

專輯論文

有依據就可信嗎？探討引用信源對 人工智慧生成內容 (AIGC) 可信度感知的影響

閔晨^a、趙攀^b、余文婷^b

^a 華科技大學新聞與信息傳播學院，武漢市，中國大陸

^b 香港理工大學人文學院，九龍，香港

摘要

生成式人工智慧的快速發展加劇了全球範圍的資訊失序風險，如何提高公眾對人工智慧生成虛假資訊的辨識能力，已成為社會各界亟待解決的重要議題。近年來，New Bing、文心一言等平台透過為人工智慧生成內容 (artificial intelligence generated content, AIGC) 標註引用信源來增強內容透明度與可信度。然而迄今為止，尚未有研究探討提供引用信源能否有效提升公眾對AI生成內容的辨識能力，特別是當面對

閔晨，華科技大學新聞與信息傳播學院智能傳播系講師。研究興趣：智能傳播、數位治理。電郵：minchen_com@hust.edu.cn

趙攀，香港理工大學人文學院語言科學及技術學系兼職研究助理。研究興趣：可持續傳播、人機互動、公共輿論。電郵：21108382g@connect.polyu.hk

余文婷，香港理工大學人文學院語言科學及技術學系助理教授。研究興趣：公共輿論、健康傳播、人機互動。電郵：wenting.yu@polyu.edu.hk

論文投稿日期：2024年12月1日。論文接受日期：2026年3月10日。

《傳播與社會學刊》，(總)第78期(2026)

AI可能生成的虛假錯誤資訊時。基於啟發式－系統加工模型 (heuristic-systematic model, HSM) 和信源可信度理論，本研究考察了AIGC內容正確性、引用信源可信性、以及態度一致性對個體感知AIGC可信度的作用。透過兩項組間線上實驗(N = 848)，研究發現：引用信源會提高AIGC虛假資訊的誤導性。在無引用來源時，公眾對正確資訊的可信度感知顯著高於錯誤資訊；而在有引用來源時，正確資訊與錯誤資訊的可信度感知差異不顯著。此外，當AIGC為虛假錯誤資訊時，提供高可信度引用來源反而會提升其可信度感知。同時，確認偏誤效應在公眾接觸AIGC時依然顯著存在。研究結論對AIGC的治理具有重要的理論與實踐啟示。

關鍵詞：生成式人工智慧、引用信源可信性、AIGC可信度感知、AIGC虛假錯誤資訊

Special Issue Article

When Transparency Backfires: Source Citations, Confirmation Bias, and the Perceived Credibility of AI-Generated Misinformation

Chen MIN ^a, Pan ZHAO ^b, Wenting YU ^b

^a Department of Intelligent Communication, School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan, Mainland China

^b Department of Language Science and Technology, Faculty of Humanities, Hong Kong Polytechnic University, Kowloon, Hong Kong

Abstract

The rapid development of generative artificial intelligence (AIGC) has exacerbated the risk of information disorder worldwide, raising an urgent question of how to improve the public's ability to detect AI-generated misinformation. In recent years, platforms such as New Bing and Ernie Bot

Chen MIN (Assistant Professor). Department of Intelligent Communication, School of Journalism and Information Communication, Huazhong University of Science and Technology. Research interests: intelligent communication, digital governance. Email: minchen_com@hust.edu.cn

Pan ZHAO (Part-Time Research Assistant). Department of Language Science and Technology, Faculty of Humanities, The Hong Kong Polytechnic University. Research interests: sustainability communication, human-computer interaction, public opinion. Email: 21108382g@connect.polyu.hk

Wenting YU (Assistant Professor). Department of Language Science and Technology, Faculty of Humanities, The Hong Kong Polytechnic University. Research interests: public opinion, health communication, human-computer interaction. Email: wenting.yu@polyu.edu.hk

Article History: Received on 1 December 2024. Accepted on 10 March 2026.

Communication and Society, 78 (2026)

have attempted to enhance the transparency and credibility of AIGC by providing source citations. However, little is known about whether citing sources effectively improves the public's ability to discern AI-generated content, particularly when facing fabricated or misleading information. Drawing on the heuristic-systematic model (HSM) and source credibility theory, this study examines the effects of content accuracy, source credibility, and attitude congruence on individuals' perceived credibility of AIGC. Based on two between-subjects online experiments (N = 848), the findings show that source citations increase the misleading potential of AIGC. Without citations, participants rated accurate information as significantly more credible than inaccurate information; with citations, this difference became nonsignificant. Moreover, when AIGC presented false information, citing highly credible sources further increased perceived credibility. In addition, confirmation bias remained robust when individuals processed AIGC. These findings provide important theoretical and practical implications for the governance of AIGC.

Keywords: generative artificial intelligence, source credibility, perceived credibility of AIGC, AI-generated misinformation

有依據就可信嗎？

引言

近年來，生成式人工智慧 (generative artificial intelligence, GENAI) 技術取得了突破性進展。以 ChatGPT、文心一言、DeepSeek 為代表的大語言模型透過對海量數據的學習與模式識別，能夠根據用戶輸入自動生成結構化、邏輯性強且內容豐富的文本 (Wu et al., 2023)。這些模型不僅在日常問答、寫作輔助、教育輔導等場景中展現出強大的能力，還在醫療、法律、新聞等專業領域逐步得到應用 (Davis, 2020; Jayavardhini et al., 2024; Malsia & Loku, 2024; Pavlik, 2023; St John et al., 2024; The Asset, 2024)。GENAI 的高效性和便捷性推動了其在全球範圍內的快速普及。數據顯示，澳大利亞和美國的 GENAI 使用率分別為 49% 和 45%，印度高達 73%，中國和新加坡分別為 60% 和 56%，而在香港，88% 的知識型工作者已將該技術應用於日常工作 (Microsoft Hong Kong, 2024; Prophet, 2024; Salesforce, 2025)。

然而，伴隨其廣泛應用，人工智慧生成內容 (artificial intelligence-generated content, AIGC) 的可信度問題也逐漸引起社會各界的關注。由於 GENAI 的生成機制依賴於大規模數據訓練，其輸出內容可能存在事實錯誤、虛假資訊或誤導性表達 (Baidoo-Anu & Ansah, 2023; Dempere et al., 2023; Laymouna et al., 2024)。在缺乏專業知識或事實核查的情況下，用戶往往難以準確判斷 AIGC 的正確性，這不僅可能動搖其對 AI 系統的信任基礎，還會增加資訊誤導與決策偏差的風險。在此背景下，深入研究影響用戶對 AIGC 可信度感知的關鍵因素，具有重要的理論價值和現實意義。

資訊來源作為影響內容可信度的重要因素，長期以來受到傳播學、心理學等領域的廣泛關注。已有研究表明，來自政府機構、知名媒體或學術權威的資訊，往往更容易獲得受眾的信任；而匿名、來源不明、專業性較低或聲譽較差的資訊，則容易被質疑甚至忽視 (Zhang et al., 2016)。因此，許多學者開始關注信源對 AIGC 可信度感知的影響 (Amos & Zhang, 2024; Kirk & Givi, 2025; Kocoń et al., 2023)。然而，現有研究多將「AI」本身視作信源展開探討 (Amos & Zhang, 2024; Kocoń et al., 2023; La-Rosa Barrolleta & Sandoval-Martín, 2024)，鮮少關注 AIGC

《傳播與社會學刊》，（總）第78期（2026）

所引用的外部資訊來源對可信度感知的影響。事實上，部分GENAI模型在生成內容時會主動或應用戶請求提供來源引用，以增強資訊的權威性與說服力。例如，微軟Copilot不僅檢索資料庫中與查詢相關的網頁與內容，還在結果中附加來源鏈接，方便用戶訪問原始網頁獲取更詳盡的資訊與上下文。引用來源本身的可信度差異，可能對用戶對AIGC的整體信任判斷產生重要影響，但這一作用機制迄今尚缺乏系統的實證檢驗。更重要的是，隨著GENAI降低虛假資訊生產與傳播的門檻 (Kreps et al., 2022)，虛假內容若附帶可信度高的來源引用，其誤導性可能更甚 (Menczer et al., 2023)。因此，亟需探討AIGC本身與其引用信源在可信度感知上的交互作用，以彌補理論與實踐的不足。

基於此，本文在中國內地開展了兩項在線實驗，旨在系統檢驗AIGC引用信源的可信度對AIGC可信度感知的影響。本研究不僅有助於揭示用戶在AI資訊環境下的信任機制，也為優化AI內容設計、提升用戶資訊素養提供了理論依據與實踐路徑。

