的判断。本文就此提出以下研究假设：

H4：相比视觉隐蔽的标识，视觉显著的标识将提升人们判断新闻真伪的准确率。

H5：相比内容笼统的标识，内容明确的标识将提升人们判断新闻真伪的准确率。

此外，本文尝试考察标识的呈现特征和内容属性对于智能核查效果的协同性影响。遵循C–HIP模型的逻辑，一方面，醒目的标识有助于唤起人们的注意力；另一方面，内

容明确的标识则通过提供具有针对性的风险描述进一步协助人们更好地理解标识所提示的具体风险，最终引起信念、态度乃至行为上的改变。也就是说，标识的呈现与内容可凭借不同方式作用于个体判断的形成，并在彼此之间产生一定的交互。考虑到既有研究中相关证据的匮乏，我们提出以下探索性问题：

RQ1：标识的视觉显著度与内容的明确程度如何共同作用于人们判断新闻真伪的准确率？

最后我们注意到，警示标识在人工智能生成错误核查结论的情况下意义尤为重大。而智能核查生成错误结论，又可分为将真实信息误判为假（也即“错分”，commission），或把虚假内容误认为真（也即“漏分”，omission）两种情形（Chanda & Banerjee，2024）。已有研究初步显示，当人工智能出现“错分”时，用户容易受其结果影响而对信息内容做出错误判断（Schemmer et al.，2022）。如果说标识作为一种辅助性决策工具，可在人们从环境中获取的信息之外提供进一步的线索补充，那么它在人工智能事实核查任务中能发挥多大的警示作用还将有赖于核查结论的正确性以及新闻自身的内容属性。为了探索三者之间的相互关联，本文最后提出一组研究问题：

RQ2：标识的显著度对人们就（a）新闻准确度评价和（b）分享意愿的影响如何随新闻真实性和人工智能核查结论的正确性发生变化？

RQ3：标识内容的明确度对人们就（a）新闻准确度评价和（b）分享意愿的影响如何随新闻真实性和人工智能核查结论的正确性发生变化？

三、研究方法

（一）样本

我们委托问卷调查平台“见数”于2025年4月15日至17日面向中国网民样本开展了一项在线调查实验。本研究按照2025年1月中国互联网信息中心发布的第55次《中国互联网络发展状况统计报告》中网民的性别与年龄构成选取配额样本。研究共回收问卷492份，其中450份为通过了注意力检测的有效答卷，有效率为91.5%，数据无缺失值。经过G*Power测算，在效应值为0.1、统计效力为0.95的情况下，该混合设计所需的最低样本量为436，本研究达到了最低样本量的要求。有效参与者中包含218名女性（48.4%）与232名男性（51.6%），年龄在18至70岁之间（M = 42.71，SD = 12.88），348名参与者具有大学本科以上学历（77.3%）。

（二）研究设计和实验程序

实验采用2（新闻真实性：真vs.假）×2（事实核查结论：肯定vs.质疑）×2（标识视觉显著度：显著vs.隐蔽）×2（标识内容明确度：明确vs.笼统）的四因子组内组间混

合设计，其中新闻真实性与核查结论为组内因素，标识的显著度与明确度为组间因素。

参与者在阅读知情同意书并同意参与研究后，首先填答一份基准调查，完成人工智能素养和信任及健康和科学素养等变量水平的测量。随后他们被随机分配至基于AI标识显著度和明确度形成的四个组别之一，进入新闻准确度判断任务。任务包含四个试次，其中每个试次都包含下述步骤：参与者阅读一段与健康话题相关的网络新闻，报告此前是否接触过这则新闻、感知到的新闻准确度与分享意愿，并判断新闻信息真实与否；随后阅读一段人工智能大语言模型提供的针对该新闻内容的事实核查信息；最后重新评估新闻的准确度与分享意愿、判断新闻信息真实与否。四条新闻在新闻真实性（真 vs. 假）与人工智能事实核查对新闻真实性的判断（肯定 vs. 质疑）之间进行平衡，即（真，肯定）、（假，质疑）、（真，质疑）、（假，肯定）四种条件下的实验材料以完全随机的顺序各出现一次。前两种条件下，人工智能提供的核查结论正确，后两种情况下，核查结论错误。四组被试所接触的新闻和事实核查内容完全一致，差别仅在于事实核查内的警示标识是否以视觉显著的方式予以呈现以及提示内容是否具体明确。最后，调查测量了参与者的社会人口学特征、思维模式以及对人工智能事实核查的使用与评价等，并就实验所提供的新闻材料与事实核查信息的真实性进行事后解释。

