A JOURNAL BY SISSA MEDIALAB

ARTICLE

Visible sources and invisible risks: exploring the impact of AI disclosure on perceived credibility of AI-generated content

Teng Lin and Yiqing Zhang

Abstract

With the widespread use of AI-generated content (AIGC) on social media, its potential to spread misinformation poses threats to the public. Although AI disclosure is widely promoted as a transparency measure to prompt critical evaluation, its effectiveness in science communication remains controversial. This study conducted a within-subjects experiment (N = 433) to examine how AI disclosure affects perceived credibility of science communication texts and the moderating roles of readers' negative attitudes towards AI and audience involvement. The experiment manipulated AI disclosure labels and information veracity. The results revealed a truth-falsity crossover effect: AI disclosure significantly reduced the perceived credibility of correct information while unexpectedly increasing the perceived credibility of misinformation. Negative attitudes towards AI significantly moderated these effects, whereas audience involvement exerted only limited influence. These findings highlight the complex and sometimes counterproductive consequences of AI disclosure in science communication and suggest implications for cue-based processing, algorithm aversion, and the design of disclosure mechanisms.

Keywords

Science and media; AI tools in science communication; Digital science communication

Received: 6th October 2025
Accepted: 7th January 2026
Published: 9th March 2026

1 ▪ Introduction

The widespread use of AI-generated content (AIGC), defined as content produced either independently or with the assistance of artificial intelligence [Huang & Wang, 2023], presents a significant challenge, as such content may contain highly persuasive misinformation that humans struggle to detect [Zhou et al., 2023; Shoaib et al., 2023]. This misinformation can be attributed to AI hallucinations — that is, content generated by a model that appears plausible but is inconsistent with facts due to limitations in training data or reasoning capabilities [Huang et al., 2025] — and to flawed or malicious prompts provided by users [Chen & Shu, 2024]. The proliferation of such convincing falsehoods poses risks to public decision-making and can erode trust in the information environment, thereby creating an urgent need for strategies to help audiences identify AI-generated misinformation [Weidinger et al., 2022]. In response, several countries have enacted laws requiring content creators to disclose AIGC. For example, China mandates that service providers add markings to content generated or synthesised by AI [Cyberspace Administration of China, 2025]. Similarly, The EU AI Act requires AI disclosure. Major social media platforms have implemented regulations to label such content [Clegg, 2024].

AI disclosure refers to the use of perceptible cues to inform audiences that AI is involved in the production of information, with the aim of helping them identify misinformation by encouraging critical evaluation of content [Wittenberg et al., 2024]. However, there is ongoing debate regarding whether AI disclosure achieves its intended effects, and its potential risks remain underexplored.

The literature on AI disclosure presents conflicting findings. Some studies suggest that AI disclosure indiscriminately reduces perceived credibility [Altay & Gilardi, 2024; Toff & Simon, 2025], whereas others find it to be effective or harmless [Liu et al., 2023; Kirkby et al., 2023]. Critically, these studies have largely overlooked the high-risk context of science communication, in which audiences must evaluate specialised knowledge but often lack the ability to verify it independently [Hodson et al., 2023], and where misinformation can pose safety risks [Denniss & Lindberg, 2025]. Furthermore, studies within science communication have focused on persuasiveness [Lim & Schmälzle, 2024; Beckmann et al., 2025; Reis et al., 2024], while neglecting credibility — the prerequisite for fostering public trust in scientific evidence [Intemann, 2023]. Addressing this gap, and recognizing that text generation represents a highly efficient yet challenging form of AIGC on social media [Chen & Shu, 2024], this study examines the impact of AI disclosure on the perceived credibility of science communication texts.

2 ▪ Literature review

2.1 ▪ *Existing research on AI disclosure*

AI disclosure is commonly implemented in the form of visible warning labels and is intended to encourage audiences to more carefully scrutinise the content [Wittenberg et al., 2024; Shoaib et al., 2023], but its effectiveness remains contested. On one hand, some research suggests it can be a useful evaluative cue. Studies in marketing and consumer contexts, for example, find that disclosure does not necessarily reduce authenticity and can even enhance credibility [Kirkby et al., 2023; Cheng et al., 2022]. Psychological studies similarly indicate

that AIGC labels can prompt audiences to make more cautious judgements about information quality [Liu et al., 2023].

Conversely, a significant body of research argues that AI disclosure often backfires. In journalism, it can paradoxically undermine public trust and reduce perceived accuracy [Toff & Simon, 2025; Altay & Gilardi, 2024]. This negative effect appears to be particularly pronounced in science communication, where disclosure has been shown to consistently lower audience evaluations on measures like perceived effectiveness, argument strength, and perceived quality across various science topics, including vaccines, vaping, and medical advice [Karinshak et al., 2023; Lim & Schmälzle, 2024; Beckmann et al., 2025; Reis et al., 2024]. Furthermore, AI disclosure may decrease audiences' perceptions of the reliability and persuasiveness of the content, their empathy and their willingness to follow health advice [Reis et al., 2024; Teigen et al., 2024]. While existing research shows that AI disclosure often dampens evaluations of scientific information, it has focused primarily on persuasive judgements rather than the more fundamental construct of perceived credibility.

In summary, three important gaps in the previous research remain. First, previous studies have focused on persuasive judgements while overlooking the more fundamental construct of credibility. Persuasion refers to changes in attitudes or behaviours, whereas credibility relates to whether information is perceived as reliable and trustworthy [Kumkale et al., 2010; Attaran et al., 2015]. This oversight is particularly problematic in science communication, where the primary goal is not merely persuasion but fostering rational trust, a process premised on the audience's recognition of credibility [Intemann, 2023]. Therefore, to understand the true impact of AI disclosure, research must examine credibility directly.

Second, although studies consistently show that AI disclosure in science communication lowers audiences' evaluations of information, the underlying mechanism remains unclear. The Elaboration Likelihood Model (ELM) proposes that audiences may process information either via the peripheral route — relying on external features or heuristic cues — or via the central route, which involves careful analysis of content and credibility [Petty & Cacioppo, 1986]. It remains uncertain whether the label acts as a simple negative heuristic, fostering generalised mistrust, or whether it prompts systematic processing that helps audiences better identify misinformation. This distinction is critical in science communication. If reduced evaluations result from heuristic scepticism, disclosure may foster generalised mistrust of all information, thereby undermining communication efforts [Hmielowski et al., 2014]. By contrast, if they result from careful verification, disclosure may help audiences identify misinformation without impeding the spread of correct information [Vraga & Bode, 2018]. Prior studies have not resolved this, as most studies have failed to differentiate disclosure's impact on correct information versus misinformation. Therefore, a critical gap remains in understanding how AI disclosure shapes perceived credibility across these two distinct information types.

Finally, most existing evidence is based on Western samples, overlooking how AI disclosure may function differently in Chinese social media environments, where audiences hold more favourable attitudes towards AI [Song, 2023; Kuai, 2025]. Addressing these gaps, this study examines how AI disclosure influences perceived credibility of both correct information and misinformation in science communication on Chinese social media, posing the following research questions.