文獻綜述

AI生成虛假資訊的風險與內在機理

GENAI是指基於生成對抗網絡 (GAN)、大型預訓練模型 (GPT)、生成擴散模型 (GDM) 等技術，透過對大規模資料集的學習和識別，自動生成文本、圖像、音視頻、代碼等多類型內容的方式或技術 (Fui-Hoon Nah et al., 2023)。與判別式AI不同，GENAI透過預測不同元素間的機率分佈來實現認知模擬，即根據給定輸入預測接下來可能出現的文本，並透過機率計算輸出結果，並在人類即時回饋資料中持續最佳化，不斷改進生成品質 (Cao et al., 2023)。目前，GENAI已發展到能夠遵循人類價值觀、具有更強的細節生成能力、拒絕知識範圍外問題、並具備建模對話歷史及上下文學習的能力。隨著人工智慧在算力、演算法和資料技術方面的持續發展，AIGC在語言風格、用詞用語、邏輯語法等方面也越來越接近人類智慧，有時已達到人類難以分辨的程度 (Jakesch et al., 2023)。

有依據就可信嗎？

生成式AI對資訊生態的深度參與也伴隨著顯著的虛假資訊風險。首先，由於GENAI的核心機制依賴機率預測，其輸出並非基於事實驗證，而是基於訓練語料的統計模式。這導致模型在缺乏相關知識或面對超出訓練範圍的問題時，容易生成結構嚴謹、語言流暢但實質錯誤或虛構的資訊，即所謂「幻覺」現象(hallucination) (Zhang et al., 2025)。其次，GENAI生成內容的可靠性高度依賴訓練資料的品質與結構。以GPT模型為例，從GPT-1.0(約4.8GB)到GPT-4.0(約數萬億tokens)，訓練資料量呈指數級增長。然而，資料規模的龐大並不必然保障其準確性、無害性及公平性。當訓練語料中存在虛假、過時或帶有偏見的資訊時，這些問題會在模型訓練、知識整合和內容生成的過程中被留存甚至放大。更重要的是，當前模型訓練高度依賴於來自開放網絡平台的非結構化資訊，涵蓋網頁、社群媒體、線上論壇、維基百科等。這些來源雖廣泛但品質參差，缺乏系統的真實性驗證與統一的審查機制，其時效性與完整性亦難以保障。即便部分GENAI系統(如New Bing、Perplexity)已具備即時檢索和附帶來源引用的功能，資訊源本身的結構性缺陷仍可能導致生成內容存在事實錯誤或誤導風險。Walters和Wilder(2023)發現，AIGC在生成引用和信源時，並非真正檢索文獻或驗證資訊，而是依賴語言模型對主題相關性和文本模式的預測。這種「猜測式引用」機制容易產生偽造或不準確的信源。GPT-3.5生成的文獻綜述中，偽造引用的比例高達55%；儘管隨著模型迭代，GPT-4在生成文獻引用時的偽造率(18%)顯著降低，但偽造和錯誤引用問題依然存在。此外，模型在資料清洗、語義重組與邏輯生成的內部處理中，即使原始資訊真實，也可能因語義歧義或重構失真而造成內容的偏差與誤導。

AIGC可信度感知

I. 作為重要啟發式線索的資訊來源

可信度是用戶基於多種因素所形成的綜合判斷，包括可信性、準確性、公正性、深度、可靠性、偏見和完整性(Johnson & Kaye, 2014)。傳統的資訊可信度研究主要聚焦於信源可信度(source credibility)和資訊內容可信度(message credibility)兩大維度(Nedelcu & Balaban, 2021;

《傳播與社會學刊》，（總）第78期（2026）

Pornpitakpan, 2004)。信源可信度理論 (Hovland et al., 1953) 指出，受眾對資訊的信任程度受到信源專業性 (expertise) 和可信賴性 (trustworthiness) 的影響。前者指信源所具備的知識、技能和能力，後者則關乎信源的意圖是否真誠、公正。高專業性和高可信賴性的信源更容易說服受眾並改變其態度。內容可信度則關注資訊本身的品質，包括事實準確性、邏輯嚴密性和論據的可靠性等。根據經典的啟發式–系統加工模型 (heuristic-systematic model, HSM)，在啟發式加工下，信源的可信度會顯著影響人們對資訊的評價與判斷 (Todorov et al., 2002)。

相較於系統式加工，啟發式加工是一種低動機、低認知投入的捷徑，個體依賴信源聲譽、呈現形式等表層線索快速形成判斷。作為一種心理捷徑，啟發式思維常在人們認知資源有限或資訊處理動機不足時被使用 (Chaiken, 1980)。大量研究發現，在識別與評估虛假資訊時，人們更有可能通過信源線索 (信源相似性或可信度) 或社交線索 (如點讚) 等心理捷徑來進行判斷 (Ali et al., 2022; Traberg et al., 2024)。隨著 AIGC 技術的不斷進步，其生成內容在語言流暢性、邏輯結構和表達方式上已高度接近人類作品，這使得用戶僅憑內容本身難以有效區分資訊的真實來源 (Kreps et al., 2022)。這種高度仿真的特性顯著增加了用戶進行系統性加工的難度與認知成本，從而促使他們更依賴啟發式線索來評估 AIGC 的可信度。

已有研究發現，AIGC 的可信度受到信源這一啟發式線索的顯著影響，但影響方向並不一致。一方面，受「機器啟發式」(machine heuristic) 正面印象的影響，個體普遍傾向於將機器視為比人類更客觀、公正和可靠的信源。這種心理捷徑使得人們將客觀性、準確性、無偏見等特性與 AI 生成內容關聯起來，從而提升其感知可信度 (Huang & Wang, 2023; La-Rosa Barrolleta & Sandoval-Martín, 2024)。另一方面，與「機器啟發式」相反，一些研究發現人們對 AIGC 仍存在先入之見，「AI 生成」標籤本身成為引發警惕的啟發式線索，顯著降低內容的感知可信度 (Altay & Gilardi, 2024; Amos & Zhang, 2024)。如 Lermann Henestrosa 和 Kimmerle (2024) 發現，將 ChatGPT 生成的科學文章標註為人類作者，能提升受眾的擬人化感知、智慧感知及一定程度的可信度。Reis 等 (2024) 將同一條醫學建議分別以「人類醫師」、「AI」、「人類 + AI」作

有依據就可信嗎？

為資訊來源標籤呈現，結果表明：只要資訊來源標籤中出現AI，實驗參與者對建議的「可靠性」評分明顯降低，遵循建議的意願也顯著減少。此外，也有研究表明，AI與人類信源在可信度上並無顯著差異（Huschens et al., 2023）；一項元分析亦指出，AI與人類在總體說服效果上相當（Huang & Wang, 2023）。

II. 第一層信源（AI系統）與第二層信源（引用信源）

隨著AIGC技術的不斷演進，其信源結構已超越單一的「人類-AI」二元模式，演變為一個多層級體系。具體而言，AI系統本身構成第一層信源，直接向用戶提供生成內容；同時，AI在生成內容時引用的外部資訊（如新聞報導、學術論文、官方網站數據等）構成第二層信源。這一「雙重信源」模式使得AIGC與傳統媒介存在顯著差異：在傳統媒介中，資訊的信源往往單一且顯性，而AIGC不僅在表層提供由自身生成的內容，還在深層依賴並整合其他來源的資訊。第二層信源的存在，不僅為用戶提供了溯源與驗證的可能，也可能成為塑造AIGC整體可信度感知的重要線索。然而，現有研究大多集中於AIGC作為資訊提供者（第一層信源）對受眾態度和信任的影響，較少關注其引用資訊來源（第二層信源）的作用機制。這一研究缺口值得深入探討，因為引用信源往往來自人類生成內容，不僅可能增強資訊的權威感與可驗證性，還可能在一定程度上降低用戶對AI信源的負面偏見，進而改變其整體可信度評估。因此，本文提出第一個研究假設：

H1：相較於不提供引用信源，提供引用信源會提高公眾對AIGC的可信度感知。

與傳統媒介中顯性且單一的信源呈現方式不同，AIGC在生成內容時引用的信源往往來自多樣化的網絡資源，這些資源在專業性與可信賴性上可能存在顯著差異。例如，在回答健康醫學相關的問題時，AI的引用來源既可能包括權威的科研機構、政府或醫院官網、知名科普平台，也可能包含專業性較低的健康類自媒體。儘管AIGC通常從與問答主題高度相關的網頁中抽取資訊，但主題相關性並不必然等同於專業性，更不能確保其在事實準確性與論述嚴謹性上的可靠性。此外，網絡資訊質量參差不齊，即便來源具備較高的專業性，其可信性仍可

