可解释人工智能及其在自动化事实核查中的应用

● 谭心瑶　闫文捷

摘　要：随着社交媒体内部虚假信息威胁的加剧，结合人工智能的自动化事实核查成为治理虚假信息的重要手段之一。本文基于对可解释人工智能概念的梳理，对可解释人工智能应用于事实核查领域的优势与潜力展开研究发现，融合了大语言模型与证据检索工具的自动化事实核查具备良好的可解释性特征。一方面，它所提供的事实核查结论凭借充分、可供查证的证据支持而容易得到用户的信任；另一方面，它所内含的声言拆分与信源呈现的论证结构亦有助于用户自主复现事实核查流程。可解释人工智能在应用中也可能带来一定误导风险，技术与素养并重的多元策略将帮助用户更安全、高效地与人工智能协作，使自动化事实核查真正成为公众防范虚假信息、做出理性判断的重要支持。

关键词：可解释人工智能；事实核查；大语言模型；虚假信息；人工智能素养
DOI:10.15997/j.cnki.qnjz.20250703.002

一、引言

如今，社交媒体日益成为公众接触新闻信息的主要途径之一。与传统媒体不同，社交媒体能使普通用户以更低的成本参与内容生产，其去中心化特征加上转评赞等社会背书线索和算法的加持，令非权威来源的信息内容有机会迅速传播并引发广泛关注[1]，造成虚假信息的扩散。随着社交媒体影响力的不断提升，其虚假信息传播问题对公共秩序的威胁日益加剧[2]。遏制虚假内容的泛滥，已经成为新闻传播研究者与从业者面临的严峻挑战。

由专业新闻机构和从业人员开展的事实核查，通过对已发布信息和声言的准确性进行专业评估，并向公众展示评估过程与所用信源[3]，为纠正虚假内容、提升公共信息质量提供了一套备选方案。然而，开展事实核查需要一定专业技能[4]，无论在所需时间还是人力成本等方面均远远高于虚假信息的生产要求。这不仅导致事实核查只能覆盖极小部分的网络虚假内容，而且往往错失最理想的干预时机。虚假信息的影响力往往在传播初始阶段达到最大化[5]，但专业事实核查很难在这一阶段对虚假信息的扩散形成有效干预。此外，由于事实核查难以像虚假信息一般持续性地大范围传播，它对错误内容的纠偏多以短期效果为主[6]。

在以上现实条件下，如何凭借技术手段的介入实现事实核查范围和效率的提升成为应重视的问题。自动化核查，即将自然语言处理与机器学习等人工智能（AI）技术应用于事实核查任务，日益成为专业事实核查的必要环节和有益补充。目前，通过多种路径实现的自动化事实核查已能以较高准确度对信息的真实性做出判断。然而，仅给出判断结论并不足以使其赢得用户的信任——若 AI 提供的结论缺乏可理解的论证，用户便难以判断结论是否可靠，而其纠错效果也将受限。因此，如何在自动化事实核查实践中高效运用可解释人工智能，成为这一领域的核心问题之一。

二、“可解释人工智能”的概念与实践

（一）理念、视角与类别

自人工智能概念被提出起，研究者即致力增进

系统做出决策时的可解释性，即构建可解释的人工智能系统（explainable AI，XAI）。对可解释性的需求源自开发 AI 的核心目标，即 AI 应当为用户提供可信的信息和可靠的判断，并且能使用户核验 AI 的决策与建议是否值得遵循[7]。进入深度学习时代，AI 模型的复杂度大幅提升，而模型的可解释性也遭受前所未有的威胁，不仅一般用户难以理解深度学习模型如何做出决策，具有专业知识的模型设计者也难以实际把握模型的内在工作机制[8]。实现可解释的人工智能成为当前 AI 领域的一项重要课题。

人工智能的可解释性通常可以从技术和用户两个视角加以审视。技术视角侧重于模型实际的工作机制，即“模型的决策过程是否透明、可追溯”；用户视角强调的则是模型做出决策的理由是否充分，“用户是否理解、信任模型的决策过程”[9]。