RQ1: How does AI disclosure affect audiences' perceived credibility of AI-generated scientific information on Chinese social media, i.e., does its influence reflect heuristic scepticism towards all AIGC or careful verification that differentiates between misinformation and correct information?

The distinction between heuristic scepticism and careful verification can be examined by assessing the differential impact of disclosure on correct and incorrect information. Heuristic scepticism would decrease the credibility of both, whereas careful verification would selectively decrease credibility for misinformation while maintaining or increasing it for correct information. Given that the stated purpose of AI disclosure is to encourage more careful scrutiny through source cues [Shoaib et al., 2023], we propose the following hypotheses:

H1a: AI disclosure reduces audiences' perceived credibility of AI-generated misinformation in science communication texts.

H1b: AI disclosure increases audiences' perceived credibility of AI-generated correct information in science communication texts.

2.2 ▪ The negative attitudes towards AI and AI disclosure

The effect of AI disclosure may depend on the audience's prior knowledge, both of AI itself and the specific content domain [Vasse'i & Udoh, 2024; Toff & Simon, 2025]. For instance, familiarity with AI and AIGC can moderate the disclosure effect [Luo et al., 2019]. However, while the role of knowledge has been explored, considerably less is known about the role of pre-existing attitudes.

Cognitive Dissonance Theory suggests that when audiences' knowledge, attitudes, and behaviours are inconsistent, they are motivated to adjust their behaviour or knowledge [Cooper, 2019]. Thus, audiences with negative attitudes towards AI [Negative Attitudes Toward Artificial Intelligence or NATAI] (a construct capturing predispositional scepticism towards AI technologies) may perceive AIGC as untrustworthy, or alternatively, revise their attitudes towards AI when the information itself is perceived as credible. However, there is limited empirical evidence on this potential moderating effect. Therefore, we propose the following research question:

RQ2: To what extent does the effect of AI disclosure on audiences' perceived credibility of AIGC depend on their negative attitudes towards AI?

Building on research on algorithm aversion [Burton et al., 2020], we expect that pre-existing negative attitudes towards AI will shape audience responses to AI disclosure. The AI label may activate these negative biases, causing audiences to transfer their scepticism onto the content itself. Therefore, we propose the following hypotheses:

H2a: Negative attitudes towards AI negatively moderate the effect of AI disclosure on the perceived credibility of AI-generated misinformation.

H2b: Negative attitudes towards AI negatively moderate the effect of AI disclosure on the perceived credibility of AI-generated correct information.

2.3 ▪ *Involvement and the effectiveness of AI disclosure*

Beyond attitudes, audience involvement — defined as the perceived relevance of the information [Zaichkowsky, 1985] — may also moderate disclosure effects. Competing theoretical predictions create uncertainty regarding the role of audience involvement. The Elaboration Likelihood Model (ELM) suggests that low-involvement audiences tend to rely on peripheral cues and would thus be more influenced by the AI label [Petty & Cacioppo, 1986]. However, the concept of anchoring could also apply. Anchoring suggests that people's evaluations are biased by perceived technological credibility, which then serves as a reference point [Furnham & Boo, 2011]. The AI label, therefore, may function as a powerful initial reference point that biases evaluations, even for high-involvement audiences.

Given these competing predictions, audiences' involvement may shape the impact of AI disclosure. Examining this possibility is crucial for understanding the mechanism through which AI disclosure influences audience evaluations. Accordingly, we propose the following research question:

RQ3: To what extent does audiences' information involvement moderate the effect of AI disclosure on their perceived credibility of AIGC?

Based on the ELM, this study conceptualises the AI disclosure label as a peripheral cue. Its influence is expected to be weaker for more highly involved audiences who are likely to focus on the message content. Therefore, we propose the following hypotheses:

H3a: Involvement significantly negatively moderates the effect of AI disclosure on the perceived credibility of AI-generated misinformation.

H3b: Involvement significantly negatively moderates the effect of AI disclosure on the perceived credibility of AI-generated correct information.

Integrating the foregoing theories and research hypotheses, we propose a research model illustrating the effect of AI disclosure and its moderating factors, as shown in Figure 1.

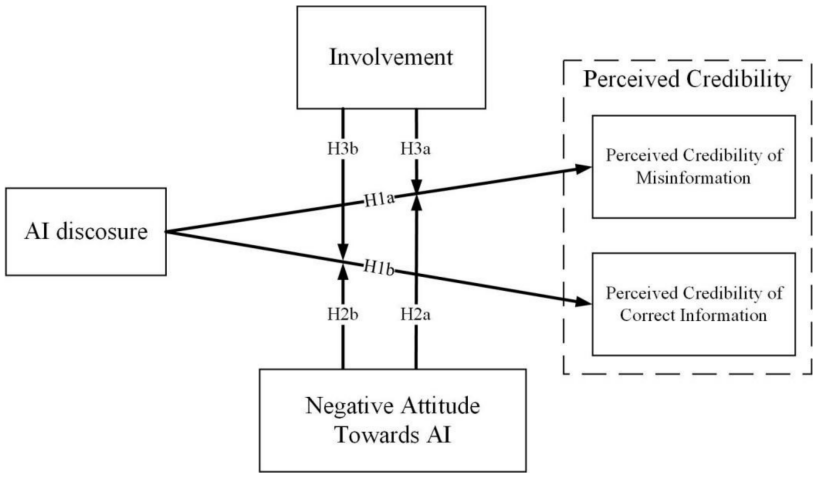

Figure 1. Research model and theoretical hypotheses.

3 ▪ Method

3.1 ▪ *Research design*

We employed a within-subjects design using an online survey to simulate the experience of browsing popular science content on social media. This approach was chosen to control for individual differences that strongly influence credibility judgements, such as attitudes and information literacy. By having each participant evaluate content under different conditions, this design enables a precise analysis of within-person variation in perceived credibility, while reducing between-person variability.

The independent variable in this study was AI disclosure. The dependent variables were the perceived credibility of correct information and misinformation, with negative attitudes towards AI and involvement serving as moderators. To manipulate the independent variable, we used GPT-4 to generate the experimental materials and randomly assigned the AI disclosure labels. Participants subsequently rated the credibility of the materials and reported on the moderating variables and demographics.

3.2 ▪ *Experimental materials and pretest*

3.2.1 ▪ *Source of materials*

Two representative topics were selected: food health and safety (Topic 1) and disease prevention and control (Topic 2). These topics were chosen for their high relevance and ecological validity within the Chinese social media context [Zhu et al., 2022], where they frequently feature both misinformation and official health policy [State Council of the People's Republic of China, 2016].

Experimental materials were drawn from the official *Science Rumour Debunking Platform* (https://piyao.kepuchina.cn/), operated by the China Association for Science and Technology, which publishes expert-reviewed monthly lists of debunked rumours. From January 2023 to March 2024, we identified 10 articles related to Topic 1 and 12 to Topic 2. Using simple random sampling, we selected 8 from each topic, which were then rewritten by AI to create the final stimuli.

3.2.2 ▪ *Rewriting process*

We used ChatGPT-4 to adapt the selected texts into Sina Weibo-style posts, generating both correct and misleading versions.

Correct information was generated using the prompt: "*Based on the typical style of popular science posts on Sina Weibo, retain the core scientific arguments of the following debunking article and rewrite it into a conversational science communication text*".