（三）实验素材

本文选用的四条微博新闻分别关于维生素缺乏引发季节性皮肤问题、防晒产品的健康隐患、畸形草莓的食品安全和蓝牙耳机辐射致癌等话题。新闻内容全部选自中文网络，同时接受过浸大事实查核、浙江辟谣平台和科普中国等权威机构对于内容真实性的专业验证。

四条新闻以微博截图形式呈现，长度约250字、形式一致。我们进而将新闻上传至DeepSeek，要求大语言模型“对上述新闻进行事实核查并生成简要的报告”。为了使AI提供的事实核查与专业核查在形式上尽可能保持一致，我们参考《事实核查手册》（魏星，2023），要求AI将核查的内容“重塑为一系列问题或主张，并分别寻找材料、数据、文件等外部信源来检验该说法”。报告中的每一部分都需要包括核查所针对的主张、核查结果、理据来源和可信度评估，以此获取原始的AI事实核查文本。由于实验涉及人工智能事实核查给出错误结论的情况，我们对其中两条事实核查文本进行修改，使之得到相反的结论，即错误地肯定虚假新闻或质疑真实新闻。具体而言，在原始AI事实核查的基础上，我们要求DeepSeek修改核查结论，同时保持核查的理据数量、语言风格、语篇结构和内容篇幅（约450字）不变，即在控制了文本质量与结构不变的前提下，利用AI生成一段结论相反的事实核查信息。在此过程中，大语言模型将汇集真实存在的网络虚假信息，或采用刻意误导的方式给出错误的解释。例如，对于“仅使用一次防晒产品，氧苯酮即可在2小时内达安全阈值”这一主张，正确的AI核查将指出这一主张系基于“涂抹面积达到身体表面的75%，每天涂抹4次”的使用条件，远高于防晒产品的实际日常用量，从而对主张

提出质疑；而错误的核查则忽略以上基本前提，仅向人们展示支持该主张的论据，得出误导性结论。以上事实核查的全部内容均采用DeepSeek聊天页面截图的形式予以呈现。

在以上核查内容基础上，我们将四种不同类别的警示标识添加至截图内部。在视觉显著度方面，本文参考当前DeepSeek提供标识的两种方式，分别使用“答案末尾的高亮标识”（“显著标识”组）与“页面底部的灰字标识”（“隐蔽标识”组）两类标识。在内容明确度方面，“笼统标识”组采用DeepSeek通用的标识文本，提示被试“内容由AI生成，请仔细甄别”，“明确标识”则参考其他领域的标识设计及Anthropic等头部人工智能公司在标识超链接中进行的风险说明，在通用标识文本的基础上，进一步“建议用户交叉验证信息，核查信息质量，警惕算法偏见与AI幻觉”，以此实现对虚假信息的风险来源与应对风险的实践建议做出具体明确的说明（详见附录①）。

（四）因变量的测量

新闻准确度评价。被试在阅读完每则新闻后被要求回答：“据您所知，这则消息在事实上有多准确？”（1 = 完全不准确，10 = 十分准确），并在阅读完人工智能做出的事实核查后使用相同的量表，再次评价新闻的准确度。

新闻判断的准确率。被试在每一次对新闻准确度做出主观评价后，同时被要求判断新闻真实与否（0 = 不真实，1 = 真实）。我们据此计算出被试对于新闻真实性做出准确判断的百分比，用以测量他们判断新闻真伪的准确率。

新闻分享意愿。在接触人工智能提供的事实核查前后，被试均需回答以下问题来汇报他们对原始新闻的分享意愿：“如果您在浏览微博、微信等社交平台时（再次）遇到这则消息，您在多大程度上会把它转发给他人”（1 = 完全不可能，7 = 极其可能）。

标识评估。被试完成所有的实验测试后，回答3个有关AI标识的问题。首先，他们需回顾事实核查提供的人工智能生成合成内容标识出现的位置（1 = 页面底部，2 = 答案末尾，3 = 不记得/不确定）。其次，他们分别对“有关人工智能生成合成内容的标识有多醒目”（1 = 十分不醒目，10 = 十分醒目）和“标识中的提示信息有多具体”（1 = 十分模糊，10 = 十分具体）做出回答。这些问题旨在用于对警示标识的操纵进行检验。

四、研究结果

在开始正式的研究假设检验之前，我们进行了随机化和操纵检验。首先，人们对人工智能所做决策的态度可能受到其人工智能素养（Huang & Ball，2024）、人工智能信任（Ashoori & Weisz，2019）、机器启发式信念（Sundar & Kim，2019）及反思性思维水平（Zhang et al.，2025）的影响；同时由于本研究采用的材料为与健康话题相关的新闻，故而还需要考虑科学素养的影响。随机化检验结果显示，四个条件组的被试围绕社会人口变量以及