基于以上两种视角，人工智能模型的可解释性可具体分为两种类型，并由相应的指标予以度量呈现。对于技术的强调要求模型内置的决策过程具有良好的透明度，即“事前可解释性”（ante-hoc explainability），能够从模型内部为用户提供可理解并有意义的解释[10]，也因此更注重输出的稳定性和完备性等系统功能性指标[11]。对于用户体验的重视则更为关注从“黑箱”外部入手分析模型的决策结果，据此在模型做出决策后构建有针对性的解释[12]，即“事后可解释性”（post-hoc explainability）。基于 AI 用户实际体验的可解释性指标包括感知透明度、可问责性与因果性等[13]。其中，因果性（causability）是衡量人工智能事后可解释性的关键指标[14]。具有高因果性的 AI 模型在做出决策的同时，解释说明决策所依循的因果路径，例如在分类任务中列举与分类相关的关键观察特征、在预测任务中提供可理解的推理等。这些解释性信息旨在协助用户理解 AI 模型输入数据与输出结果之间的逻辑关系、知晓 AI 做出决策的原因，并在此基础上决定是否采纳 AI 决策。

（二）大语言模型：黑箱中的可解释性

用户视角对于构建可解释的人工智能至关重要。用户是否接受并信任人工智能对决策的解释，通常取决于他们能否从解释中获取逻辑自洽、连贯的因果推断以及相关的信息支持，而不仅仅是人工智能的研究者们倾向于关注的 AI 系统自身的算法透明度和稳定性等[15]。早期通过可视化、特征权重等技术手段实现的事后可解释性[16]，尽管能够为用户提供与 AI 决策相关的线索，但是正确解读线索仍然需要用户具备一定的专业知识。自然语言生成技术（natural language generation，NLG）通过将非语言输入信息转化为人类自然语言形式的文本输出，令 AI 系统能够直接为用户提供易于理解的解释[17]，进一步降低了用户解读的门槛。传统的自然语言生成技术依赖专家设计和维护的规则，因此使用 NLG 实现的事后解释需要针对不同类别的 AI 系统分别制定输出规则，这限制了解释系统的通用性。相对而言，近年来取得突破性进展的大语言模型（large language model，LLM）则有能力为多种不同任务生成符合语境的解释[18]，通用性更强、维护成本更低，提供的解释风格与人类文本也更加相近[19]，因而成为实现事后可解释性的一种理想工具。

大语言模型尤其契合用户视角对可解释人工智能的要求：当前的 GPT-4o、DeepSeek-R1 等模型均具备以自然语言形式模拟推理路径，即“思维链”（chain-of-thought）的能力。用户可通过阅读模型生成的推理链条理解模型决策的内在逻辑；并且，针对具有不同知识背景与需求的用户，大语言模型亦可根据语境生成定制化解释、支持进一步追问与交互[20]，有利于用户对模型决策的深入理解，降低误解的风险[21]。

需要指出的是，单就技术而言，大语言模型依赖于规模庞大的训练数据与参数配置，内部运作机制难以被直接观察或理解，因此属于典型的“黑箱”。它能否胜任事后解释任务在技术层面上尚待考证，比如有证据表明大语言模型提供的解释可能不符合实际的内部推理过程，而只是事后生成的合理化理据[22]。但不可否认的是，以自然语言形式生成解释性文本的能力提升了大语言模型成为适用于事后解释任务之工具的可能性[23]。它在用户感知下的突出的“可解释性”令大语言模型所提供的论断被广泛接受，并因此可能被运用于自动化事实核查的任务之中。

三、可解释人工智能在自动化事实核查中的应用

（一）自动化事实核查的实现方式

自动化事实核查（automated fact-checking）旨在运用人工智能技术识别、验证及修正虚假声言[24]。根据AI模型结构与判断方式的不同，自动化事实核查大致包含四类实现方式。首先，“语义特征表示”和“辅助任务设计”基于虚假信息的关联特征，诸如标题风格[25]、情感词使用[26]及信息在社交情境中引发的用户行为[27]等做出真实度判断。其次，“内部知识推断”利用知识图谱等模型的结构化知识[28]，采用逻辑推理评估信息真伪。最后，“基于外部特征的事实核查”通过提取事实性宣称，随之检索和收集外部证据等方式，做出信息真实度的判断并为此提供完整的解释[29]。