Misinformation was generated using the prompt: "*Based on rumours mentioned in the debunking article, simulate a pseudo-scientific popular science post in the style commonly found on Sina Weibo, ensuring that logical fallacies align with real misinformation*".

This process resulted in four accurate and four misleading posts (two of each per topic). All materials underwent manual verification: two researchers independently reviewed each post

to ensure factual accuracy in the case of correct information, or consistency with the original rumour in the case of misinformation, resolving any discrepancies through discussion.

3.2.3 ▪ *Pretest*

To control for potential confounds, we conducted a pretest (N = 50) on 16 draft stimuli. The pretest served two purposes: (1) to confirm that the two main topics were comparable in terms of audience familiarity and engagement, and (2) to select the final set of stimuli based on their comprehensibility and perceived credibility.

Paired-sample t-tests confirmed no significant differences in familiarity (t = -1.091, p = 0.281) or engagement (t = 0.590, p = 0.558) between the two topics, ensuring their comparability. We then filtered the materials using predefined criteria. Stimuli were retained if they met the following conditions: comprehensibility scores exceeded 3.0; perceived credibility for correct information exceeded 3.5; and the average credibility score for misleading information fell between 1.5 and 2.5. This process was designed to select correct information that was seen as credible and misinformation that was deceptive but not overly convincing, while ensuring all final texts were easy to understand.

This resulted in a final set of eight posts: four accurate and four misleading (two per topic). The specific topics included carcinogenic risks of food and food washing (Topic 1), and heart health and eye disease prevention (Topic 2).

3.2.4 ▪ *AI disclosure operationalization*

The independent variable, AI disclosure, was manipulated through a fully random assignment process implemented by the survey platform, such that each accurate and misleading post could appear either with or without an AIGC label. The label stated "*Attention: The content was detected as being generated by AI*" and appeared in red at the top of the text to ensure a clear distinction from the body content. Materials were formatted to resemble Sina Weibo posts; however, source-related cues were removed to avoid social endorsement effects. Specifically, user avatars, usernames, repost counts, likes, and comments were excluded. The final layout of the stimuli is shown in Figure 2 (participants viewed the Chinese version; an English version is provided for readers' reference).

3.3 ▪ *Sample*

Data for this study were collected via the Credamo platform between March and May 2024. All participants provided informed consent prior to commencing the survey, and the study was approved by the University's Academic Ethics Review Committee. To ensure data quality, we implemented an instructional manipulation check (e.g., "Please select option 1 for this item") and excluded responses with completion times shorter than one-third of the sample median. After applying these criteria, we retained 433 valid responses from an initial sample of 536 participants, yielding a retention rate of 80.8%.

Sample characteristics are presented in Table 1. To assess representativeness of the sample, we compared its demographic profile with that of Weibo users, the platform context of this study. Sina Weibo is one of the largest social media platforms in China, functionally similar to

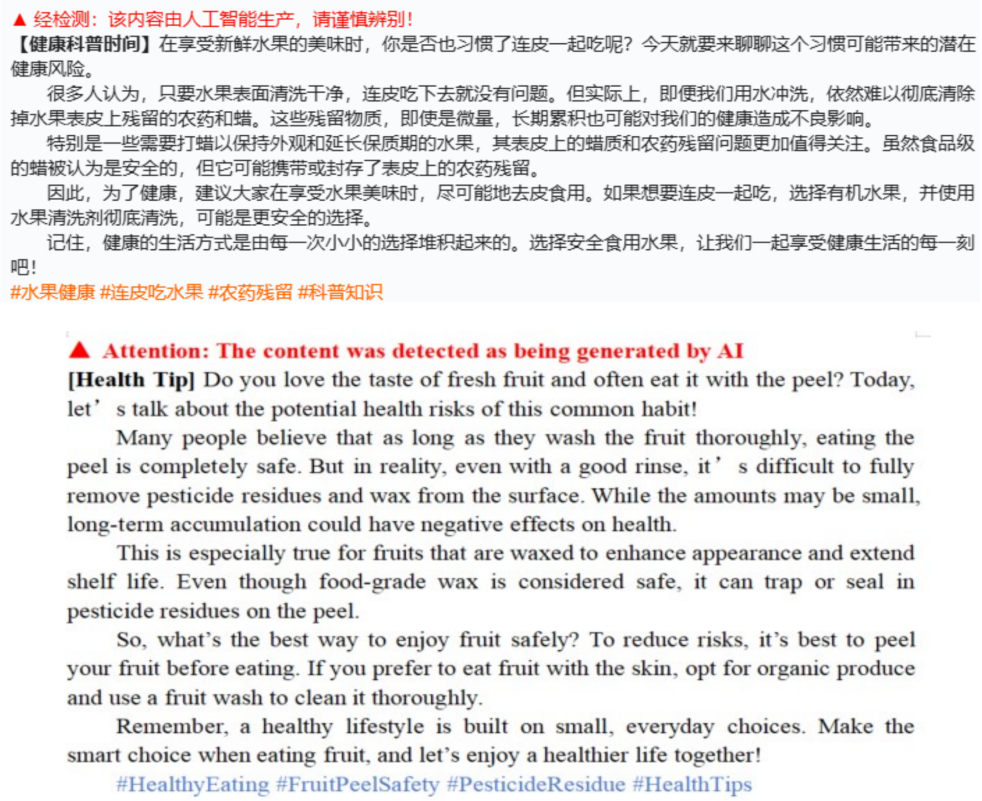

Figure 2. Sample of experimental materials received by subjects.

Table 1. Overall characteristics of samples.

Variable	*Category*	*N*	*Percentage*
Gender	Male	155	35.8
	Female	278	64.2
Age	under 20	17	3.9
	21–30	224	51.7
	31–40	140	32.3
	41–50	32	7.4
	Age 51 and older	20	4.6
Education	General high-school & below	22	5.1
	Associate degree	42	9.7
	Bachelor's degree	295	68.1
	Postgraduate degree	74	17.1

X (formerly Twitter). It is primarily text-based, making it a particularly relevant setting for studying the dissemination of AI-generated text. According to official statistics, Weibo users are 55% female, with the largest age segment being 21–30 (48%), followed by 31–40 (29%) [Sina Weibo, 2021].

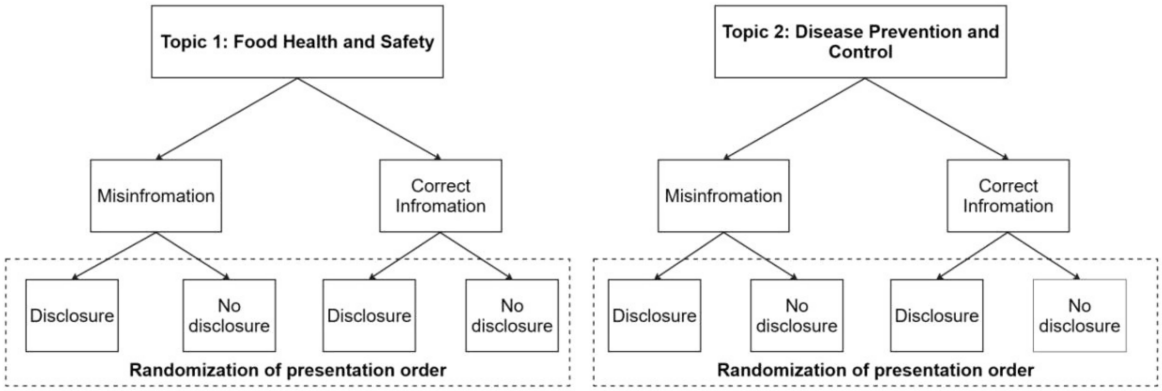

Figure 3. Experimental materials to which subjects were exposed.