上述个体特征等方面均不存在显著差异，随机化成功。

其次，操纵检验结果显示，在注意到提示标识的被试者中，“显著标识”组的被试更倾向于正确指出标识出现在“答案末尾”的位置，“隐蔽标识”组的被试则更倾向于指认标识出现在“页面底部”（$\chi^2(1, N = 380) = 30.64, p < .001$），同时前者对标识显著度的评分显著高于后者（$t(448) = 7.48, p < .001$）；在标识提示内容更加明确的条件下，被试感知到的标识中的提示信息也更为具体（$t(448) = 3.63, p < .001$）。综合以上检测结果，本文成功操纵了警示标识的视觉显著度及其内容的明确程度。

（一）人工智能事实核查的影响

我们首先通过配对样本t检验，比较被试在接触事实核查信息前后对新闻事实准确度的评价。当人工智能提供的事实核查判定新闻为真时，被试接触事实核查后对新闻的准确度评价（$M = 8.35, SD = 1.31$）较接触前（$M = 7.02, SD = 1.63$）显著提升，$t(449) = -19.61, p < .001$, Cohen's $d = -.93$；当人工智能提供的事实核查判定新闻为假时，被试接触事实核查后对新闻的准确度评价（$M = 4.79, SD = 2.22$）较接触前（$M = 7.06, SD = 1.76$）显著下降，$t(449) = 20.39, p < .001$, Cohen's $d = .96$。简言之，人们对新闻准确度的评价随着人工智能事实核查的结论呈现出规律的变化，H1a和H2a均得到支持。

进而不难想象，核查结论的准确度势必影响人们对新闻真实性判断的准确率。的确，当人工智能得出正确的核查结论时，被试新闻判断的准确率显著提升（$M_{后} = 82.3\%, SD = 26\%, M_{前} = 49.8\%, SD = 22\%, t(449) = -23.42, p < .001$, Cohen's $d = -1.10$）；相反，当人工智能核查出现错误时，被试判断新闻真伪的准确率则随之显著下降（$M_{后} = 20.8\%, SD = 27\%, M_{前} = 51.8\%, SD = 31\%, t(449) = 18.54, p < .001$, Cohen's $d = .87$）。H3a和H3b成立。

与新闻评价形成相互映照的是，人们对新闻的分享意愿同样在接触事实核查信息前后出现系统性差异：被试因人工智能事实核查判定新闻为真而更乐于在网上与他人分享新闻（$M_{后} = 5.57, SD = 1.23, M_{前} = 4.57, SD = 1.64, t(449) = -16.39, p < .001$, Cohen's $d = -.77$）；相反，对人工智能核查结论为假的新闻则分享热情不高（$M_{后} = 3.27, SD = 1.67, M_{前} = 4.57, SD = 1.60, t(449) = 15.31, p < .001$, Cohen's $d = .72$），H1b和H2b均得到验证。

（二）预测AI标识的警示效果

接下来，我们转入分析AI标识在人工智能应用于事实核查任务场景时，其自身特征的警示效果。我们首先通过双因素协方差检验，在控制了核查前新闻判断准确率的基础上，围绕显著度和明确度各不相同的标识条件组之间在新闻判断准确率方面最终的均值差异做出比较。结果显示，虽然标识显著度具有显著的主效应（$F(1,450) = 4.17, p < .05, \eta^2 = .01$），但有悖于H4，“显著标识”组的被试（$M = 50.3\%, SD = 19.09\%$）判断新