四种核查方式的内在逻辑预示着它们在可解释性方面存在显著差异，其中基于外部证据的自动化事实核查被认为具有天然的可解释性优势[30]。这类模型在调用搜索工具展开自主证据检索的基础上，无需针对特定宣称进行训练，即可做出真实度判断[31]，同时基于对证据的整合，形成完整的因果推断链条，为用户提供对判断结果的解释说明。

值得注意的是，专业领域对AI模型能力的评估多聚焦于准确度与跨领域适用性等指标，往往忽略了用户对模型决策的接受过程。虽然一些基于文本特征的模型在预测信息真伪时表现出较高的准确度[32]，但已有研究表明，用户对AI决策的接受程度往往并不取决于决策本身的准确度，而是有赖于AI为决策提供合理、清晰、易于理解的解释与依据。例如，不同类别AI纠偏效果的差异可能恰恰源自它们可解释程度的差异[33]。当AI仅展示有关信息真实度的判断结果，而未提供任何理据时，判断本身几乎无法影响用户分享虚假信息的倾向[34]；相反，当AI为判断提供解释，尤其是在说明中使用专家意见、逻辑论证与经验证据时，用户对其做出的判断会更加信任[35]。只有当用户认为AI提供的纠偏信息比原始宣称更加可信时，才可能实现纠正错误信念并改变传播行为的目标。因此，审视人工智能技术在虚假信息治理中的角色呈现理应超越它在信息检测任务中的表现，从可解释性视角纳入用户对这一技术应用的信任、接受和采纳过程的考察和理解。

（二）可解释人工智能嵌入事实核查的实践路径

可解释性有助于增进用户对AI决策的信任，从而提高人工智能介入虚假信息核查时的实际纠偏效果。可解释人工智能可通过多种实践路径被嵌入事实核查的过程。

首先，AI系统有能力为用户提供信息真实度判断中的逻辑论证链条，以自然语言的形式完成并呈现从证据到结论的完整推理过程，从而令自动化核查提供的解释性信息更易于理解。

根据图尔敏论证模型（Toulmin model of argument），良好的论证包含三个基本要素：主张（claim）、资料（grounds）和依据（warrant）。其中，主张指需要确立的结论，资料指用于支持结论的事实基础，依据则指运用资料支持主张的合理性来源[36]。这一模型被认为能够反映现实情境中多数论证的结构，因此常被自然语言处理中的论证挖掘借鉴，用于论证文本的分析与生成[37]。同样，事实核查的论证也遵循这一结构：核查主体基于事实基础（“资料”），展开一系列逻辑论证（“依据”），最终对新闻信息的真实度做出判断（“主张”）。

在前文述及的自动化事实核查的四种实现方案中，内部知识推断和基于外部证据的事实核查均内含论证的三个基本要素。诸如此类的自动化事实核查模型一方面能够利用可验证的外部知识，为用户提供可查证的资料，完成事实举证，另一方面可在事实基础上生成包含逻辑推理在内的完整的解释文本，为用户提供清晰连贯的因果推理过程[38]。自动化核查甚至可以根据用户背景提供更加丰富的定制化论证。例如，基于GPT-4的事实核查工具可以根据用户年龄、受教育程度和政治倾向等个体特征生成不同语言风格的解释文本，而此类高度定制化的论证和解释通常被用户认为更有帮助，能够更加有效地提升用户识别虚假信息的准确率[39]。由于围绕解释文本因果逻辑的理解力因人而异，大语言模型提供的定制化论证为实现更加普遍的可解释性奠定了基础。这些都意味着此类自动化事实核查过程