《傳播與社會學刊》，（總）第78期（2026）

能因資訊發布主體的立場、公正性或事實核查能力的差異而存在不確定性。根據信源可信度理論，高可信性資訊來源是提升受眾可信度感知的關鍵因素 (Hovland et al., 1953)。類似地，MAIN模型中的代理性 (agency) 線索強調，受眾對資訊生產者身份及其可靠性的感知會顯著影響受眾對資訊整體可信度的判斷 (Sundar, 2008)。在AIGC情境下，引用信源的專業性與可信性共同構成了代理性線索的重要組成部分。高可信性引用 (第二層信源) 能夠強化受眾對資訊生產主體 (第一層信源) 的積極認知，並透過代理性路徑提升其對內容的整體信任；反之，低可信性引用則可能削弱代理性線索的正向作用，甚至引發對內容質量的懷疑。基於上述分析，本文提出第二個研究假設：

H2：相較於提供低可信性引用信源，提供高可信性引用信源將顯著提升公眾對AIGC的可信度感知。

III. AIGC虛假資訊的引用信源效應

既有研究發現，資訊來源在虛假資訊的判斷與接受過程中具有顯著作用。Walter與Tukachinsky (2020) 透過元分析發現，即便資訊本身存在錯誤，其來源被標註為可信來源時，受眾仍傾向給予較高的信任。並且，只要最初來源被視為可信，錯誤資訊對受眾判斷的影響往往會持續存在。同樣，Pennycook和Rand (2019) 指出，信源可信度常被受眾作為啟發式線索，在資訊真偽難以判斷時，可信的信源能夠強化資訊的重複效應，提高虛假資訊被接受的可能性。

如前文所述，GENAI在生成內容時可能產生事實錯誤或虛假資訊。Danry等人 (2024) 的研究表明，即便受眾具備較高的認知反思能力，當AI系統生成有誤導性但邏輯上具有說服力的解釋時，這種「欺騙性解釋」仍比「誠實但不全面的解釋」更易放大人們對虛假新聞的信任。在此情境下，Li和Yang (2024) 發現，AIGC標籤雖然能幫助用戶區分AI和人類生成內容，但對虛假資訊的遏制作用微乎其微，甚至可能在某些情況下略微增強了虛假資訊的分享意圖和感知準確性。Lu等人 (2022) 的研究也發現，當AI提供可信度提示時，用戶無論其真偽均會傾向於接受其判斷結果。這表明，用戶可能將AI本身 (第一層信源) 視為一種具備「權威屬性」的信源，從而在面對虛假資訊時削弱了自身

有依據就可信嗎？

的批判性評估意願。類似地，當AIGC呈現「可點可查」的引用資訊（第二層信源）時，受眾可能將「可溯源性」誤作「可信度」的替代指標，即便內容為假，仍可能因引用的存在與表面權威而獲得較高的可信度評價，尤其是當引用信源可信度較高時。

但是，相關研究也提示引用信源可能產生相反效果。一方面，AIGC存在幻覺引用與虛構文獻的現象（McGowan et al., 2023），在虛假資訊情境中，如果引用來源缺乏公認權威性，甚至可能激發受眾的反向懷疑（Lewandowsky et al., 2012）。另一方面，Pehlivanoglu等人（2022）採用兩項實驗研究了分析性推理能力與新聞信源可信度對真實與虛假新聞文章的判斷影響。當新聞來自可信性來源時，受眾更加傾向分辨真偽，信源資訊本身顯著影響了其可信度判斷。因此，若GENAI提供引用來源，反而在一定程度上會提高人們的驗證動機或批判性思維，從而降低人們對虛假資訊的接受度。

基於此，本研究提出以下研究問題：

RQ1：當AIGC為虛假資訊時，提供引用信源是否影響公眾對AIGC的可信度感知？

RQ2：當AIGC為虛假資訊時，引用信源的可信性是否影響公眾對AIGC的可信度感知？

IV. 意見(不)一致、引用信源與AIGC可信度感知

動機性推理理論（motivated reasoning）指出，個體的既有信念與態度在資訊處理過程中發揮關鍵作用（Kunda, 1990）。人們往往傾向於接觸和相信與自身立場一致的資訊，而對立場相衝突的資訊則表現出抵觸與懷疑（Redlawsk, 2002）。在虛假資訊情境下，這種一致性偏好尤為顯著。當資訊與個體立場相符時，其更易被接受與記憶；反之，當資訊與其立場衝突時，個體則更可能啟動反向推理，對資訊進行批判性審視並尋找反駁證據（Tandoc et al., 2020）。

在AIGC情境下，立場一致性與信源存在如何相互作用，仍有待深入探討。Lu等人（2022）的研究發現，當AI提供可信度提示或信源資訊時，無論該評價是否準確，用戶往往會將自身對新聞真實性的判斷與AI的評價對齊。這表明，當AI生成與受眾立場一致的內容時，引用信

《傳播與社會學刊》，（總）第78期（2026）

源可能被視為進一步支持既有信念的啟發式線索，從而提升其可信度感知。相反，當AI生成的內容與受眾立場不一致時，引用信源的作用可能更為複雜。一方面，引用信源作為潛在權威資訊，可能在一定程度上緩解因立場不一致帶來的認知抗拒，促使受眾進行更系統化的資訊加工；另一方面，引用信源也可能成為受眾質疑與反駁的焦點，尤其是在其可信度較低時，從而強化既有立場。基於此，本研究提出第三個研究問題：

RQ3：在AIGC與受眾立場一致或不一致的情境下，引用信源的存在是否會顯著影響其對錯誤與正確資訊的可信度感知？

研究方法與結果

為檢驗上述研究假設，本研究開展了兩項線上實驗。實驗一選擇「地震雲能否預測地震」作為實驗材料，考察GENAI生成答案的正確性及引用信源的可信性對AIGC感知可信度的影響，採用3（引用信源：無來源 vs. 低可信性 vs. 高可信性） $\times$ 2（答案正確性：正確回答 vs. 錯誤回答）的組間實驗設計。為檢驗實驗的外部效度，實驗二採用與實驗一相同的實驗設計，選擇「氣候變化是否只是自然變化的一部分」作為實驗材料。

實驗一

I. 實驗設計及實驗流程

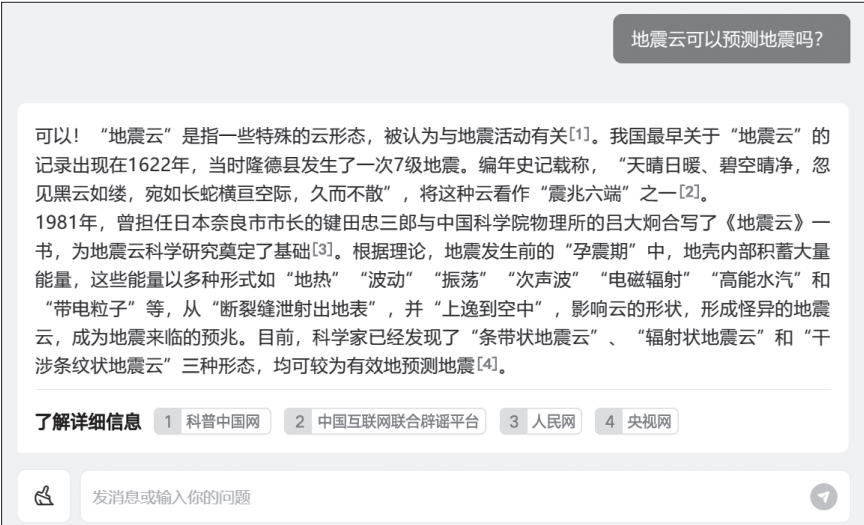
實驗採用雙因素組間實驗設計，參與者被隨機分配至六個實驗組之一。每位參與者將觀看一張模擬人類與會話式AI（conversational AI）互動的對話圖片，這些圖片在AI生成答案的「正確性」和「引用信源」兩個維度上存在差異。為建構實驗圖片，研究首先明確實驗話題。前測一列舉了10條易被公眾誤解的健康與科學類觀點，這些觀點在學術界已有定論，且廣泛被專業平台所澄清，如「地震雲可以預測地震」、「氣候變化只是自然變化的一部分」、「喝咖啡會導致骨質疏鬆」等。前測一