3.4 ▪ *Procedure*

Before starting, participants were informed they would read and evaluate eight popular science posts across two topics, some correct and some incorrect, using a 5-point Likert scale.

The procedure was identical for both topic blocks. Participants first reported their involvement with the upcoming topic. They were then presented with four related posts in a randomised order to minimise order effects. After reading each post, they immediately rated its perceived credibility before proceeding. In total, each participant who completed the study viewed and evaluated eight posts, as illustrated in Figure 3.

After completing all materials, participants reported their negative attitudes towards AI and provided demographic information. At the end of the study, they were debriefed and informed which posts contained incorrect information.

3.5 ▪ *Measures*

AI disclosure was the independent variable, comprising two levels: disclosed and undisclosed. Each condition contained both correct information and misinformation.

Perceived credibility was the dependent variable, which was measured separately for correct information and misinformation. Following previous research [Wölker & Powell, 2021], participants rated the credibility of each post on a 5-point Likert scale (1 = not at all reliable, 5 = extremely reliable). To reduce participant fatigue and to reflect the rapid-judgement context of social media, we measured perceived credibility using a single item per post. This is appropriate because perceived credibility meets the criteria of a "doubly concrete construct" — both the object (specific text material) and the attribute (credibility) are concrete, singular, and easily conceptualised [Bergkvist & Rossiter, 2007].

Negative attitudes towards AI served as a moderating variable and were measured with the General Attitudes towards Artificial Intelligence Scale (GAAIS) [Schepman & Rodway, 2023]. Since the study focuses on negative attitudes towards AI, we used only the Negative GAAIS subscale. Participants rated a series of statements on a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree). Items covered AI's practicality, ethical

concerns, error rates, and potential social impact. To ensure accuracy, we translated the scale into Chinese. Two university students independently back-translated into English, and comparisons with the original scale revealed no ambiguities. The eight-item subscale demonstrated good internal consistency, indicating acceptable reliability (Cronbach's $\alpha = 0.789$). An exploratory factor analysis (EFA) using principal axis factoring with varimax rotation supported structural validity: KMO = 0.733, $p < .001$; two common factors with eigenvalues > 1 explained 38.84% of total variance; factor loadings ranged from 0.44 to 0.83 with no substantial cross-loadings.

Involvement also served as a moderating variable and was measured using four items adapted from the revised Personal Involvement Scale [Zaichkowsky, 1994]. Items assessed perceived importance, attention, value, and prior knowledge, rated on a 5-point Likert scale (1 = not at all, 5 = very much). The scale demonstrated good reliability (Cronbach's $\alpha = 0.771$). EFA supported structural validity: KMO = 0.776, $p < .001$; two factors with eigenvalues > 1 explained 42.75% of total variance; factor loadings ranged from 0.38 to 0.63, with no substantial cross-loadings.

Demographic control variables included gender, age, and education level.

4 ▪ Result

4.1 ▪ *Main effect: impact of AI disclosure on perceived credibility*

To test whether AI disclosure reduces the perceived credibility of misinformation (H1a) and enhances that of correct information (H1b), we conducted a linear mixed-effects model (LMM) in SPSS 18.0 to account for the repeated-measures design. We specified AI disclosure and text type as fixed factors and perceived credibility as the dependent variable. Covariates included topic, age, education, and gender. To account for individual differences in baseline credibility, we specified random intercepts for each participant. The results of this analysis, including tests of fixed effects and parameter estimates for each predictor, are presented in Tables 2 and 3.

The analysis revealed a significant interaction effect between AI disclosure and information type ($F = 213.91$, $p < .001$), indicating that the effect of disclosure depended on whether the

Table 2. Type III tests of fixed effects.

Effect source	df_1	df_2	F
Intercept	1	429.144	536.892***
AI disclosure	1	3023.184	63.912***
Type	1	3022.685	141.046***
AI disclosure*Type	1	3022.685	213.91***
Topic	1	3023.871	105.079***
Age	1	429.020	1.184
Education	1	429.495	2.305
Gender	1	429.102	3.808

Note. Results from the linear mixed-effects model.
* $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3. Estimates of fixed effects.

Effect	β	*SE*	*t*	*df*	*95% CI*
Intercept	4.030***	0.166	24.221	461.497	[3.703, 4.357]
AI disclosure (No vs Yes)	0.209***	0.045	4.688	3022.785	[0.122, 0.297]
Type (Misinformation vs Correct)	0.087	0.045	1.944	3022.537	[-0.001, 0.174]
AI disclosure × Type	-0.922***	0.063	-14.626	3022.685	[-1.046, -0.799]
Topic	-0.323***	0.032	-10.251	3023.871	[-0.385, -0.261]
Age	0.031	0.029	1.088	429.020	[-0.025, 0.088]
Education	-0.055	0.036	-1.518	429.495	[-0.126, 0.016]
Gender	0.100	0.051	1.952	429.102	[-0.001, 0.201]

Note. Results from the linear mixed-effects model. * $p < .05$, ** $p < .01$, *** $p < .001$.

information was correct or not. The estimates of fixed effects are detailed in Table 3, where correct information with AI disclosure specified as the reference group. Relative to this reference group, correct information without disclosure was perceived as significantly more credible ($\beta = 0.209$, $p < .001$). This finding indicates that AI disclosure reduced the credibility of correct information, leading to the rejection of H1b. For misinformation, the pattern of coefficients indicated higher perceived credibility when an AI disclosure label was present than when it was absent, relative to the reference condition ($\beta = -0.922$, $p < .001$), a finding that contradicts H1a.

In summary, these findings address RQ1 by showing a crossover interaction contrary to our hypotheses: AI disclosure reduced the perceived credibility of correct information while increasing the perceived credibility of misinformation. Demographic covariates were not statistically significant.

4.2 ▪ *Moderating effects: negative attitudes towards AI and involvement*

This section reports the analyses conducted to test whether negative attitudes towards AI (H2) and audience involvement (H3) moderate the effect of AI disclosure on perceived credibility. To analyse the within-subject repeated-measures design, we used Model 2 of the MEMORE macro (version 2.1) for SPSS [Montoya, 2019]. In the model, negative attitudes towards AI and involvement were specified as moderators. The dependent variable was defined as the difference in perceived credibility between AI-disclosed and non-disclosed information within the same topic and information type. Accordingly, we conducted four tests.