与传统的专业核查具有基本一致的论证结构，符合人类判断信息真伪时遵循的论证逻辑，能够满足用户视角下可解释人工智能的基本要求。

其次，自动化事实核查提供的解释性信息不仅有望成为可被用户感知的可信度线索，增强其对事实核查结论的信任，还具备引导用户对信息真实性进行自主复核的潜能。对于一段有待检测的信息，AI 系统首先将其拆解转化为一系列独立的、可分别验证的次级宣称 [40]。例如，对于“美国食品药品监督管理局（FDA）发布的研究显示，防晒霜中的化学成分可通过皮肤吸收进入血液，血药浓度远超安全基准”这一声言，AI 模型可以将其拆分为“FDA 是否发布了研究”“研究的具体结论与数据是否符合声言陈述”“研究结论是否存在争议或补充”三项待核查的问题。已有关于“助推”的研究显示，仅仅提供网络虚假信息大量存在的简单提示即可引导人们对信息真实性进行反思，进而减少分享虚假信息的倾向 [41]。我们可以合乎逻辑地推论，在信息拆分过程中，原始宣称所忽略的隐含推理与关键细节（如实验条件、结论争议性等）被重新提出 [42]，这将有助于用户以此为切入点展开批判性思考。我们注意到，现有文献中尚未出现相关的经验证据以支持自动化事实核查有助于引导用户自主识别虚假信息这一论断，值得学者们做出深入的探索。

完成声言拆分后，AI 模型将次级宣称进一步转化为用于检索的关键词或有待核查的问题，调用搜索工具检索证据 [43]。具备联网功能的大语言模型可为用户提供可访问的来源链接，以便用户直接跳转至模型完成事实核查时参考的外部网页，例如宣称中引文的原始出处、相关机构的官方网站、在线百科和第三方发布的事实核查等。提供信息来源既为自动化事实核查结论提供了证据支持，也提升了核查结论的可验证性。

综上所述，可解释人工智能在事实核查领域展现了突出的应用潜力。融合大语言模型与证据检索工具的自动化事实核查具备良好的可解释性特征。一方面，它所提供的事实核查结论凭借充分、可供查证的证据支持而容易得到用户的信任；另一方面，它所内含的声言拆分与信源呈现的论证结构亦有助于用户自主复现事实核查流程。基于这些特性，可解释人工智能系统有望在专业核查之外，为用户提供事实核查的补充性服务。

四、可解释人工智能的风险与应对

（一）事后可解释性的局限及其伴随的风险

大语言模型因其能够生成自然语言推理链并提供决策证据而具有良好的可解释性，也因此有望在自动化事实核查中赢得更多的用户信任。这种可解释性得以实现的基础在于，模型首先得出判断结论，继而再使用自然语言赋予结论以流畅、完整、合乎逻辑的推理论证，因此属于事后可解释性的范畴。不同于事前可解释性，事后可解释性无需在自身与模型的预测性能之间进行取舍，并且更加注重满足用户在解释形式方面的需求。然而，事后解释所呈现的逻辑推理链条往往基于 AI 模型对大量论证文本的学习，而并非直接展现模型内部的推理过程 [44]。这意味着，即使 AI 做出错误判断，事后解释方法也能够赋予它看似合理、逻辑连贯的解释，掩盖原本出现错误的环节。这些貌似合理流畅的事后解释通过增强用户对于答案说服力的感知，实际上造成诱使用户接受错误结论的风险 [45]。例如，在进行伴随联网搜索的事实核查任务时，大语言模型可能检索到来源不可靠的信息并将其作为证据，但模型生成的论证文本并不会对此做出提示，最终，AI 的回应在形式上具备因果逻辑，实则建立在错误依据之上。用户在缺少对证据进行复核的情况下，很容易被 AI 基于不可靠论据得出的错误结论所误导 [46]。

这种风险可能由于 AI 提供的解释信息的因果性特征而被进一步放大。大语言模型擅长呈现关于决策的因果解释，但它实际上的因果推理能力很有限，因而时常在执行因果推理任务时出现幻觉 [47]。尽管开发可解释人工智能的目的在于协助用户更加准确地做出判断，例如在 AI 决策与自身对于信息真伪的判断意见相左时，AI 提供的解释可以协助人们更好地理解 AI 模型的决策过程 [48]，可解释人工智能的因果性特征却可能在一定程度上增加用户对 AI 决策的盲目信任 [49]，落入可解释性陷阱（Explainability Pitfalls，EPs）。在这种情况下，人们并不真正具备理