有依據就可信嗎？

邀請了100位參與者，評估他們對每個觀點的態度、興趣程度及熟悉程度。綜合考慮後，最終選定公眾態度分歧較大（方差較大）的「地震雲能否預測地震」作為正式實驗話題。其次，操縱答案的正確性。參考人工智慧平台回答及相關科普資訊，研究團隊撰寫了三則與科學一致的正确答案（如「地震雲與地震發生無關」）與三則觀點有誤的答案（如「地震雲可能預示地震即將發生」）。為排除其他文本特徵（如論證品質、感知的作者身份、篇幅長度）對實驗結果的干擾，研究開展了前測二，並透過配對樣本 t 檢定，篩選出在上述特徵上無顯著差異的一組正誤答案作為正式實驗材料。在此基礎上，研究進一步操縱引用信源的可信性。前測三羅列了20個真實網站作為可能的AI資訊引用來源（如「華美生物」、「百家號」、「商談偽科學」、「120搜尋網」、「人民網」、「央視網」等）。這些來源均曾在ChatGPT、New Bing、文心一言等會話式人工智慧平台回答科學類問題時被引用。前測共招募了200位參與者，並隨機分配每人對其中10個網站進行可信度評分（1 = 完全不可信，7 = 完全可信）。根據評分結果，研究從中篩選出得分最高的四個引用來源與得分最低的四個引用來源，分別用於後續實驗材料中正確答案與錯誤答案的資訊出處，以操控引用信源的可信性。為驗證高可信性與低可信性信源之間的顯著差異，研究進一步對高可信組中得分最低的網站（「科普中國」， $M = 6.08$, $SD = .93$ ）與低可信組中得分最高的網站（「商談偽科學」， $M = 4.35$, $SD = 1.34$ ）進行了配對樣本 t 檢定（ $t = -10.03$, $p < .001$ ），結果表明兩類來源在受眾認知中的可信度存在顯著差異。

每位參與者在實驗開始前被告知其將參與一項關於智慧資訊認知的調查研究。首先，參與者需報告其對「地震雲預測地震」的態度、熟悉度，以及生成式AI的使用頻率與熟悉程度等問題。隨後，參與者進入AI生成答案的閱讀階段。實驗設置情境為：有一天他們對「地震雲能否預測地震」感到疑惑，於是向人工智慧諮詢了這一問題，實驗呈現其與AI的對話截圖。為確保參與者充分理解文本內容，系統設置圖片閱讀不少於30秒。閱讀結束後，參與者需完成感知可信度等相關問卷，並回憶AI所提供的答案以進行注意力檢測。最後，參與者填寫性別、年齡與受教育程度等基本資訊。所有參與者均為自願參與，並簽署知情同意書，實驗結束後獲得相應報酬。

圖一 實驗材料示例圖（實驗一）



II. 實驗變數及測量

AIGC感知可信度。參考Lee等人(2020)提出的量表，採用語義差異法測量參與者對實驗材料中AI生成答案的可信度感知。量表包括五對雙極性形容詞：「不可信的/可信的」、「不可信服的/可信服的」、「不值得信賴的/值得信賴的」、「不準確的/準確的」、「無用的/有用的」。每項採用五點量表評分，五項平均得分作為感知可信度指標，得分越高表示感知可信度越高 ($M = 4.22$, $SD = .70$, $\alpha = .87$)。

態度一致性。態度一致性透過參與者的既有立場與AI生成答案的立場是否一致進行編碼。「完全不同意」或「比較不同意」被視為持有正確立場，「比較同意」或「完全同意」被視為持有錯誤立場。若持有錯誤立場的參與者隨機進入正確答案組，或持有正確立場的參與者隨機進入錯誤答案組，則被判定為態度不一致（記作0）；反之為態度一致（記作1） ($M = .58$, $SD = .49$)。

控制變數。本研究的控制變數主要包括：生成式AI使用頻率、生成式AI感知熟悉度和話題熟悉度。生成式AI使用頻率主要詢問參與者在日常生活中使用生成式AI的頻率有多高（1 = 從不，5 = 總是） ($M = 2.96$, $SD = .75$)；生成式AI感知熟悉度主要詢問參與者認為自己對生成式AI的熟悉與了解程度有多高（1 = 完全不熟悉，5 = 非常熟悉） ($M = 4.03$,

有依據就可信嗎？

$SD = .72$)；話題熟悉度主要詢問受試者對「地震雲能否預測地震」這一話題的了解程度(1 = 完全不了解，5 = 非常了解)($M = 3.38$, $SD = 1.16$)。

III. 實驗參與者

本研究委託專業調查實驗平台，透過購買樣本服務的方式開展線上實驗。基於事前效力分析(prior power analysis)，本研究所需最小樣本量為360人(預期效應大小 = .25，顯著性水準 = .05，統計效力 = .90)，共招募實驗參與者500名。剔除未通過注意力檢測($N = 18$)、在「地震雲預測地震」問題上沒有明確立場($N = 71$)，以及回答時間過短或過長的問卷後($N = 36$)，最終獲得有效樣本數據375人(47.5%男性， $M_{\text{年齡}} = 30.97$ ， $SD_{\text{年齡}} = 8.00$ ，91%獲得本科及以上學歷)。與中國互聯網發展狀況統計報告(中國互聯網絡信息中心，2026)相比，樣本中受過高等教育者及年輕人比例略高。不同組別在性別($F(5, 369) = .13$, $p = .99$)、年齡($F(5, 369) = .67$, $p = .65$)、學歷($F(5, 369) = 1.26$, $p = .28$)以及個人月收入($F(5, 369) = 1.75$, $p = .12$)等人口統計變量上均不存在顯著差異，表明實驗分組具有良好的隨機性與可比性。

IV. 數據分析

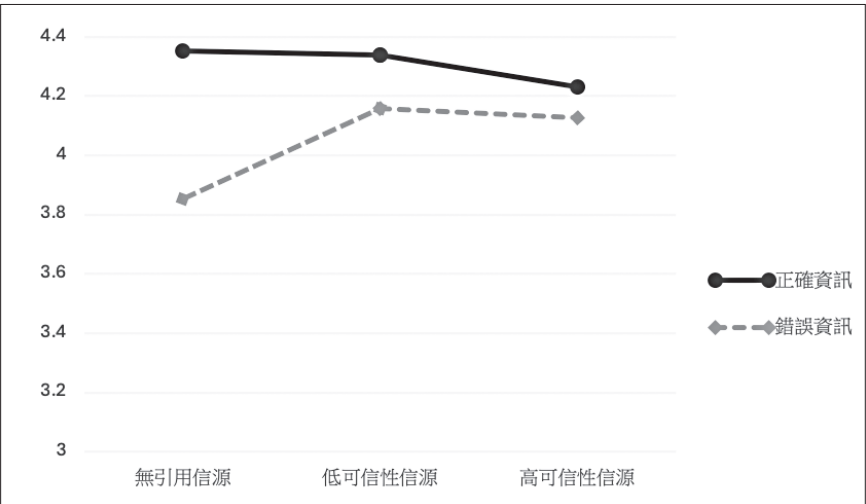
多因素方差分析結果(見表一)顯示，引用信源可信性對AIGC感知可信度的主效應不顯著($F(2, 360) = 1.57$, $p = .21$)。成對比較結果表明，與未提供引用信源($M_{\text{無來源}} = 4.10$, $SD_{\text{無來源}} = .06$)相比，提供引用信源($M_{\text{低可信性}} = 4.25$, $SD_{\text{低可信性}} = .06$; $M_{\text{高可信性}} = 4.18$, $SD_{\text{高可信性}} = .06$)並未顯著提高AIGC的感知可信度，假設1不成立。低可信性與高可信性引用信源之間的可信度感知差異也並不顯著，假設2不成立。引用信源可信性與答案正確性的交互作用顯著($F(2, 360) = 3.24$, $p = .04$)。在缺乏引用信源的情況下，正確答案($M = 4.35$, $SD = .08$)的可信度感知顯著高於錯誤答案($M = 3.85$, $SD = .08$)；而當引用信源存在時，無論其可信性高低，正確答案與錯誤答案的可信度感知差異均不顯著。進一步分析發現，當GENAI生成正確答案時，引用信源的有無對可信度感知無顯著影響；然而，當GENAI生成錯誤答案時，提供引用信源會顯著提升其可信度感知，且該效應不受引用信源可信性高低的影響($M_{\text{無來源}} = 3.85$, $SD_{\text{無來源}} = 0.08$; $M_{\text{低可信性}} = 4.16$, $SD_{\text{低可信性}} = .08$; $M_{\text{高可信性}} = 4.13$, $SD_{\text{高可信性}} = .08$) (見圖二)。

《傳播與社會學刊》，（總）第 78 期（2026）

表一 多因素方差分析

	實驗一						實驗二					
	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2		<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	η^2	
答案正確性	1	5.94	14.64	.00	.04		1	37.91	74.66	.00	.14	
引用信源	2	.64	1.57	.21	.01		2	.48	.94	.39	.00	
態度一致性	1	8.51	20.97	.00	.06		1	23.18	45.65	.00	.09	
答案正確性 × 引用信源	2	1.32	3.24	.04	.02		2	2.09	4.12	.02	.02	
答案正確性 × 態度一致性	1	14.49	35.68	.00	.09		1	13.89	27.35	.00	.06	
引用信源 × 態度一致性	2	.34	.84	.44	.01		2	.57	1.13	.33	.01	
引用信源 × 答案正確性 × 態度一致性	2	.37	.90	.41	.01		2	1.44	2.83	.06	.01	
話題熟悉度	1	.01	.02	.90	.00		1	.09	0.17	.68	.00	
生成式 AI 使用 頻率	1	5.50	13.55	.00	.04		1	1.19	2.34	.13	.00	
生成式 AI 感知 熟悉度	1	4.03	9.94	.00	.03		1	.01	.02	.89	.00	
對 AI 的態度							1	7.84	15.45	.00	.03	
誤差	360	.41					457	.51				
總計	375						473					