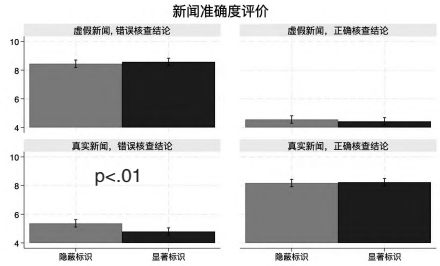

图 1　新闻真实性、核查结论正确性与标识视觉显著度对新闻准确度评价的交互影响

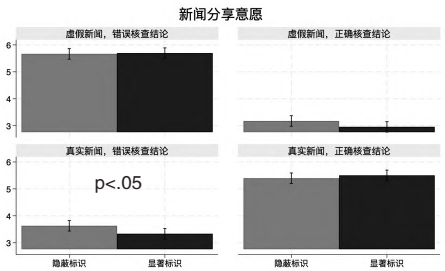

图 2　新闻真实性、核查结论正确性与标识视觉显著度对新闻分享意愿的交互影响

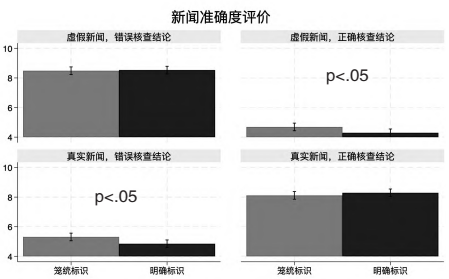

图 3　新闻真实性、核查结论正确性与标识内容明确度对新闻准确度评价的交互影响

闻真伪的准确率显著低于“隐蔽标识”组的被试（M = 52.8%，SD = 19.09%）。同时，标识内容的明确度对人们判断新闻真伪的准确率既未显示出显著的主效应（F（1,450）= .01，*ns*）（H5），也不存在和显著度之间的交互效应（F（1,450）= .46，*ns*）（RQ1）。

基于以上观察，我们进一步检验标识作用的发挥如何依赖于智能核查和新闻自身的内容属性，尤为关注标识如何在人工智能针对真假新闻生成准确度不一的核查结论时，对人们的新闻评价和分享行动产生差异化的影响。为此，我们估测了两组多层线性模型（Multilevel Linear Modeling），其中模型一用于预测新闻准确度评价，模型二用来估测新闻转发意愿。新闻层次的主要变量包括新闻自身的真实性和人工智能针对每则新闻进行事实核查后得出的结论正确与否，此外，我们控制了人们对每条新闻的熟悉程度以及在接触事实核查前对新闻准确度的初步评价或转发意愿。个体层次的主要变量为警示标识的视觉显著度和内容的明确度。最后，在主效应之外，我们在模型中纳入了新闻真实性、核查结论正确性与两项组间元素之间的交互项，并控制了与之相关的低阶交互效应，以评估标识特征与新闻真实性及人工智能核查的协同影响。

首先，如表 1 的结果显示，虽然 AI 标识的显著度和明确度对新闻准确度评价和分享意愿的影响效果均未达到统计显著的水平，但两者之间存在彼此交互（新闻准确度评价：B = .44，SE = .21，p < .05，分享意愿：B = .42，SE = .17，p < .05）：视觉显著而内容笼统的标识令人们在判断新闻真伪（B = –.36，SE = .15，p < .05）和分享新闻（B = –.31，SE = .12，p < .05）时都倾向于更加保守，在标识内容明确具体的情况下，视觉上是否显著不再对新闻评价（B = .09，SE = .15，*ns*）与分享（B = .11，SE = .12，*ns*）产生影响。其次，正如我们所推断的那样，标识特征对新闻准确度评价（围绕显著度的三阶交互：B = .90，SE = .36，p < .05；围绕明确度的三阶交互：B = 1.07，SE = .36，p < .01）和分享意愿（围绕显著度的三

表 1 新闻准确度评价的多层线性模型估测结果

变量	模型一新闻准确度评价	模型二分享意愿
固定效应截距	6.05 (.23) ***	4.12 (.15) ***
新闻层次固定效应（*N* = 1800）		
新闻真实性 （1 = 真实）	−2.82 (.22) ***	−1.95 (.16) ***
核查结论正确性 （1 = 正确）	−3.70 (.22) ***	−2.47 (.16) ***
曾听说过该新闻（1 = 是）	.14 (.10)	.18 (.08) *
新闻准确度评价（核查前）	.34 (.02) ***	——
新闻转发意愿（核查前）	——	.34 (.02) ***
个体层次固定效应（*N* = 450）		
标识视觉显著度 （1 = 显著）	−.10 (.21)	−.19 (.17)
标识内容明确度 （1 = 明确）	−.17 (.21)	−.18 (.17)
交互效应		
显著度 × 明确度	.44 (.21) *	.42 (.17) *
新闻真实性 × 核查结论正确性	6.17 (.31) ***	4.06 (.22) ***
新闻真实性 × 显著度	−.71 (.25) **	−.33 (.18)
新闻真实性 × 明确度	−.51 (.25) *	−.17 (.18)
核查结论正确性 × 显著度	−.26 (.25)	−.25 (.18)
核查结论正确性 × 明确度	−.43 (.25)	−.14 (.18)
新闻真实性 × 核查结论正确性 × 显著度	.90 (.36) *	.65 (.26) *
新闻真实性 × 核查结论正确性 × 明确度	1.07 (.36) **	.41 (.26)
随机效应		
截距方差	.34 ***	.39 ***
模型拟合度		
χ^2 (*df* = 14)	1878.62 ***	1636.51 ***
边际 R^2	.49	.45
条件 R^2	.54	.54