To evaluate the significance of the interaction effects, we applied a non-parametric bootstrapping procedure with 5,000 resamples to construct 95% confidence intervals. To reduce potential multicollinearity, all variables involved in the interaction terms were mean-centred before analysis. The results are reported in Tables 4 and 5. Table 4 shows the conditions under which the two moderating variables exhibit moderating effects in which directions. Table 5 presents the simple slope results, which reflect the conditional effects of AI disclosure when negative attitudes towards AI are at three levels (the mean and ± 1 SD). Figures 4 and 5 are visualisations of Table 5.

The moderating effect of negative attitudes towards AI on the perceived credibility of misinformation was topic-dependent. This moderating effect was significant and negative for

Table 4. Test results of moderating variables.

Dependent variable	*Topic*	*Moderator variable*	β	*SE*	*t*	R^2
PCM	T1	NATAI	-0.523***	0.010	-5.256	0.25
		Involvement	-0.237	0.125	-0.190	
	T2	NATAI	-0.052	0.074	-0.706	0.13
		Involvement	0.213*	0.093	2.286	
PCCI	T1	NATAI	-0.205*	0.092	-2.221	0.11
		Involvement	-0.135	0.116	-1.165	
	T2	NATAI	-0.175*	0.078	-2.251	0.11
		Involvement	-0.041	0.097	-0.418	

Note. Results from the model. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 5. Simple slope results for negative attitudes towards AI.

Dependent variable	*Topic*	*NATAI*	*t*	*Effect*
PCM	T1	1.767	14.8558	1.417***
		2.470	15.6188	1.053***
		3.174	7.2197	0.689***
PCCI	T1	1.767	-2.1079	-0.187*
		2.470	-4.9708	-0.312***
		3.174	-4.9178	-0.437***
	T2	1.767	0.1773	0.013
		2.470	-1.9757	-0.104*
		3.174	-2.9697	-0.221*

Note. Results from the model. * $p < .05$, ** $p < .01$, *** $p < .001$.

Topic 1 ($\beta = -0.523$, $p < .001$) but not for Topic 2 ($\beta = -0.052$, $p > .05$), providing only partial support for H2a. As shown for Topic 1 in Table 5 and Figure 4, although AI disclosure increased the perceived credibility of misinformation across levels of negative attitudes, this effect was significantly weaker among participants with stronger negative attitudes towards AI.

Supporting H2b, negative attitudes towards AI significantly moderated the effect of disclosure on the perceived credibility of correct information. This negative moderation was significant for both Topic 1 ($\beta = -0.205$, $p < .05$) and Topic 2 ($\beta = -0.175$, $p < .05$). As shown in Table 5 and Figure 5, stronger negative attitudes consistently amplified the credibility penalty imposed by disclosure on correct information. Simple effects analyses showed this penalty was significant at medium and high levels of negative attitudes across both topics. However, at low levels of negative attitude, the penalty reached significant only for Topic 1, suggesting the effect is topic-dependent among audiences who are less sceptical of AI.

In response to RQ2, these findings confirm that negative attitudes towards AI moderate the effect of disclosure, but this moderation is complex and topic-dependent. For misinformation, the moderating effect emerged in only one of the two topics. For correct information, negative attitudes consistently amplified the negative impact of disclosure on correct information, with some topic-dependence at lower levels of negative attitudes.

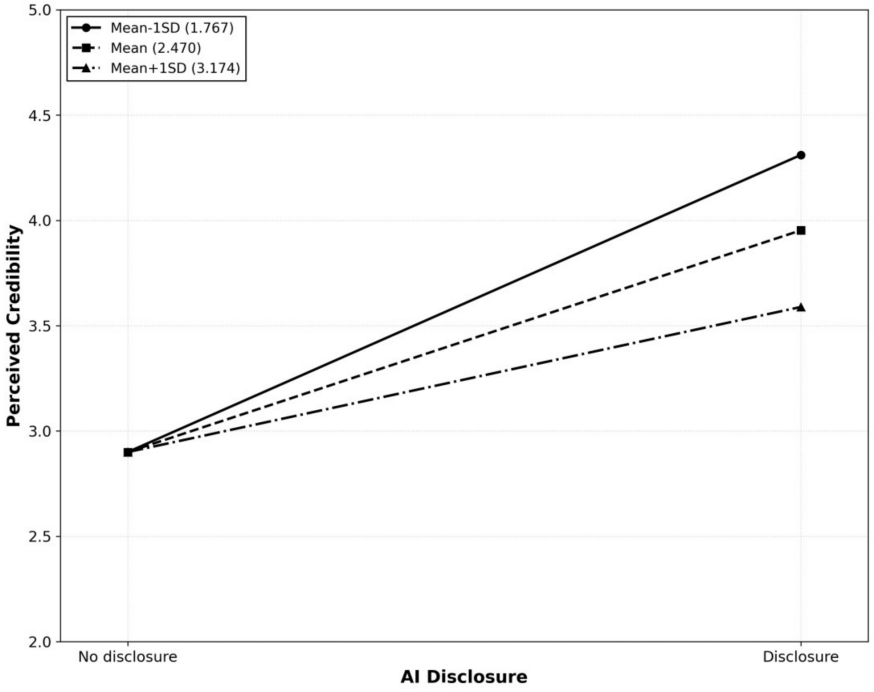

Figure 4. Moderating effect of negative attitudes towards AI on perceived credibility of misinformation.

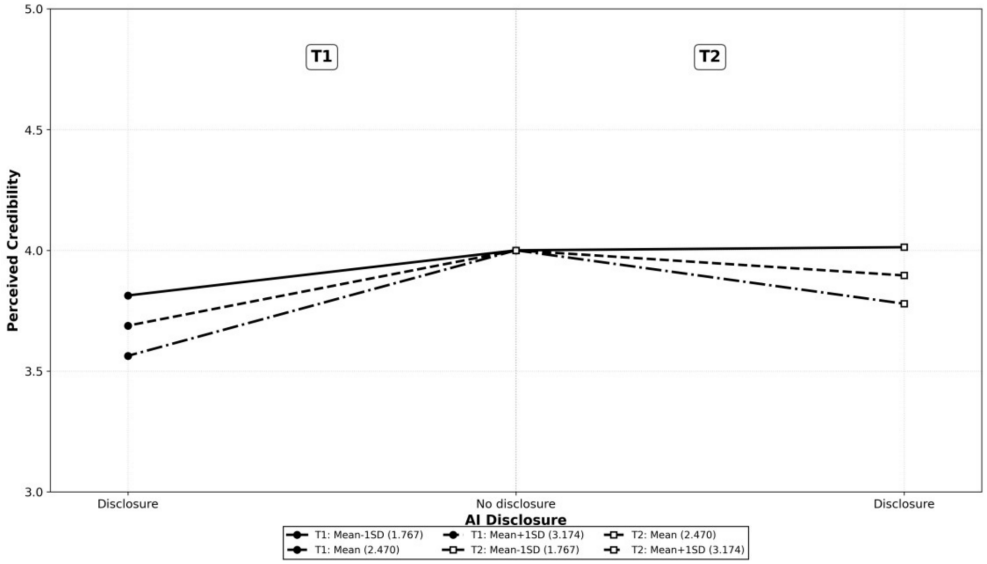

Figure 5. Moderating effect of negative attitudes towards AI on perceived credibility of correct information.