圖二 引文來源可信性與內容正確性對公眾 AIGC 可信度感知的交互效應圖（實驗一）



有依據就可信嗎？

針對研究問題三，結果顯示態度一致性對AIGC感知可信度的主效應顯著 ($F(1, 360) = 20.97, p < .001$)。態度一致的AIGC ($M = 4.34, SD = .05$) 可信度顯著高於態度不一致的AIGC ($M = 4.01, SD = .05$)。引用信源與態度一致性的交互作用不顯著 ($F(2, 360) = .84, p = .44$)。即無論是否提供引用信源，只要AIGC與受眾態度一致，其可信度均較高。但是，態度一致性與資訊正確性的交互作用顯著 ($F(1, 360) = 35.68, p < .001$)。當資訊與個體態度不一致時，正確答案 ($M = 4.34, SD = .07$) 的可信度顯著高於錯誤答案 ($M = 3.68, SD = .08$)；而當資訊與個體態度一致時，正確答案 ($M = 4.27, SD = .07$) 與錯誤答案 ($M = 4.41, SD = .06$) 之間的感知可信度差異不顯著。

此外，研究發現，生成式AI使用頻率 ($F(1, 360) = 13.55, p < .001$) 與生成式AI感知熟悉度 ($F(1, 360) = 9.94, p < .001$) 對感知可信度均具有顯著的正向影響，表明對AI技術接觸越頻繁、主觀認知程度越高的用戶，往往對AIGC的可信度感知更高。但是，話題熟悉度對AIGC可信度感知的影響不顯著 ($F(1, 360) = .02, p = .90$)。

V. 結論

實驗一結果表明，單純提供引用信源並不能顯著提升公眾對AIGC的感知可信度，無論引用信源可信性高低。然而，引用信源在個體對AIGC可信度判斷中扮演著關鍵的調節作用。具體而言，當GENAI輸出錯誤資訊時，無論其所附引用來源的可信度高低，其存在本身都會一定程度上「掩蓋」資訊的錯誤性，從而顯著提升受眾對該錯誤內容的可信度評價；相較之下，對於正確資訊，引用信源的有無則並未帶來顯著影響。這表明，引用信源的存在並非必然促進理性判斷，反而在某些情境下可能削弱公眾對錯誤資訊的識別能力，從而對AI生成內容的接受過程產生誤導作用。此外，研究進一步發現，公眾在接受AI生成資訊時仍然表現出明顯的動機推理傾向。當AI生成內容與個體既有態度一致時，即使內容本身是錯誤的，受眾也更傾向於賦予其較高可信度；而當資訊與個體態度不一致時，即使其為正確答案，也面臨顯著的可信度損失。然而，本研究未發現引用信源會顯著影響態度一致性与可信度感知之間的關係，這意味著態度一致性對AIGC可信度的影響在有無引用信源的條件下保持穩定。

《傳播與社會學刊》，（總）第78期（2026）

實驗一透過以「地震雲能否預測地震」為研究議題，初步回答了本文的主要研究假設和研究問題。儘管該議題公眾感興趣、易混淆程度高，但其本質更接近偏向科學常識的流言，且使用單一話題可能限制了研究結果的外部效度與普適性。因此，實驗二在保持實驗邏輯一致的基礎上更換研究話題，以進一步檢驗研究結論的穩定性與適用性。

實驗二

I. 實驗設計與實驗流程

實驗二在設計思路與實驗一一致，但更換實驗話題為「氣候變化是否只是自然變化的一部分」。選擇該話題有兩點原因：第一，相較於相對中立且純粹的科學流言（「地震雲」），公眾對氣候變化議題往往持有預設立場或政治傾向，這種「立場先行」的心理圖式可能會干擾個體對資訊真實性的客觀判斷。因此，透過引入這一話題，實驗二可以驗證實驗一中所發現的「引用信源誤導效應」是否具有跨情境的穩健性。即在面對社會化傾向更強、議題更熟悉、認知衝突更明顯的議題時，引用信源這一啟發式線索是否依然能夠繞過個體的理性加工，從而影響受眾對AI生成資訊的信任。第二，前測一結果顯示，公眾對該議題的熟悉度更高，且對正確與錯誤觀點的判斷比例相對均衡，有利於實驗材料的平衡設置。選定該話題後，研究者參照權威資料，分別撰寫三則正確答案（「氣候變化與人類活動有關」）與三則錯誤答案（「氣候變化只是自然變化的一部分」），並根據前測結果篩選實驗材料，以排除其他資訊特徵干擾。引用信源的操縱方式與實驗一相同：每則答案透過腳註形式標註四個低可信性或四個高可信性引用來源。最終，製作出六張模擬人與會話式AI互動的實驗圖片，實驗流程亦與實驗一保持一致。

II. 實驗變數及測量

AIGC感知可信度 ($M = 3.96, SD = .82, \alpha = .89$) 和態度一致性 ($M = .47, SD = .50$) 的測量方式與實驗一相同。控制變數方面，除話題熟悉度 ($M = 3.91, SD = .82$)、生成式AI使用頻率 ($M = 2.68, SD = .84$)、生成式AI感知熟悉度 ($M = 3.70, SD = .79$) 外，實驗二新增「公眾對人工

有依據就可信嗎？

智慧的態度」作為控制變數。此前研究指出，該變數會顯著影響其對人工智慧生成內容的感知(Grassini & Koivisto, 2024)。本研究參考Bergdahl等人(2023)開發的量表，在五點量表上測量參與者對八項陳述(如「人工智慧可以為國家提供新的經濟機遇」，1 = 完全不贊同，5 = 完全贊同)的贊同程度，其平均得分代表參與者對AI的整體態度，得分越高表示態度越積極($M = 3.86$, $SD = .57$, $\alpha = .73$)。

III. 實驗參與者

實驗二同樣委託專業調查公司在線上開展。共招募650名參與者，剔除未通過注意力檢測者($N = 57$)、在「氣候變化是否只是自然變化的一部分」問題上無明確立場者($N = 78$)、以及回答時間異常者($N = 42$)，最終獲得473份有效問卷(男性佔47.6%， $M_{\text{年齡}} = 30.15$ ， $SD_{\text{年齡}} = 8.22$ ，79.8%具有本科及以上學歷)。與實驗一一致，參與者總體受教育程度較高且以年輕群體為主。方差檢驗結果顯示，實驗參與者在不同情境中分佈均衡，符合隨機化要求。

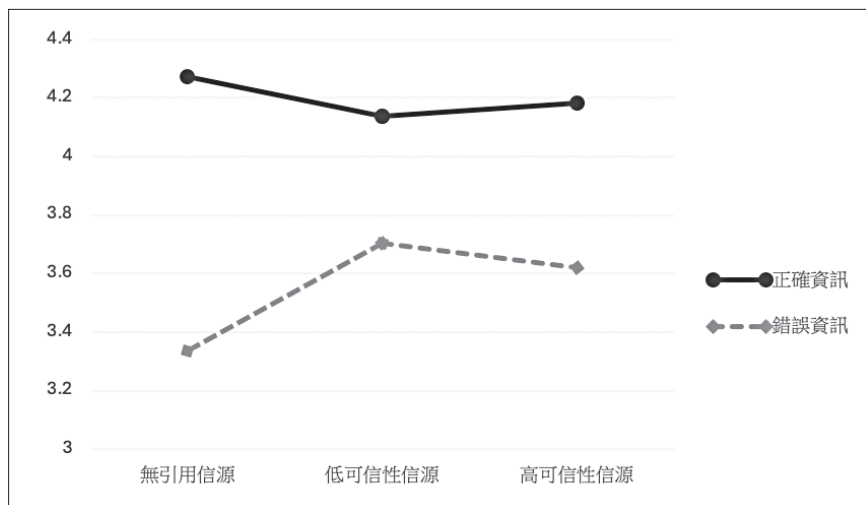
IV. 數據分析

多因素方差分析表明，引用信源的主效應不顯著($F(2, 457) = .94$, $p = .39$)，但其與資訊正確性之間存在顯著交互作用($F(2, 457) = 4.12$, $p = .02$)。具體而言，在資訊錯誤的情境下，提供引用信源(無論可信度高低)會顯著提升公眾對AI生成內容的可信度感知($M_{\text{無來源}} = 3.34$ ， $SD_{\text{無來源}} = .09$ ； $M_{\text{低可信性}} = 3.70$ ， $SD_{\text{低可信性}} = .09$ ； $M_{\text{高可信性}} = 3.62$ ， $SD_{\text{高可信性}} = .09$) (見圖三)。而在資訊正確的情境下，提供引用信源對AIGC感知可信度的影響並不顯著。對AIGC感知可信度的主效應顯著($F(1, 457) = 45.65$, $p < .001$)，但其與引用信源的交互作用不顯著($F(2, 457) = 1.13$, $p = .33$)。態度一致性與資訊正確性之間的交互作用顯著($F(1, 457) = 27.35$, $p < .001$)。公眾對與其態度不一致的錯誤答案的可信度($M = 3.09$, $SD = .09$)顯著低於態度一致的錯誤答案($M = 4.01$, $SD = .06$)。控制變數方面，實驗二特別考察了受眾對AI的態度與AIGC可信度感知的影響。結果表明，公眾對AI的態度越積極，其對AIGC的可信度感知越高($F(1, 457) = 15.45$, $p < .001$)。值得注意的是，儘管氣候議題