解 AI 所提供的解释性信息的能力，无法就信息的可靠性做出自主鉴别，但只是呈现解释性信息便可增加使用者对 AI 决策的信任 [50]。

可解释性所诱发的人们对 AI 决策过度信任的程度与解释性信息的呈现方式有关。当解释性信息包含具体的证据示例，而不仅仅以自然语言的形式做出解释说明时，AI 决策更容易诱发专家与非专家用户在虚假信息检测任务中的过度信任 [51]。

而对 AI 的过度信任在 AI 决策出现错误的情况下则可能带来始料未及的回火效应。具体到事实核查这一场景，回火效应表现为，事实核查信息不仅没能纠正个体基于虚假信息形成的错误认知，反而加深了他们原有的错误信念，甚至促生出新的错误观点。在检测虚假信息的任务中，AI 决策可能出现错分（commission error）和漏分（omission error）两类错误。错分是指将真实信息误判为虚假信息，漏分则是指将虚假信息误认为真实信息，较之于不提供解释性说明的 AI 工具，当可解释人工智能模型出现错分时，用户更有可能受到 AI 结论的影响而做出错误判断 [52]。由于传统的事实核查通常由专业机构与核查人员实施完成，研究者一般不考虑事实核查自身的内容与结论出现错误的情况，但当大语言模型等可解释人工智能技术被引入事实核查实践，特别当 AI 工具成为可供普通用户进行事实验证的手段时，研究者和实践者则必须把 AI 得出错误结论的情况纳入实际的考量。

（二）风险应对：技术与素养

为了减轻甚至规避将可解释人工智能运用于事实核查任务时带来的潜在风险，提供相关服务的企业、平台和组织需要在设计自动化事实核查工具时便将模型可能的失误纳入考量。有研究表明，在大语言模型的回复中嵌入警告标签，提示用户“AI 可能生成包含不准确信息的回复”可有效增加用户准确识别 AI 幻觉的可能，并且不影响真实内容的感知真实性 [53]。循此思路，由 AI 驱动的自动化事实核查亦可在其回复中嵌入提示信息，引导用户对模型的论证过程进行复核。除此之外，人工智能工具的设计者还应考虑将具备复核功能的 AI 模型嵌入自动化事实核查的整体架构之中。部分 AI 模型已能够实现对模型幻觉的检测与评估，标记出未经证实的声言并据此生成提示性信息，在多模型协作中结合此类模型，被认为有助于减轻 AI 幻觉的影响 [54]。在此基础上，若在自动化事实核查的 AI 模型架构内引入此类负责复核的模型工具，并在事实核查的解释论证中嵌入复核结果，或可有助于减轻 AI 错误结论对使用者的负面影响。

在通过技术途径改进 AI 模型、减少 AI 误判概率的同时，也有必要引导用户以批判性的方式理解和接受可解释人工智能工具提供的论证，保持对其潜在错误的警惕态度。对此，相关领域的研究者提出人工智能素养（AI literacy）概念，以此强调人类具备与人工智能相关的各个领域（例如原理、应用与伦理等）的专业知识、技能与能力之重要性 [55]。人工智能素养也常被视作个体与 AI 进行有效互动的必要前提 [56]。针对目前应用最为广泛的生成式人工智能，研究者进一步提出生成式人工智能素养（Generative AI literacy）概念，用于度量用户使用 ChatGPT 等生成式人工智能工具的能力，同时观察到，那些拥有较高生成式人工智能素养的用户在人机协作中的表现更加理想 [57]。