Finally, audience involvement demonstrated only limited and topic-dependent moderation. H3a was not supported, as the moderating effect on misinformation was non-significant in Topic 1 ($\beta = -0.237$, $p > .05$) and significant but in the opposite direction to that hypothesised in Topic 2 ($\beta = 0.213$, $p < .05$). H3b was also rejected, as involvement did not significantly moderate the effect on correct information in either topic.

4.3 ▪ *Robustness checks*

To further assess the robustness of the main findings, we conducted two robustness checks. First, re-estimating the main model with experimental materials as a random effect showed that material-specific characteristics did not significantly influence credibility assessments (variance = 0.042, $p = .246$). Second, adding participants' prior knowledge as a covariate did not alter the results; the disclosure-by-type interaction remained highly significant ($F(1,3027) = 215.785$, $p < .001$), while prior knowledge itself was not a significant predictor ($p = .061$). These checks provide additional evidence that the observed interaction reflects a stable pattern rather than an artefact of the specific materials or participants' prior knowledge, strengthening confidence in our findings.

5 ▪ Discussion

5.1 ▪ *AI disclosure and the crossover effect: misinformation gains and correct information diminishes in credibility*

This study's main finding — that AI disclosure increased the perceived credibility of misinformation while diminishing it for correct information — diverges from the prevailing view in the literature. Prior research, particularly in high-stakes domains like health and news, has largely documented that disclosure has a uniformly negative impact, reducing audience ratings of perceived effectiveness, argument strength, and information quality [Karinshak et al., 2023; Lim & Schmälzle, 2024; Beckmann et al., 2025]. Similar effects have been documented in finance, politics, digital health advice, and prostate cancer information [Reis et al., 2024; Teigen et al., 2024; Hershenhouse et al., 2025]. Findings in lower-stakes contexts are more mixed. When audiences perceive AI as approachable, disclosure may even enhance the perceived credibility of information [Cheng et al., 2022]. However, when content sources mix human and AI input, disclosure can still elicit distrust [Jakesch et al., 2019]. Conversely, our identification of a credibility redistribution, rather than a simple penalty, offers novel evidence for science communication.

In response to RQ1, our study identified a truth-falsity crossover effect. To our knowledge, it is a pattern of credibility redistribution not previously documented in the literature. Rather than eliciting simple heuristic scepticism or systematic verification, AI disclosure in science communication appears to function as an interpretive cue that reallocates credibility — enhancing it for false claims while diminishing it for true ones. Framed within the Elaboration Likelihood Model (ELM), this finding is consistent with a two-stage processing account [Petty & Cacioppo, 1986]. At an initial stage, the AI label may act as a positive heuristic, reducing cognitive scrutiny and allowing misinformation to pass credibility checks. Subsequently, audiences with sufficient motivation may engage in more systematic evaluations of higher-order attributes like explanatory quality and contextual relevance, which ultimately shapes their final judgement and explains the credibility redistribution [Fähnrich et al., 2023].

This credibility reallocation can be further interpreted through Sundar's [2008] MAIN model, which posits that technological affordances activate mental heuristics that shape users' credibility evaluations. One such affordance is agency. Specifically, human agency activates the authority heuristic, whereby users place trust in information provided by experts, authoritative institutions, or celebrities. Machine agency activates the machine heuristic, in

which users perceive information as more objective, accurate, and unbiased. In this case, the AI disclosure signals machine agency, activating a heuristic that AI is objective and unbiased, which may enhance the credibility of seemingly factual misinformation. Simultaneously, it weakens the human agency cues of expertise and accountability, which can devalue the nuanced explanatory quality expected from correct scientific information.

Beyond this, stereotypes about AI may further account for the devaluation of correct information, as explained by the Stereotype Content Model (SCM). The SCM posits that people's stereotypes organised along two primary dimensions: warmth and competence. AI disclosure can activate the common stereotype of AI as highly competent but low in warmth — cold, mechanical, and lacking empathy [McKee et al., 2023]. This perception is particularly salient in the context of science communication [Wang et al., 2025], and it can negatively bias an audiences' assessment of its explanatory credibility [Li et al., 2025, 2024], acting as a filter even when they are processing the information carefully.

The results of this study diverge from those reported in Western contexts, which may be linked to cross-cultural differences in public cognition and attitudes towards AI. In China, a long-standing discourse of technological optimism [Richter et al., 2025], together with official narratives that portray AI as a symbol of national progress and development, has fostered greater acceptance of and more positive attitudes towards AI [Song, 2023; Kuai, 2025]. Such conditions may reinforce the heuristic of AI as objective and advanced in the initial stage of information processing. At the same time, the prevalent stereotype of AI as cold, mechanical, and lacking warmth also resonates strongly within Chinese society [Li et al., 2025], which can lead audiences to downgrade their evaluations of the explanatory quality of correct information.

This study advances understanding of AI disclosure in three ways. First, it demonstrates that AI disclosure in science communication produces a crossover effect contingent on information veracity, providing different empirical evidence for the inherent assumption that people have widespread doubts about AI. Second, by integrating information-processing models such as the ELM and MAIN with a social-cognitive perspective, it shows how cue reweighting and stereotypes interact to redistribute credibility. Finally, it underscores the role of cultural context and boundary conditions in shaping the effects of AI disclosure.

5.2 ▪ *Attitude-driven credibility: the asymmetric effects of negative attitudes towards AI and the extension of algorithm aversion*

This study highlights the asymmetric effects of negative attitudes towards AI. While stronger negative attitudes amplified the credibility penalty for correct information, they only attenuated — but did not eliminate — the credibility-enhancing effect of disclosure for misinformation. These findings contribute to the literature on algorithm aversion by illuminating complex moderating mechanisms related to information accuracy. This theory emphasises humans' irrational rejection of algorithmic decisions and low tolerance for algorithmic errors [Dietvorst et al., 2015; Burton et al., 2020]. Thus, audiences with high negative attitudes towards AI should exhibit increased distrust towards AIGC following disclosure, regardless of its accuracy. However, our results suggest a more nuanced reality: even among audiences with strong negative attitudes towards AI, AI disclosure can still slightly enhance the perceived credibility of misinformation in some contexts — an effect that, while attenuated, did not disappear. At the same time, strong negative attitudes

intensified scepticism towards correct information. This suggests that algorithm aversion is not a uniform rejection of AI, but rather a task-dependent and asymmetric response, with its effect varying in strength and direction.

This asymmetry aligns with recent research on task specificity, which suggests that algorithm acceptance depends on the task's nature [Schaap et al., 2024]. In science communication, this may manifest as a dual-pathway process. For seemingly objective, data-based facts (even when false), AI's perceived competence as a rational tool may temporarily suppress aversion. However, when evaluating the interpretive and explanatory quality of correct information, negative attitudes may reinforce the 'high competence-low warmth' stereotype, activating distrust in the AI source and, by extension, the content itself.