《傳播與社會學刊》，（總）第78期（2026）

的話題熟悉度 ($M = 3.91, SD = .82$) 顯著高於實驗一中的「地震雲」議題 ($M = 3.38, SD = 1.16$)，話題熟悉度對AIGC可信度的影響仍不顯著 ($F(1, 457) = .17, p = .68$)。綜上，實驗二在排除了一系列潛在干擾變數後，再次驗證並拓展了實驗一的主要結論。

圖三 引文信源可信性與內容正確性對公眾AIGC可信度感知的交互效應圖（實驗二）



V. 討論

透過更換研究話題並採用相同的實驗設計，實驗二再次驗證了實驗一的核心發現，顯著提升了研究結論的穩健性與外部效度。在話題選擇上，實驗二從偏向純粹科學常識的「地震雲」切換為具有高度公共爭議性與立場偏好屬性的「氣候變化」議題。這一跨議題的驗證表明，引用信源的誤導效應不僅存在於知識真空的科普情境，在價值立場先行、話題熟悉度、社會討論度較高的公共議題中依然穩定存在。

首先，引用信源的可信性本身並不直接影響AIGC感知可信度，假設1和2在兩個實驗中均未獲得支持。其次，實驗二進一步證實了引用信源與答案正確性之間的交互作用：當資訊錯誤時，無論引文信源的可信性高低，提供引用信源均會顯著提高AI生成答案的感知可信度，這一發現再次揭示了引用信源在誤導性資訊傳播中的潛在放大效應。即使在受眾對話題更為熟悉、具備更高辨識基礎的情境下，引用線索

有依據就可信嗎？

依然能有效誘導認知惰性，透過賦予資訊某種「符號性權威」來干擾用戶的真實性判斷。此外，實驗二也觀察到答案正確性與態度一致性之間的顯著交互作用，與自身態度一致的錯誤資訊被賦予更高的可信度，進一步佐證了態度一致性偏差在資訊判斷中的關鍵作用。

討論與總結

隨著 AIGC 深度嵌入知識生產與傳播，虛假或錯誤資訊的生成幾乎成為其內在機制的必然產物。在「可解釋性」與「透明性」成為 AI 治理核心原則的語境下，提供引文鏈接與資訊來源不僅是技術開發者的一種策略性設計，更被寄予厚望作為對抗虛假資訊擴散、重建知識溯源鏈條的關鍵手段。但迄今為止，鮮有研究關注引文信源能否及如何影響公眾對 AI 生成虛假資訊的判斷。基於此，本研究透過兩個線上實驗，系統考察了資訊來源與態度一致性在公眾對 AIGC 可信度感知中的作用機制。研究結果不僅有助於揭示用戶在 AI 生成內容情境下的認知加工與判斷邏輯，同時也為 AIGC 虛假資訊治理提供了實證依據，對促進資訊生態的有序與健康發展具有重要意義。本研究的發現可從以下三個方面進行深入探討：

引用信源的「透明度悖論」

既有研究證實，提供信源能夠增強資訊的透明性和可驗證性，從而提升其可信度感知 (Peifer & Meisinger, 2021; Winterlin et al., 2020)。然而，本研究揭示了一個關鍵的「透明度悖論」：AI 所提供的引用信源，並非總能有效提升可信度，反而可能在特定情境下成為誤導受眾、加劇認知迷失的工具。我們的研究發現，引用信源本身並未對公眾的 AIGC 可信度感知產生直接影響，但其存在卻顯著提高了人們對 AI 生成的虛假資訊的可信度。更重要的是，這一效應與引用信源的可信性高低無關，其存在本身就具備強大的誤導性。

這一發現挑戰了傳統的信源可信度理論 (Hovland et al., 1953)，即信源的可信性 (如專業性、公正性) 是影響說服力的關鍵變數。這表

《傳播與社會學刊》，（總）第78期（2026）

明，在AIGC情境下，引用信源的「存在性」可能比其「可信性」更具決定性。這是因為AIGC所生成的引用信源被賦予「符號性權威」的特徵，即其所傳遞的並非實質性內容，而是一種足以改變受眾判斷的象徵性信任標識。Li和Yang（2024）研究發現，「AI生成」標籤作為第一層信源，對虛假資訊的遏制作用有限。本研究進一步深化了這一結論，即作為第二層信源的引用信源同樣具備強大的啟發式作用，且方向更具誤導性。AI不僅能夠通過自身標籤與身份屬性影響用戶的可信度感知，更能藉助外部信源的嫁接來完成自我偽裝與合法化。

值得注意的是，在我們的兩個實驗中，AI生成內容的正確性對AIGC可信度感知的主效應顯著，這意味著公眾能夠識別AI生成資訊的正確與否。然而，即使在這種「知情」的背景下，引用信源仍然能夠提高虛假資訊的可信度感知。這一發現與Ecker等人（2022）的研究高度契合，他們的研究表明，當錯誤資訊得到「支持性證據」（哪怕是虛假的）時，即便受眾已經意識到資訊本身存在問題，其可信度仍會顯著增強。因此，AI所提供的引用信源，無論其真實質量如何，都可能被視作一種虛假的支持性證據，利用人們對透明度和可驗證性的本能需求，操縱其認知。

引用信源誘發認知惰性風險

本研究的發現不僅揭示了引用信源的「透明度悖論」，更進一步指向了引用信源誘發的認知惰性與系統性信任風險。研究結果表明，在缺乏引用信源的情況下，個體能夠較為有效地識別AI生成內容的正誤；然而，一旦提供引用信源，其對虛假資訊的識別能力便顯著下降。這一效應在不同實驗情境中均得到驗證。無論受眾對相關議題是否熟悉、是否具備專業背景，在AIGC所呈現的「可溯源」表象面前，其批判性判斷能力均受到不同程度的削弱。即便是對話題較為熟悉的個體，也未能對帶有引用的虛假資訊形成有效免疫。這表明，AIGC並非僅通過內容本身塑造認知，更是通過「引用信源」這一符號性機制，牽引個體進入認知惰性與錯誤信任的陷阱。

有依據就可信嗎？

實際上，公眾對AI本身存在一種先驗信任。正如Spitale等人(2023)的研究所示，在人類自身固有缺陷和對「事實」認定標準多元化的前提下，人們更傾向於將人工智慧虛構的資訊作為事實看待。這種對機器的系統性信任在我們的研究中也得到了印證。在實驗二中，受試者對AI整體持積極態度($M = 2.14$, $SD = .57$, $1 =$ 完全不贊同, $5 =$ 完全贊同)，並且在兩個實驗中均系統性地高估了AIGC的可信度(在五點量表上，任何情況下得分均超過3)。這種「先驗信任」使得用戶在接收到AI生成內容時，傾向於賦予其較高的可信度，從而為認知惰性埋下了伏筆。

在此基礎上，引用信源作為第二層信源進一步放大了AI的說服效力。引用信源不僅充當了資訊補充，更作為一種啟發式線索，喚起人們對「透明度」和「證據」的本能需求，誘導用戶進入依賴性加工模式。在高效便捷與即時反饋的交互環境中，用戶逐漸弱化獨立的分析、推理與權衡動機，轉而直接採納AI所提供的結論。由此，引用信源並未提升用戶的驗證能力，反而通過提供一種認知支撐的幻象，使個體誤以為資訊具有更高的可靠性，進而放棄查證與批判性思維。這種遞進式信任機制解釋了為何即便在「知情」(能夠識別真偽)的情況下，引用信源仍然顯著提升了虛假資訊的可信度，最終導致用戶陷入認知迷失。