注：两个模型均为线性混合效果回归模型，采用限制性最大似然法（REML）估测，表格中为非标准化回归系数，括号内为标准误差。$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$。

阶交互：$B = .65$, $SE = .25$, $p < .05$）的影响均受限于人工智能核查结论的准确性，并且因新闻自身的真实性而呈现出规律性差异。RQ2 和 RQ3 得到回应。

具体而言，如图 1 和图 2 所示，视觉效果夺目的 AI 标识在人工智能核查误判真实新闻为假时，可显著降低人们对新闻准确度的评价，即增加“错分”的可能（$B = -.58$, $SE = .19$, $p < .01$），并降低其新闻转发意愿（$B = -.30$, $SE = .14$, $p < .05$）。与此同时，内容明确的标识既可在人工智能核查误判真实新闻为假时，增强“错分”的风险（$B = -.46$, $SE = .19$, $p < .05$），也可在人工智能正确识别出虚假新闻时，降低人们误信假消息的可能（$B = -.39$, $SE = .19$, $p < .05$）（见图 3）。

五、结论和讨论

在数字化转型的当下，大语言模型等人工智能技术被不断深入地应用于事实核查领域。相比由新闻记者和职业核查人员等人类行动者所主导的专业事实核查，人工智能驱

动的事实核查无疑在提升核查效率与规模等方面表现理想。但由于机器幻觉等问题的存在，公众在运用LLM等人工智能工具自主开展事实核查的过程中面临被AI误导的风险。为了增进公众对公共信息的理性认知，为人工智能生成合成内容添加标识便成为国际社会普遍公认的必要之举。本文的核心关切有两重。首先，人工智能应用于日常事实核查任务时如何影响公众的新闻判断？其次，针对人工智能核查内容的标识以何种样貌呈现有望实现警示公众免受AI误导的目标？为此本文通过一项在线调查实验，检验呈现特征与内容明确度不一的AI标识，在何种经验条件下作用于人工智能事实核查的展开，从而影响公众围绕其新闻判断和分享倾向做出改变。

研究结果表明，人们对于新闻真实性的判断受到人工智能事实核查的左右。这不仅体现在接触智能核查后的新闻准确度评价与分享意愿随着事实核查对新闻真伪的判定而规律性地提升或降低（H1、H2），还表现在智能核查结论正确与否直接决定着人们识别新闻真伪的准确度（H3）。那么面对人工智能事实核查生成错误结果，并因此催生公众误解这一可能，添加AI标识能在多大程度上起到警示用户谨慎甄别核查内容的作用？围绕这个问题，双因素协方差分析的结果显示，内容明确或视觉醒目的警示标识不足以单独提升人们在接触人工智能事实核查后识别新闻真伪的准确率，标识的显著度甚至可能对辨识率起到不利的影响（H4、H5），两者之间也未呈现出任何显著的相互作用（RQ1）。这些初步的观察从表面看似乎有别于C–HIP模型等标识制度领域相关研究的主要结论，同时它们也在提示我们，针对人工智能事实核查，希望通过添加标识而减轻使用者被AI误导的程度，进而达到警示目的的作用机制，远比我们预计的要复杂。一个可能的原因是，标识在人工智能事实核查场景中的具体效果不仅受制于核查结论是否准确，还有赖于新闻自身真实与否。

本文的数据印证了这一预期。我们观察到，视觉显著、内容明确的标识在人工智能误将真实新闻错判为假消息，即出现“错分”的情况下，可能事与愿违地加剧AI误导公众、损害真实新闻可信度及传播扩散的负面影响；明确具体的AI标识只有在人工智能正确识别出虚假新闻的条件下，才有望进一步避免公众对假消息的误信误判，发挥出积极的警示作用（RQ2、RQ3）。综合这些研究结果来看，一旦人工智能得出新闻内容不实的结论，为其添加引人注意、内容又有针对性的AI标识，都可在实际上起到强化公众警觉的作用，进而令人们在新闻评价和传播行动方面都表现得更趋谨慎保守。换言之，呼应C–HIP模型等相关研究的主张，视觉显著、内容明确的标识的确倾向于对人们产生更可见的影响，只是在人们借助人工智能核验新闻真伪的场景下，其影响效果发生的实际程度和具体方向均受限于智能核查得出的判定结论。