5.3 ▪ *Practice optimisation and risk mitigation: dual-labelling mechanisms and tiered disclosure systems*

The findings suggest AI disclosure may create what may be described as a heuristic trust trap, whereby the label activates a schema of objectivity, leading audiences to assume that the information is factually sound by default, potentially diminishing the likelihood of thorough verification. This indicates a single disclosure label may be insufficient.
A promising direction for future research and policy design is to test whether a dual-label system — combining AI disclosure with a risk-verification cue — can better balance efficiency and safety in audiences' cognitive processing. For instance, alongside an AI disclosure, a risk warning label such as "*This information has not been independently verified; please interpret with caution*" may prompt audiences to consider potential risks and engage in more systematic evaluation. While the present study did not test such mechanisms, future work could assess whether dual-label systems provide more comprehensive cognitive scaffolding for audiences.

In a similar vein, the findings suggest another possible avenue for exploration: the development of graded and categorised disclosure systems, which could be evaluated for their effectiveness in accommodating varying levels of informational complexity and risk. In domains where information structures are relatively flat and the dimension of professionalism is singular, such as brand communication [Kirkby et al., 2023], existing single-label disclosure may be sufficient, as audiences can map AI's technical function directly onto credibility judgements.

By contrast, in domains requiring multidimensional verification and adherence to social or ethical standards, credibility may need to be conceptualised along two dimensions: factual credibility, driven by technical accuracy, and interpretative credibility, driven by explanatory quality and contextual relevance. Future research could examine whether distinct disclosure cues for these dimensions help mitigate the credibility risks associated with AI-generated scientific content.

6 ▪ Conclusion

This study examines how AI disclosure in science communication affects the perceived credibility of correct information and misinformation, with a particular focus on the moderating roles of audience attitudes and involvement. We found a truth-falsity crossover

effect: disclosure increases the perceived credibility of misinformation while reducing that of correct information. This effect is significantly moderated by negative attitudes towards AI, whereas audience involvement showed no consistent moderating effect.

This study contributes to the AI disclosure literature in several ways. First, it provides empirical evidence of a dynamic redistribution of credibility between correct and misinformation, extending information-processing models such as ELM and MAIN by highlighting cue reweighting in credibility judgements. Second, the asymmetric moderating role of negative attitudes towards AI advances research on algorithm aversion, suggesting that its impact may depend on task characteristics and the information type [Dietvorst et al., 2015]. Finally, this study offers practical implications by proposing dual-labelling and tiered disclosure systems to reduce misinformation risks while maintaining transparency [Wittenberg et al., 2024].

This study has several limitations. First, it focuses on text-based disclosure, while multimodal formats may shape credibility judgements differently [Unal et al., 2022]. Second, only two science-related topics were examined, limiting generalisability across domains. Third, the experimental design removed social cues such as "likes" to control for confounding variables, but this may have reduced ecological validity, as prior work shows that social endorsement cues influence credibility judgements [Borah & Xiao, 2018]. Finally, although robustness checks ruled out item-level variance or prior knowledge as alternative explanations, these analyses also have constraints: the relatively small number of stimuli limited the power to detect material effects, and prior knowledge was assessed with a single item. Future research should address these limitations by employing multimodal formats, a broader range of topics, naturalistic settings, and more comprehensive knowledge measures.

Future studies should systematically test AI disclosure across multimodal formats and diverse topics to examine whether the crossover effect generalises to other contexts. Mixed-methods approaches, combining real-world social media data with controlled experiments, may also provide richer insights into disclosure effects under naturalistic conditions. Finally, our exploratory interpretation that cultural narratives underlie the divergence between our findings and prior Western studies should be empirically examined. Comparative cross-cultural research could clarify whether the observed crossover effect reflects universal processing mechanisms or culture-specific credibility frameworks.

Acknowledgments

As the study investigates how AI disclosure affects perceived credibility of AI-generated content in science communication, ChatGPT-4 was used to generate the experimental materials.

Funding
The authors received no specific funding for this work.

References

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. https://doi.org/10.1093/pnasnexus/pgae403

Attaran, S., Notarantonio, E. M., & Quigley, C. J. (2015). Consumer perceptions of credibility and selling intent among advertisements, advertorials, and editorials: a persuasion knowledge model approach. *Journal of Promotion Management*, *21*(6), 703–720. https://doi.org/10.1080/10496491.2015.1088919

Beckmann, S. A., Link, E., & Bachl, M. (2025). "ChatGPT, is the influenza vaccination useful?" Comparing perceived argument strength and correctness of pro-vaccination-arguments from AI and medical experts. *JCOM*, *24*(02), A04. https://doi.org/10.22323/2.24020204

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184. https://doi.org/10.1509/jmkr.44.2.175

Borah, P., & Xiao, X. (2018). The importance of 'likes': the interplay of message framing, source, and social endorsement on credibility perceptions of health information on Facebook. *Journal of Health Communication*, *23*(4), 399–411. https://doi.org/10.1080/10810730.2018.1455770

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Chen, C., & Shu, K. (2024). Can LLM-generated misinformation be detected? *arXiv*. https://doi.org/10.48550/arXiv.2309.13788

Cheng, X., Bao, Y., Zarifis, A., Gong, W., & Mou, J. (2022). Exploring consumers' response to text-based chatbots in e-commerce: the moderating role of task complexity and chatbot disclosure. *Internet Research*, *32*(2), 496–517. https://doi.org/10.1108/intr-08-2020-0460

Clegg, N. (2024, February 6). Labeling AI-generated images on Facebook, Instagram and Threads. *Meta*. https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/

Cooper, J. (2019). Cognitive dissonance: where we've been and where we're going. *International Review of Social Psychology*, *32*(1), 7. https://doi.org/10.5334/irsp.277

Cyberspace Administration of China. (2025). Regulations on marking generative AI-generated and synthetic content. https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

Denniss, E., & Lindberg, R. (2025). Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International*, *40*(2), daaf023. https://doi.org/10.1093/heapro/daaf023

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Fähnrich, B., Weitkamp, E., & Kupper, J. F. (2023). Exploring 'quality' in science communication online: expert thoughts on how to assess and promote science communication quality in digital media contexts. *Public Understanding of Science*, *32*(5), 605–621. https://doi.org/10.1177/09636625221148054

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

Hershenhouse, J. S., Mokhtar, D., Eppler, M. B., Rodler, S., Storino Ramacciotti, L., Ganjavi, C., Hom, B., Davis, R. J., Tran, J., Russo, G. I., Cocci, A., Abreu, A., Gill, I., Desai, M., & Cacciamani, G. E. (2025). Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer and Prostatic Diseases*, *28*(2), 394–399. https://doi.org/10.1038/s41391-024-00826-y

Hmielowski, J. D., Feldman, L., Myers, T. A., Leiserowitz, A., & Maibach, E. (2014). An attack on science? Media use, trust in scientists, and perceptions of global warming. *Public Understanding of Science*, *23*(7), 866–883. https://doi.org/10.1177/0963662513480091

Hodson, J., Reid, D., Veletsianos, G., Houlden, S., & Thompson, C. (2023). Heuristic responses to pandemic uncertainty: practicable communication strategies of "reasoned transparency" to aid public reception of changing science. *Public Understanding of Science*, *32*(4), 428–441. https://doi.org/10.1177/09636625221135425

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication*, *73*(6), 552–562. https://doi.org/10.1093/joc/jqad024

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 42. https://doi.org/10.1145/3703155