根据上文所展望的自动化事实核查的 AI 模型架构，AI 模型所提供的解释论证包括可供复核的证据以及建立在此基础上的逻辑推理链条。具有批判意识的事实核查使用者在接触自动化核查及其结论时，既要自觉关注证据是否可信，也需对 AI 呈现的逻辑推理之有效性等问题展开自主的评估。培养公众的 AI 素养，则不仅需要提升他们对 AI 技术原理与可供性等一般性知识的理解，还应结合具体的应用场景为使用者提供实践指引。在本文关注的自动化事实核查情境中，用户批判地接受 AI 核查的程度，不仅取决于他们对 AI 工具内在运作逻辑的认知，还有赖于他们在这一特定的任务背景下看待 AI 的方式——是把 AI 当作提供最终真相的权威，还是把它视作事实核查人机协作实践中的辅助环节。由于自动化事实核查工具为用户展示搜索证据和论证过程，以及评估信息真实性时可供参考的维度，它们因此被认为具有提升用户媒介素养的长期效益 [58]。这更加显示出引导公众以审慎、自主的

方式使用自动化事实核查工具的重要意义。

五、结语

可解释人工智能在针对虚假信息的自动化事实核查中具备突出的应用潜力。一方面，利用可解释人工智能技术开展事实核查，可以实现对海量的网络虚假信息和声言做出快速识别和判断。另一方面，基于大语言模型的自动化事实核查还可提供并呈现与声言真实性相关的完整的逻辑推理链条。本文从用户视角出发，强调可解释人工智能所提供的事实核查基于其文本特征与论证结构如何增进用户对核查结论的理解与信任，从而成为对专业事实核查实践的有益补充。

与此同时，可解释人工智能所驱动的自动化事实核查同样伴随着一定风险。可解释性陷阱揭示出AI所提供的看似流畅并且符合逻辑的解释说明可能在现实应用中诱发使用者对AI决策的过度信任。这意味着即便AI得出的结论存在错误，人们仍然倾向于信任AI提供的结论，并因此而受其误导。不难想象，当AI为错误信息做出表面合理的辩护时，人们基于错误信息而形成的错误信念有可能因此而得到强化，无形中增加了后续性纠偏工作的难度。目前的技术水平无法保证AI驱动的自动化事实核查一定能够得出准确的结论，而大语言模型所采用的事后解释方法亦不能保证它所做出的解释性说明真实地反映了模型内部的推理过程。如何从技术完善与素养提升等方面着手解决自动化事实核查造成的潜在性误导风险，是人工智能的设计者和教育者亟待解决的问题。

【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”（批准号：21BXW065）的成果，同时得到北京师范大学中央高校基本科研业务费专项资金资助（编号：1233300010）】

参考文献：

[1] Chen S, Xiao L, Kumar A. Spread of misinformation on social media: What contributes to it and how to combat it[J]. Computers in Human Behavior, 2023, 141: 107643.

[2] Muhammed T S, Mathew S K. The disaster of misinformation: A review of research in social media[J]. International Journal of Data Science and Analytics, 2022, 13(4): 271–285.

[3] Graves L. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking[J]. Communication, Culture & Critique, 2017, 10(3): 518–537.

[4] Hassan N, Adair B, Hamilton J T, et al. The quest to automate fact-checking[C]//Proceedings of the 2015 Computation+ Journalism Symposium. New York: The Computation+ Journalism Symposium, 2015.

[5] Starbird K, Dailey D, Mohamed O, et al. Engage early, correct more: How journalists participate in false rumors online during crisis events[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018: 1–12.

[6] Burel G, Farrell T, Mensio M, et al. Co-spread of misinformation and fact-checking content during the Covid-19 pandemic[C]//Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12. Springer International Publishing, 2020: 28–42.

[7] Scott A C, Clancey W J, Davis R, et al. Explanation capabilities of production-based consultation systems[J]. American Journal of Computational Linguistics, 1977: 1–50.

[8] Xu F, Uszkoreit H, Du Y, et al. Explainable AI: A brief survey on history, research areas, approaches and challenges[C]//Natural Language Processing and Chinese Computing: 8th cCF International Conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8. Springer International Publishing, 2019: 563–574.

[9] Lipton Z C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16(3): 31–57.