值得注意的是，隨著生成式技術的發展，AI生成的虛假資訊日益難以與真實資訊區分。在效率邏輯的驅動下，用戶極有可能停止質疑與分析，直接接受並傳播AIGC的內容，從而在知識獲取過程中造成誤解與曲解，形成新的「知識迷霧」。這種對機器的過度依賴不僅會削弱人類的批判性能力，還可能導致整體認知能力的退化(彭蘭，2023)。

AIGC的可信度感知與確認偏誤的穩定性

此外，本研究發現公眾在接受AIGC時仍然表現出明顯的確認偏誤效應。當生成內容與個體既有態度一致時，即使其為錯誤資訊，用戶也會賦予其更高的可信度；而當資訊與個體態度不一致時，即使其為正確答案，也會遭遇可信度損失。這一發現與經典的動機推理理論一

《傳播與社會學刊》，（總）第78期（2026）

致 (Kunda, 1990)，表明用戶在面對AIGC時會主動通過信念一致性來維持認知平衡。

值得關注的是，本研究發現引用信源並未顯著改變確認偏誤對AIGC可信度感知的作用。無論是相對中立的「地震雲」議題，還是社會爭議性較強的「氣候變化」，確認偏誤效應在不同情境下均顯著存在。雖然引用信源作為啟發式線索顯著提升了虛假資訊的總體可信度，並可能誘發用戶的認知惰性，但它並未進一步放大或削弱確認偏誤效應。這表明，確認偏誤似乎作為一種穩固的認知偏向，能夠獨立於引用信源而發揮作用。這種雙重作用機制揭示了AIGC可信度判斷的複雜性：人們既受到引用信源等外部線索的影響，也受到個體內在動機的自我強化影響。

實際上，AI本身的核心訓練範式是基於人類反饋的強化學習 (reinforcement learning from human feedback, RLHF)，其本質是讓模型通過與人類的互動來學習「甚麼才是好的答案」。為了獲得高分，模型會傾向於生成那些能得到人類「點讚」或「好評」的內容，這導致AI系統在面對用戶的已有信念時，會傾向於提供支持性、迎合性的資訊，而非挑戰性的客觀事實 (Gao et al., 2023)。方師師、唐巧盈 (2023) 將這種迎合行為概括為「語境坍塌」，即會通過迎合和拓展用戶有意或無意的錯誤，創造出一個看似可信實則充滿不實內容的交互語境。在這種背景下，人類固有的心理偏向與AI的內在機制共同作用，最終構建了一個封閉的、充滿偏見的資訊繭房，使得用戶在資訊消費過程中更容易陷入認知偏誤的泥潭。

理論與實踐意義

透過創新性地考察引用信源對AIGC感知可信度的影響，本研究為AIGC時代的資訊傳播理論提供了重要補充，其主要理論意義體現在以下兩個方面：第一，拓展了信源理論在AIGC語境下的應用範式。以往關於AIGC可信度的研究，多聚焦於AI系統本身作為第一層信源對用戶認知的影響 (Huschens et al., 2023; Sun et al., 2024)，而對引用信源這一第二層信源的作用機制關注不足。本研究系統考察了引用信源在

有依據就可信嗎？

AIGC 虛假資訊傳播中的獨特作用，發現其不僅是簡單的資訊補充，更是一種具有「符號性權威」的啟發式線索。

尤為重要的是，本研究透過在實驗二中引入具有高度公共爭議性與立場偏好屬性的「氣候變化」議題，與實驗一偏向科學常識的「地震雲」議題形成互補。實驗結果在不同屬性議題間表現出高度的一致性，有力證明了引用信源的誤導效應並非特定話題下的偶然現象，而是在「事實判斷」與「價值判斷」兩類認知情境中均穩定存在的「透明度悖論」。這種跨議題的穩健性發現，拓展了經典信源理論的邊界，為理解多層信源在複雜資訊環境中的作用提供了更具普適性的理論視角。

第二，本研究深化了對 AIGC 可信度感知的認知心理理解。研究表明，個體對 AIGC 內容的可信度感知不僅受外部啟發式線索影響，還與內在的動機推理及認知偏誤相互作用。這揭示了 AIGC 可信度感知的雙重機制：一方面，外部啟發式線索為認知判斷提供捷徑；另一方面，內在心理偏向形成自我強化效應，使確認偏誤在用戶認知中持續存在。該發現為制定更有效的 AIGC 資訊治理策略提供了實證依據。

本研究同樣具有顯著的實踐指導意義。首先，引用信源這一透明化手段在實踐中並未完全達到預期效果。無論信源可信度高低，其存在本身都可能對 AI 生成的虛假資訊產生誤導，揭示了所謂的「透明度悖論」：當信源本身存在不確定性或偏誤時，透明度不僅無法提升可信度，反而可能成為更具欺騙性的工具。因此，未來 AI 治理應在提供信源的同時，附加信源可信度評估或潛在風險提示，以優化資訊甄別機制。其次，本研究發現，在 AI 提供引用的高效便捷環境下，用戶更容易產生認知惰性，放棄對資訊來源的批判性審查。這表明，提升公眾的 AIGC 媒介素養至關重要，應透過教育和引導幫助用戶跳出認知惰性，主動審視和批判 AI 生成內容，從而在複雜數位環境中保持批判性思維，避免陷入由 AI 編織的認知迷霧。

研究不足與未來研究方向

本研究也存在一定的局限性。首先，樣本偏向年輕且受教育程度較高。根據 Quest Mobile (2024) 行業報告，中國排名前列的生成式人

《傳播與社會學刊》，（總）第78期（2026）

工智慧應用用戶中，69.3%年齡集中在25至45歲，71.6%來自三線及以上城市。此外，生成式人工智慧產品在金融、教育及設計等專業社區中使用較為頻繁，而這些用戶往往具有較高的AI聊天機器人使用經驗。這一樣本特徵限制了研究結果在更廣泛的中國網際網絡用戶群體中的外部效度。未來研究可考慮擴展樣本覆蓋不同年齡、教育水平及地域的用戶，以增強結果的普適性。

其次，本研究選取「氣候變化」和「地震雲」作為研究議題，而這類話題本身存在較大態度分歧且已有大量科學證據支撐。在缺乏引用信源的情況下，個體在一定程度上仍能夠辨識資訊的真偽。然而，隨著生成式人工智慧技術的發展，AI可能生成難以被普通用戶識別的虛假資訊，尤其是在不熟悉領域或公共危機事件中。此類情境下，引用信源是否仍對可信度感知產生顯著影響，未來可進一步進行實證研究。

最後，根據HSM，個體在資訊加工過程中可能既受到啟發式線索的直接影響，也透過系統性加工路徑間接作用於可信度判斷。由於本文的研究設計更強調引用信源作為啟發式線索的作用，而非對比啟發式與系統性加工兩種路徑，因此未在假設與模型中納入中介機制。未來研究可進一步探討這一潛在路徑，以揭示引用信源如何透過認知加工模式影響AIGC可信度感知，從而深化對變數間內在作用機制的理解。

披露聲明

本文作者未報告潛在的利益衝突。

Disclosure Statement

No potential conflict of interest was reported by the authors.

研究經費資助

本文獲國家自然科學基金「資訊失序風險下生成式人工智慧內容標識效應評估與優化策略」（72404097）和教育部人文社會科學研究一般項目「政治安全視角下生成式AI資訊失序風險的識別與治理研究」

有依據就可信嗎？

(24YJCZH216) 的資助。另外，本文也獲香港理工大學啟動項目「驗證新聞查證量表」(A0051092) 的資助。

Funding

This article was funded by the National Natural Science Foundation of China “Evaluation and Optimization Strategy of Generative AI Content Labeling Effects Under the Risk of Information Disorder” (72404097), and the General Program of Humanities and Social Sciences Research of the Ministry of Education “Research on Identification and Governance of Generative AI Information Disorder Risks from the Perspective of Political Security” (24YJCZH216). In addition, this study was also supported by the Startup Program of The Hong Kong Polytechnic University “Validating a News Verification Scale” (A0051092).

ORCID

閔晨 (Chen MIN) <https://orcid.org/0000-0001-8770-3398>

趙攀 (Pan ZHAO) <https://orcid.org/0009-0006-0541-0510>

余文婷 (Wenting YU) <https://orcid.org/0000-0002-1137-4759>

參考文獻

中文部分 (Chinese Section)

Quest Mobile (2024 年 12 月)。《中國互聯網發展年鑑 (2024–2025 年度核心趨勢洞察)》。取自 *Quest Mobiel*，<https://www.questmobile.com.cn/research-book/annualreport2025>。

Quest Mobile (2024, December). *China Internet Development Yearbook (Core Trend Insights 2024–2025)*. Quest Mobile. <https://www.questmobile.com.cn/research-book/annualreport2025>.