我们可尝试从两个方面来理解这样的研究结果。首先，公众对人工智能作为新兴传播主体的认识与接受值得关注。有学者提出，在数字媒介环境下，科学技术等非传统力量有望成为崭新的知识来源，进而构成新型认知权威，譬如大语言模型等AI应用即可被视

作非个人认知权威（non-personal epistemic authority）的代表（Bartsch et al., 2025）。具体到事实核查任务中，AI显然已具备了超越绝大多数普通用户的知识优势，无论人们对其既有的总体评价是否积极，AI都正在成为能够对使用者产生稳定说服力的事实核查来源（Chae & Tewksbury, 2024）。不难想象，为人工智能事实核查添加警示标识，虽然意在提示用户AI生成错误核查内容与结论的潜在风险，却也可能在实际上起到增进AI作为认知权威的作用。譬如，作为大语言模型的界面特征之一，醒目且内容明确的标识可能被用户视为AI工具自身专业合规的标志，强化其对AI认知权威的主观感知，故此即便在AI核查出错的情况下，也能起到提升其核查结论可信度与说服力的效果。结合本文观察到的标识特征、新闻真实性与智能核查结论之间较为复杂的三阶交互，标识效应集中出现于人工智能正确或错误地核验新闻为假这两种条件下。这一结果是否由于AI否定新闻真实性时，用户需要承担更强的认知负荷，从而促使他们更容易依赖标识特征这一便于捕捉的外在线索，进而采纳来自AI这一认知权威的决策？在缺少直接数据支持的情况下，我们只能对此作出猜想，驱动实验结果的内在机制则有待未来研究加以检验。

其次，本研究结果提示我们需正视可解释性陷阱（explainable pitfalls）的风险。根据可解释人工智能（explainable AI）的相关研究，当人们缺乏基于AI提供的解释性信息进行自主决策的能力时，这些信息可能诱发启发式推理，详尽的解释易于被视作AI具备较高性能的线索，最终强化人们对AI决策的过度信任（Ehsan & Riedl, 2024）。特别在大语言模型等可解释性较高的人工智能应用错误地将信息归入用户“关注的类别”时，对错误结论的进一步解释反而可能加深对使用者的误导（Schemmer et al., 2022）。本文中视觉显著、内容明确的警示标识，在提示人们AI核查存在风险的同时，亦可能暗示提供人工智能事实核查服务的主体对于潜在风险的认知，继而构成一种潜在的解释，即承担核查任务的AI系统具备自我监测和风险披露的能力，因而已在核查过程中对所述风险进行过一定程度的管理。我们推断，在日常的事实核查任务场景下，人们对虚假内容更加关注，也更为敏感，在普通用户缺少复核验证人工智能核查结论之能力的情况下，为智能核查结论添加醒目明确的警示标识可能在实际上成为一种可觉察的解释行为，被人们视作显示AI专业性的启发式线索，进而成为对智能核查结论的背书。以上两方面论证或许可以部分地解释本文观察到的标识呈现和内容特征在新闻真伪各异、核查结论准确性不一的情况下所呈现出的差异化效果。考虑到本文的研究目的，同时受限于现有数据，我们未对AI标识在人工智能事实核查场景下的作用机制展开细致的考察，这无疑是未来研究与实践需要进一步探究的问题。

本研究的主要局限在于实验中有关核查任务的设置与现实中人们运用AI展开事实核查的经验存在一定差距。为了控制实验中的干扰因素，被试被要求在阅读完事实核查文本后、在不借助其他外部信源的条件下，立即对原始新闻的真实性做出再次评价。这有

别于人们可能在多个信息渠道之间进行交叉验证的日常媒介使用习惯。尤其在基于大语言模型等人工智能工具展开新闻验证的场景下，AI通常会向使用者提供佐证其核查结论的外部信源。对于事实核查这一高度依赖外部信息的新闻实践来说，获取外部信源对于任务表现而言至关重要。AI标识虽能提示用户提高警觉、仔细甄别人工智能核查信息的准确性，却无法为其评估核查信息准确与否提供实际线索。创造一个更加仿真的人工智能事实核查场景对于检验现行警示标识的有效程度，以及如何对其进行进一步优化等问题的重要性无须多言。未来研究可尝试探索运用田野实验等具备较高现实性的复合研究方法做出进一步评估。与之相关，为了聚焦针对AI标识的实验操控，被试应研究者的要求卷入阅读新闻及事实核查的过程。然而在真实环境下，人们在多大程度上会主动运用AI展开自主的事实核查，以及在多大程度上直接采纳AI提供的事实核查结论，均可能受到其对相关新闻话题的关注和卷入程度等因素的影响。譬如，依循双过程模型的逻辑（Evans & Stanovich，2013），人们倾向于更加审慎地处理与评价自己更为关心的新闻及核查信息，进而有望减轻对人工智能核查结论的过度信任。虽然本文在多层线性模型中控制了被试对每则新闻话题的熟悉程度，在一定程度上侧面反映了话题本身可能对新闻评价及转发意愿的影响效果，但由于缺少对个体新闻卷入等变量的直接测量，本文无法进一步探讨在不同程度的新闻验证动机下，人们面对结论各异的AI核查以及警示标识对此作出的提示时会呈现出怎样不同的反应。这些问题均可以在未来通过研究者更具针对性的变量测量与统计控制，乃至直接的实验操控得到解答。