[10] Došilovi'c F K, Brčić M, Hlupi? N. Explainable artificial intelligence: A survey[C]//2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018: 0210–0215.

[11] Nauta M, Trienes J, Pathak S, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai[J]. ACM Computing Surveys, 2023, 55(13s): 1–42.

[12] Retzlaff C O, Angerschmid A, Saranti A, et al. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists[J]. Cognitive Systems Research, 2024, 86: 101243.

[13] Hassija V, Chamola V, Mahapatra A, et al. Interpreting black-box models: a review on explainable artificial intelligence[J]. Cognitive Computation, 2024, 16(1): 45–74.

[14] Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(4): e1312.

[15] Miller T. Explanation in artificial intelligence: Insights from the social sciences[J]. Artificial Intelligence, 2019, 267: 1–38.

[16] Band S S, Yarahmadi A, Hsu C C, et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods[J]. Informatics in Medicine Unlocked, 2023, 40: 101286.

[17] Danilevsky M, Qian K, Aharonov R, et al. A survey of the state of explainable AI for natural language processing[J]. arXiv preprint arXiv:2010.00711, 2020.

[18] Mavrepis P, Makridis G, Fatouros G, et al. XAI for all: Can large language models simplify explainable AI?[J]. arXiv preprint arXiv:2401.13110, 2024.

[19] Karanikolas N, Manga E, Samaridi N, et al. Large language models versus natural language understanding and generation[C]//Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics. 2023: 278–290.

[20] Lakkaraju H, Slack D, Chen Y, et al. Rethinking

explainability as a dialogue: A practitioner's perspective[J]. arXiv preprint arXiv:2202.01875, 2022.

[21] Cawsey A. Planning interactive explanations[J]. International Journal of Man-Machine Studies, 1993, 38(2): 169-199.

[22] Agarwal C, Tanneru S H, Lakkaraju H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models[J]. arXiv preprint arXiv:2402.04614, 2024.

[23] Liao Q V, Vaughan J W. Ai transparency in the age of llms: A human-centered research roadmap[J]. arXiv preprint arXiv:2306.01941, 2023, 10.

[24] Graves D. Understanding the promise and limits of automated fact-checking[J]. Reuters Institute for the Study of Journalism, 2018.

[25] Lee N, Li B Z, Wang S, et al. On unifying misinformation detection[J]. arXiv preprint arXiv:2104.05243, 2021.

[26] Ajao O, Bhowmik D, Zargari S. Sentiment aware fake news detection on online social networks[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2507-2511.

[27] Liu Y, Wu Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[28] Zhu B, Zhang X, Gu M, et al. Knowledge enhanced fact checking and verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3132-3143.

[29] Wang H, Shu K. Explainable claim verification via knowledge-grounded reasoning with large language models[J]. arXiv preprint arXiv:2310.05253, 2023.

[30]许旻辰, 屈丹, 司念文, 等. 社交媒体虚假信息检测技术研究综述[J/OL]. 计算机工程, 1-20[2025-05-11]. https://doi.org/10.19678/j.issn.1000-3428.0070287.

[31] Li X, Zhang Y, Malthouse E C. Large language model agent for fake news detection[J]. arXiv preprint arXiv:2405.01593, 2024.

[32] Chen M Y, Lai Y W, Lian J W. Using deep learning models to detect fake news about COVID-19[J]. ACM Transactions on Internet Technology, 2023, 23(2): 1-23.

[33] Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI[J]. International Journal of Human-computer Studies, 2021, 146: 102551.

[34] Yaqub W, Kakhidze O, Brockman M L, et al. Effects of credibility indicators on social media news sharing intent[C]//Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems. 2020: 1-14.

[35] Pareek S, van Berkel N, Velloso E, et al. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment[J]. Proceedings of the ACM on Human-Computer Interaction, 2024, 8(CSCW2): 1-31.

[36] Kneupper C W. Teaching argument: An introduction to the Toulmin model[J]. College Composition & Communication, 1978, 29(3): 237-241.

[37] Habernal I, Gurevych I. Argumentation mining in user-generated web discourse[J]. Computational Linguistics, 2017, 43(1): 125-179.