中國互聯網絡信息中心 (2026)。《第 57 次中國互聯網絡發展狀況統計報告》。中國互聯網絡信息中心。

China Internet Network Information Center. (2026). *The 57th statistical report on the development of China's internet*. China Internet Network Information Center.

《傳播與社會學刊》，(總)第78期(2026)

方師師、唐巧盈 (2023)。〈聰明反被聰明誤：ChatGPT 錯誤內容生成的類型學分析〉。《新聞與寫作》，第4期，頁31–42。

Fang, S., & Tang, Q. (2023). Smart people being misled by their own cleverness: A typological analysis of erroneous content generated by ChatGPT. *Journalism and Writing*, 4, 31–42.

彭蘭 (2023)。〈智能生成內容如何影響人的認知與創造？〉。《編輯之友》，第11期，頁21–28。

Peng, L. (2023). How AI-generated content influences human cognition and creativity? *Editorial Friend*, 11, 21–28.

英文部分 (English Section)

Ali, K., Li, C., Zain-ul-abdin, K., & Zaffar, M. A. (2022). Fake news on Facebook: Examining the impact of heuristic cues on perceived credibility and sharing intention. *Internet Research*, 32(1), 379–397.

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS nexus*, 3(10), pgae403.

Amos, C., & Zhang, L. (2024). Consumer reactions to perceived undisclosed ChatGPT usage in an online review context. *Telematics and Informatics*, 93, 102163.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62.

Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Mantere, E. S., Savela, N., & Oksanen, A. (2023). Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. *Telematics and Informatics*, 82, 102013.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv*. <https://doi.org/10.48550/arXiv.2303.04226>.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), 752–766.

Danry, V., Pataranutaporn, P., Groh, M., Epstein, Z., & Maes, P. (2024). Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. *arXiv*. <https://doi.org/10.48550/arXiv.2408.00024>.

Davis, A. E. (2020). The future of law firms (and lawyers) in the age of artificial intelligence. *Revista Direito GV*, 16(1), e1945.

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, 8, 1206936.

有依據就可信嗎？

- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29.
- Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304.
- Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (pp. 10835–10866). PMLR.
- Grassini, S., & Koivisto, M. (2024). Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. *Scientific Reports*, 14, 4113.
- Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. Yale University Press.
- Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication*, 73(6), 552–562.
- Huschens, M., Briesch, M., Sobania, D., & Rothlauf, F. (2023). Do you trust ChatGPT?—Perceived credibility of human and AI-generated content. *arXiv*. <https://doi.org/10.48550/arXiv.2309.02524>.
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. In *Proceedings of the National Academy of Sciences*, 120(11), e2208839120.
- Jayavardhini, P., Mahalakame, R. M., Srinivetha, P., & Eugene Berna, I. (2024). AI study partner: Development of an LLM and Gen AI-enhanced study assistant tool. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 8(3), 1–7.
- Johnson, T. J., & Kaye, B. K. (2014). Credibility of social network sites for political information among politically interested internet users. *Journal of Computer-Mediated Communication*, 19(4), 957–974.
- Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, 186, 114984.
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniec, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mielejszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99, 101861.
- Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1), 104–117.

《傳播與社會學刊》· (總) 第 78 期 (2026)

- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.
- La-Rosa Barrolleta, L. A., & Sandoval-Martín, T. (2024). Artificial intelligence versus journalists: The quality of automated news and bias by authorship using a Turing test. *Andlisi*, 70, 15–36.
- Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., & Lebouché, B. (2024). Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *Journal of Medical Internet Research*, 26, e56930.
- Lee, S., Nah, S., Chung, D. S., & Kim, J. (2020). Predicting AI news credibility: Communicative or social capital or both? *Communication Studies*, 71(3), 428–447.
- Lermann Henestrosa, A., & Kimmerle, J. (2024). The effects of assumed AI vs. human authorship on the perception of a GPT-generated text. *Journalism and Media*, 5(3), 1085–1097.
- Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131.
- Li, F., & Yang, Y. (2024). Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based, randomized, controlled experiment. *JMIR Formative Research*, 8, e60024.
- Lu, Z., Li, P., Wang, W., & Yin, M. (2022). The effects of AI-based credibility indicators on the detection and spread of misinformation under social influence. In *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–27.
- Malsia, E., & Loku, A. (2024). Generative artificial intelligence in health system management: Transformative insights. *Journal of Service Science and Management*, 17, 107–117.
- McGowan, A., Gui, Y., Dobbs, M., Shuster, S., Cotter, M., Selloni, A., Goodman, M., Srivastava, A., Cecchi, G. A., & Corcoran, C. M. (2023). ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326, 115334.
- Menczer, F., Crandall, D., Ahn, Y. Y., & Kapadia, A. (2023). Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence*, 5, 679–680.
- Microsoft Hong Kong. (2024, June 3). *Microsoft and LinkedIn unveil the state of AI at work in Hong Kong*. Microsoft Hong Kong. <https://news.microsoft.com/en-hk/2024/06/03/microsoft-and-linkedin-unveil-the-state-of-ai-at-work-in-hong-kong/>
- Nedelcu, D., & Balaban, D. C. (2021). The role of source credibility and message credibility in fake news engagement. Perspectives from an experimental study. *Journal of Media Research*, 14(3), 42–62.

有依據就可信嗎？

- Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism and Mass Communication Educator*, 78(1), 84–93.
- Pehlivanoglu, D., Lighthall, N. R., Lin, T., Chi, K. J., Polk, R., Perez, E., Cahill, B. S., & Ebner, N. C. (2022). Aging in an “infodemic”: The role of analytical reasoning, affect, and news consumption frequency on news veracity detection. *Journal of Experimental Psychology: Applied*, 28(3), 468–485.
- Peifer, J. T., & Meisinger, J. (2021). The value of explaining the process: How journalistic transparency and perceptions of news media importance can (sometimes) foster message credibility and engagement intentions. *Journalism and Mass Communication Quarterly*, 98(3), 828–853.
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. In *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526.
- Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades’ evidence. *Journal of Applied Social Psychology*, 34(2), 243–281.
- Prophet. (2024). *The rise of the AI-powered consumer*. Prophet. <https://prophet.com/pdf/the-rise-of-the-ai-powered-consumer/?redirectedfrom=thinking>.
- Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *Journal of Politics*, 64(4), 1021–1044.
- Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, 30(11), 3098–3100.
- Salesforce. (2025, February). *Top generative AI statistics for 2025*. Salesforce. <https://www.salesforce.com/news/stories/generative-ai-statistics/>.
- Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9(26), eadh1850.
- St John, A., Cooper, L., & Kavic, S. M. (2024). The role of artificial intelligence in surgery: What do general surgery residents think? *The American Surgeon™*, 90(4), 541–549.
- 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*. <https://doi.org/10.48550/arXiv.2403.09987>.
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). The MIT Press.
- Tandoc, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3), 381–398.
- The Asset. (2024, June 21). *HK knowledge workers tied for tops in GenAI use*. <https://www.theasset.com/article/51754/hk-knowledge-workers-tied-for-tops-in-genai-use>.

《傳播與社會學刊》·（總）第78期（2026）

- Todorov, A., Chaiken, S., & Henderson, M. (2002). The heuristic-systematic model of social information processing. In J. P. Dillard & M. Pfau (Eds.), *The persuasion handbook: Developments in theory and practice* (pp. 195–212). Sage.
- Traberg, C. S., Harjani, T., Roozenbeek, J., & Van der Linden, S. (2024). The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, 14, 4205.
- Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2), 155–177.
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045.
- Wintterlin, F., Engelke, K. M., & Hase, V. (2020). Can transparency preserve journalism's trustworthiness? Recipients' views on transparency about source origin and verification regarding user-generated content in the news. *Studies in Communication and Media*, 9(2), 218–240.
- Wu, J., Gan, W., Chen, Z., Wan, S., & Lin, H. (2023). AI-generated content (AIGC): A survey. *arXiv*. <https://doi.org/10.48550/arXiv.2304.06632>.
- Zhang, M., Chen, C., Hu, W., Chen, L., & Zhan, J. (2016). Influence of source credibility on consumer acceptance of genetically modified foods in China. *Sustainability*, 8(9), 899.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4), 1373–1418.

本文引用格式

閔晨、趙攀、余文婷 (2026)。〈有依據就可信嗎？探討引用信源對人工智慧生成內容 (AIGC) 可信度感知的影響〉。《傳播與社會學刊》，第78期，頁79–110。

Citation of This Article

Min, C., Zhao, P., & Yu, W. (2026). When transparency backfires: Source citations, confirmation bias, and the perceived credibility of AI-generated misinformation. *Communication and Society*, 78, 79–110.