值得注意的是，当下的人工智能大语言模型可能由于幻觉、偏见及引用不可靠信源等多种不同原因而得出错误的事实核查结论，本文并未对错误类别进行区分。此外，为还原AI核查出现错误的现实场景，我们以指令形式要求大语言模型生成与原核查结论相反的事实核查文本，以此获得AI核查出错条件下的事实核查材料。由于生成机制上的差异，这种以人为设定方式所得到的错误核查信息在说服力方面是否可与AI自动产制的致误核查内容相比拟，结论尚不明确。近期的实证研究表明，标识对于不同程度的AI幻觉起到的警示效果有所不同（Nahar et al.，2024）。本研究则促使我们思考，标识对于不同原因和生成机制下所产生的错误的AI事实核查是否也存在类似的差异化效果？我们对其如何识别，又当如何应对？这些问题都值得研究者在未来研究中持续地追问。

最后，我们想强调的是，本文仅就人工智能大语言模型针对健康类新闻话题展开事实核查任务时，添加AI标识的有效性进行了探讨。AI用于对时政、民生等其他领域的新闻议题开展事实核查时，为其核查内容添加标识也许会呈现出有别于本文所观察到的警示模式。更进一步，在人工智能应用于事实核查或信息整合以外的任务场景时，为AI生成合成内容添加标识势必会出现更多新的难题。将人工智能融入事实核查以及更广泛的虚假信息抗击方案是否能够带来实际回馈，在很大程度上有赖于方案设计与实施的妥当

性。当前业界采用的标识设计方案大多基于警示标识领域的研究传统，大体上预设了视觉醒目、内容明确的标识对潜在风险具有更为有效的警示效果。本文结果提示我们，至少在事实核查的任务场景下，采用标识警告人们避免因过度信任人工智能而陷入认知与决策风险的实际效果要复杂得多。这意味着传统标识理论在 AI 参与内容生产与传播的情境中需要做出相应调整，采纳了 AI 技术的事实核查平台或 AI 交互页面的设计者还需发展针对人工智能生成内容自身特性的差异化标识方案，并在实践中对不同设计方案的警示效果展开持续的校验。本文结果呼唤我们正视，AI 标识的警示效果唯有在人工智能事实核查的精确度得到严格落实、误报或漏报的可能性得以在最大限度上被降低的条件下才有望显现。在虚假信息狙击战中，单凭先进的技术工具远远不够，这些工具必须可靠可信，方能为保障公众免于由错误信息所致的认知与决策风险奠定基础。■

注释

① 感兴趣的读者可通过 OSF 网站公开获取本文附录，网址为 https://osf.io/szc4n。

参考文献

杨奇光，张宇（2025）。生成式人工智能新闻生产的边界——基于 AI 新闻评论文本语态与幻觉风险的考察。《新闻记者》，（09），21-36。

魏星（2023）。事实核查手册。检索于 https://chinafactcheck.com/wp-content/themes/youju/assets/fact-check-manua-PC.pdf。

张凌寒，贾斯瑶（2024）。人工智能生成内容标识制度的逻辑更新与制度优化。《求是学刊》，51（01），112-122。

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., & Eargle, D. (2016). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems*, 33(3), 713-743.

Anthropic. (2025). Claude is providing incorrect or misleading responses. What's going on? Retrieved January 22, 2026: https://support.claude.com/en/articles/8525154-claude-is-providing-incorrect-or-misleading-responses-what-s-going-on.

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675*.

Bartsch, A., Neuberger, C., Stark, B., Karnowski, V., Maurer, M., Pentzold, C., ... & Schemer, C. (2025). Epistemic authority in the digital public sphere. An integrative conceptual framework and research agenda. *Communication Theory*, 35(1), 37-50.