[38]何富威, 张仕斌, 卢嘉中, 等. 融合大语言模型和证据抽取的事实核查模型[J/OL]. 武汉大学学报(理学版), 1-10[2025-05-11]. https://doi.org/10.14188/j.1671-8836.2024.0067.

[39] Gabriel S, Lyu L, Siderius J, et al. MisinfoEval: Generative AI in the Era of" Alternative Facts"[J]. arXiv preprint arXiv:2410.09949, 2024.

[40] Wang H, Shu K. Explainable claim verification via knowledge-grounded reasoning with large language models[J]. arXiv preprint arXiv:2310.05253, 2023.

[41] Andl S, Akesson J. Nudging away false news: Evidence from a social norms experiment[J]. Digital Journalism, 2020, 9(1): 106-125.

[42] Chen J, Sriram A, Choi E, et al. Generating literal and implied subquestions to fact-check complex claims[J]. arXiv preprint arXiv:2205.06938, 2022.

[43] Li X, Zhang Y, Malthouse E C. Large language model agent for fake news detection[J]. arXiv preprint arXiv:2405.01593, 2024.

[44] Laugel T, Lesot M J, Marsala C, et al. The dangers of post-hoc interpretability: Unjustified counterfactual explanations[J]. arXiv preprint arXiv:1907.09294, 2019.

[45] Ajwani R, Javaji S R, Rudzicz F, et al. LLM-generated black-box explanations can be adversarially helpful[J]. arXiv preprint arXiv:2405.06800, 2024.

[46] Hannigan T R, McCarthy I P, Spicer A. Beware of botshit: How to manage the epistemic risks of generative chatbots[J]. Business Horizons, 2024, 67(5): 471-486.

[47] Gao J, Ding X, Qin B, et al. Is chatgpt a good causal reasoner? a comprehensive evaluation[J]. arXiv preprint arXiv:2305.07375, 2023.

[48] Bauer K, Hinz O, van der Aalst W, et al. Expl (AI) n it to me explainable AI and information systems research[J]. Business & Information Systems Engineering, 2021, 63: 79-82.

[49] Schemmer M, Kühl N, Benz C, et al. On the influence of explainable AI on automation bias[J]. arXiv preprint arXiv:2204.08859, 2022.

[50] Ehsan U, Riedl M O. Explainability pitfalls: Beyond dark patterns in explainable AI[J]. Patterns, 2024, 5(6).

[51] Morrison K, Spitzer P, Turri V, et al. The impact of imperfect XAI on human-AI decision-making[J]. Proceedings of the ACM on Human-Computer Interaction, 2024, 8(CSCW1): 1-39.

[52] Chanda S S, Banerjee D N. Omission and commission errors underlying AI failures[J]. AI & society, 2024, 39(3): 937-960.

[53] Nahar M, Seo H, Lee E J, et al. Fakes of varying shades: How warning affects human perception and engagement regarding LLM hallucinations[J]. arXiv preprint arXiv:2404.03745, 2024.

[54] Gosmar D, Dahl D A. Hallucination Mitigation using Agentic AI Natural Language-Based Frameworks[J]. arXiv preprint arXiv:2501.13946, 2025.

[55] Pinski M, Benlian A. AI literacy for users A comprehensive review and future research directions of learning methods, components, and effects[J]. Computers in Human Behavior: Artificial Humans, 2024: 100062.

[56] Walter Y. Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education[J]. International Journal of Educational Technology in Higher Education, 2024, 21(1): 15.

[57] Jin Y, Martinez-Maldonado R, GaŠevic' D, et al. GLAT: The Generative AI Literacy Assessment Test[J]. arXiv preprint arXiv:2411.00283, 2024.

[58] Nazim B. The use of fact-checking instruments in enhancing media [J]. Modern Problems in Education and Their Scientific Solutions, 2024,1(2): 440-444.

[谭心瑶：北京师范大学新闻传播学院硕士研究生；闫文捷（通讯作者）：北京师范大学新闻传播学院教授]