Intemann, K. (2023). Science communication and public trust in science. *Interdisciplinary Science Reviews*, *48*(2), 350–365. https://doi.org/10.1080/03080188.2022.2152244

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: how the perception that profile text was written by AI affects trustworthiness. In S. Brewster, G. Fitzpatrick, A. Cox & V. Kostakos (Eds.), *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300469

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 116. https://doi.org/10.1145/3579592

Kirkby, A., Baumgarth, C., & Henseler, J. (2023). To disclose or not disclose, is no longer the question — effect of AI-disclosed brand voice on brand authenticity and attitude. *Journal of Product & Brand Management*, *32*(7), 1108–1122. https://doi.org/10.1108/jpbm-02-2022-3864

Kuai, J. (2025). Navigating the AI hype: Chinese journalists' algorithmic imaginaries and role perceptions in reporting emerging technologies. *Digital Journalism*. https://doi.org/10.1080/21670811.2025.2502851

Kumkale, G. T., Albarracín, D., & Seignourel, P. J. (2010). The effects of source credibility in the presence or absence of prior attitudes: implications for the design of persuasive communication campaigns. *Journal of Applied Social Psychology*, *40*(6), 1325–1356. https://doi.org/10.1111/j.1559-1816.2010.00620.x

Li, Y., Wu, B., Huang, Y., Liu, J., Wu, J., & Luan, S. (2025). Warmth, competence, and the determinants of trust in artificial intelligence: a cross-sectional survey from China. *International Journal of Human-Computer Interaction*, *41*(8), 5024–5038. https://doi.org/10.1080/10447318.2024.2356909

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology*, *15*, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100058. https://doi.org/10.1016/j.chbah.2024.100058

Liu, Y., Wang, S., & Yu, G. (2023). The nudging effect of AIGC labeling on users' perceptions of automated news: evidence from EEG. *Frontiers in Psychology*, *14*, 1277829. https://doi.org/10.3389/fpsyg.2023.1277829

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: machines vs. humans: the impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience*, *26*(8), 107256. https://doi.org/10.1016/j.isci.2023.107256

Montoya, A. K. (2019). Moderation analysis in two-instance repeated measures designs: probing methods and multiple moderator models. *Behavior Research Methods*, *51*(1), 61–82. https://doi.org/10.3758/s13428-018-1088-6

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 123–205, Vol. 19). Elsevier. https://doi.org/10.1016/s0065-2601(08)60214-2

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. https://doi.org/10.1038/s41591-024-03180-7

Richter, V., Katzenbach, C., & Zeng, J. (2025). Negotiating AI(s) futures: competing imaginaries of AI by stakeholders in the US, China, and Germany. *JCOM*, *24*(02), A08. https://doi.org/10.22323/2.24020208

Schaap, G., Bosse, T., & Hendriks Vettehen, P. (2024). The ABC of algorithmic aversion: not agent, but benefits and control determine the acceptance of automated decision-making. *AI & Society*, *39*(4), 1947–1960. https://doi.org/10.1007/s00146-023-01649-6

Schepman, A., & Rodway, P. (2023). The general attitudes towards artificial intelligence scale (GAAIS): confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction*, *39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *2023 International Conference on Computer and Applications (ICCA)*, 1–7. https://doi.org/10.1109/ICCA59364.2023.10401723

Sina Weibo. (2021, March 12). *2020 Weibo User Development Report*. https://weibo.com/1642909335/K5OFGDin3

Song, B. (2023). How Chinese philosophy impacts AI narratives and imagined AI futures. In S. Cave & K. Dihal (Eds.), *Imagining AI: how the world sees intelligent machines* (pp. 338–352). Oxford University Press. https://doi.org/10.1093/oso/9780192865366.003.0021

State Council of the People's Republic of China. (2016, October 25). Healthy China 2030 plan outline. https://www.gov.cn/zhengce/2016-10/25/content_5124174.htm

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.

Teigen, C., Madsen, J. K., George, N. L., & Yousefi, S. (2024). Persuasiveness of arguments with AI-source labels. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey & E. Hazeltine (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 4076–4083, Vol. 46). https://escholarship.org/uc/item/6t82g70v

Toff, B., & Simon, F. M. (2025). "Or they could just not use it?": the dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, *30*(4), 881–903. https://doi.org/10.1177/19401612241308697

Unal, M. E., Kovashka, A., Chung, W.-T., & Lin, Y.-R. (2022). Visual persuasion in COVID-19 social media content: a multi-modal characterization. In F. Laforest, R. Troncy, L. Médini & I. Herman (Eds.), *WWW '22: Companion Proceedings of the Web Conference 2022* (pp. 694–704). Association for Computing Machinery. https://doi.org/10.1145/3487553.3524647

Vasse'i, R. M., & Udoh, G. (2024). *In transparency we trust? Evaluating the effectiveness of watermarking and labeling AI-generated content*. Mozilla Foundation. https://foundation.mozilla.org/en/research/library/in-transparency-we-trust/research-report/

Vraga, E. K., & Bode, L. (2018). I do not believe you: how providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*(10), 1337–1353. https://doi.org/10.1080/1369118x.2017.1313883

Wang, B., Shibo, B. W., & Kafle, J. (2025). When ChatGPT speaks about health: examining perceptions of warmth and competence toward AI as a health information source. *Journal of Health Communication*, *30*(10–12), 285–295. https://doi.org/10.1080/10810730.2025.2540864

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness Accountability and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2024). Labeling AI-generated content: promises, perils, and future directions. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.0319e3a6

Wölker, A., & Powell, T. E. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, *22*(1), 86–103. https://doi.org/10.1177/1464884918757072

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research*, *12*(3), 341–352. https://doi.org/10.1086/208520

Zaichkowsky, J. L. (1994). The personal involvement inventory: reduction, revision, and application to advertising. *Journal of Advertising*, *23*(4), 59–70. https://doi.org/10.1080/00913367.1943.10673459

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: understanding AI-generated misinformation and evaluating algorithmic and human solutions. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson & M. L. Wilson (Eds.), *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581318

Zhu, M., Fu, S., Zheng, D., & Li, Y. (2022). Research on social media health rumors from the perspective of literature: characteristics, dissemination and governance. *Documentation, Information & Knowledge*, *39*(05), 131–143. https://doi.org/10.13366/j.dik.2022.05.131

About the authors

Teng Lin. *School of Journalism and Communication, University of Chinese Academy of Social Sciences. Beijing, China.*
Teng Lin is a Ph.D. candidate in Journalism and Communication at the University of Chinese Academy of Social Sciences. His research focuses on science communication, with particular emphasis on science communication mediated by artificial intelligence (AI).

linteng@ucass.edu.cn @atengcc5

Yiqing Zhang. *School of Journalism and Communication, University of Chinese Academy of Social Sciences. Beijing, China.*
Yiqing Zhang is a Master's student in Journalism and Communication at the University of Chinese Academy of Social Sciences. Her research interests primarily lie in human-machine communication.
zyqing0127@163.com

How to cite

Lin, T. and Zhang, Y. (2026). 'Visible sources and invisible risks: exploring the impact of AI disclosure on perceived credibility of AI-generated content'. *JCOM* 25(01), A09. https://doi.org/10.22323/3580202601070857O3.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it