Bowles, J., Croke, K., Larreguy, H., Liu, S., & Marshall, J. (2025). Sustaining exposure to fact-checks: Misinformation discernment, media consumption, and its political implications. *American Political Science Review*, 119(4), 1864-1887.

Cabrera, M., Machín, L., Arrúa, A., Antúnez, L., Curutchet, M. R., Giménez, A., & Ares, G. (2017). Nutrition warnings as front-of-pack labels: Influence of design features on healthfulness perception and attentional capture. *Public Health nutrition*, 20(18), 3360-3371.

Chae, J. H., & Tewksbury, D. (2024). Perceiving AI intervention does not compromise the persuasive effect of fact-checking. *New Media & Society*, 14614448241286881.

Chanda, S. S., & Banerjee, D. N. (2024). Omission and commission errors underlying AI failures. *AI & Society*, 39(3), 937-960.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4), 1073-1095.

Colville, S., & Ostern, N. (2024). Trust and distrust in GAI applications: the role of AI literacy and metaknowledge. *In Proceedings of the International Conference on Information Systems* (ICIS 2024). Association for Information Systems.

DeVerna, M. R., Yan, H. Y., Yang, K. C., & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50), e2322823121.

Ehsan, U., & Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns*, 5(6), 100971.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3), 223-241.

Garry, M., Henkel, L. A., & Foster, J. L. (2024). Wires crossed? On Chatbots as threats to reality monitoring. *Journal of Applied Research in Memory and Cognition*, 13(4), 485-489.

Gong, Y., Shang, L., & Wang, D. (2024). Integrating social explanations into explainable artificial intelligence (XAI) for combating misinformation: Vision and challenges. *IEEE Transactions on Computational Social Systems,* 11(5), 6705-6726.

Habas, M., & Abu Alasal, N. S. (2025). The effectiveness of AI at rumor correction during crisis: Digital media perspective. In A. Al-Marzouqi, S. Salloum, K. Shaalan, T. Gaber, & R. E. Masa'deh (Eds), *Generative AI in creative industries* (pp. 519-531). Springer Nature Switzerland.

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015, July). The quest to automate fact-checking. *In Proceedings of the 2015 computation+ journalism symposium.* The Computation+ Journalism Symposium.

Huang, K. T., & Ball, C. (2024). The influence of ai literacy on user's trust in ai in practical scenarios: a digital divide pilot study. *Proceedings of the Association for Information Science and Technology*, 61(1), 937-939.

Laughery, K. R., Vaubel, K. P., Young, S. L., Brelsford Jr, J. W., & Rowe, A. L. (1993). Explicitness of consequence information in warnings. *Safety Science,* 16(5-6), 597-613.

Lin, H., Deng, Y., Gu, Y., Zhang, W., Ma, J., Ng, S. K., & Chua, T. S. (2025). Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. *arXiv preprint arXiv*:2502.17924.

Miller, E. R., Ramsey, I. J., Baratiny, G. Y., & Olver, I. N. (2016). Message on a bottle: are alcohol warning labels about cancer appropriate? *BMC Public Health*, 16, article number 139.

Nahar, M., Seo, H., Lee, E. J., Xiong, A., & Lee, D. (2024). Fakes of varying shades: How warning affects human perception and engagement regarding LLM hallucinations. *arXiv preprint arXiv*:2404.03745.

Pareek, S., van Berkel, N., Velloso, E., & Goncalves, J. (2024). Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), 1-31.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770-780.

Quelle, D., & Bovet, A. (2024). The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence,* 7, 1341697.

Schemmer, M., K ü hl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. *arXiv preprint arXiv:* 2204.08859.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daum é Iii, H., & Boyd-Graber, J. (2024, June). Large language models help humans verify truthfulness--Except when they are convincingly wrong. *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1459-1474).

Sun, X., Ma, R., Zhao, X., Li, Z., Lindqvist, J., Ali, A. E., & Bosch, J. A. (2024). Trusting the search: unraveling human trust in health information from Google and ChatGPT. *arXiv preprint arXiv*:2403.09987.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. *In Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1-9).

Vereschak, O., Alizadeh, F., Bailly, G., & Caramiaux, B. (2024, May). Trust in AI-assisted decision making: Perspectives from those behind the system and those for whom the decision is made. *In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Wogalter, M. S. (2018). Communication-human information processing (C-HIP) model. In M. S. Wogalter (Ed.), *Forensic human factors and ergonomics* (pp. 33-49). CRC Press.

Zhang, M., Ye, J. H., & Yang, X. (2025). Watch out for errors! Factors related to ChatGPT skepticism: A cognitive perspective. *Interactive Learning Environments*, 33(6), 3985-3999.