The Disclosure Dilemma: How AI Attribution Affects Reactions to Public Health Messages

Jacob A. Long, Tabitha Oyewole*, Maryam Goli*, Jacqueline M. Keisler*, Saud Alyaqout*, Michael D. Rodgers*, and Arielle N'Diaye*

University of South Carolina

*: Authors contributed equally and are randomly ordered.

Contact: jacob.long@sc.edu, please check jacob-long.com for updated drafts.

Abstract

This experiment (N=1,500) examines how disclosing AI generation of public health message creation affects audience reactions. Results suggest a trade-off: up-front disclosure of AI usage significantly reduces message credibility and learning (17% less information retained) but preserves institutional credibility, but if usage of AI is revealed after the fact there is substantial damage to source trustworthiness and perceived transparency. For professional-quality content with no obvious deficits, audiences did not suspect AI involvement when undisclosed.

Keywords: generative AI, artificial intelligence, public health, health communication, credibility

The Disclosure Dilemma: How AI Attribution Affects Reactions to Public Health Messages

Artificial intelligence (AI) has entered the health communication landscape, offering new capabilities for message generation but raising questions about disclosure, trust, and message reception (Karinshak et al., 2023; Lim & Schmälzle, 2023). As health organizations increasingly experiment with AI tools for content development, they face difficult decisions about whether and how to disclose AI's role in message creation. The growing integration of AI systems in health messaging raises important questions about message credibility, source transparency, and audience reception (Karinshak et al., 2023).

Health communication campaigns traditionally face resource constraints in generating sufficient high-quality, tailored messages to maintain audience engagement (Lim & Schmälzle, 2023). The creation of persuasive health messages has historically been a labor-intensive process, often creating a bottleneck in campaign implementation (Schmälzle & Wilcox, 2022). AI technologies, particularly large language models (LLMs), offer potential solutions to these constraints by efficiently generating messages that can be reviewed and deployed by health professionals (Burke-Garcia & Soskin Hicks, 2024). However, the use of AI in health message generation introduces novel considerations about source disclosure and audience trust. Health information is uniquely sensitive, with credibility playing a role in message acceptance and behavioral impact (Nan et al., 2023). The source of health information significantly influences how audiences evaluate message credibility, with traditional authorities like the CDC and healthcare professionals typically enjoying higher levels of trust than newer or less familiar sources (Karinshak et al., 2023; Wasike, 2022).

Recent research shows a fairly consistent pattern with regard to audience reactions to AI-generated health messages. Although AI systems can produce health messages that are

technically accurate and linguistically appropriate, audience perceptions may be influenced by knowledge of the source (Lim & Schmälzle, 2024). Experimental evidence suggests that when identical health content is attributed to different sources (human versus AI), audience evaluations differ, even when the actual quality of information remains constant (Karinshak et al., 2023; Lim & Schmälzle, 2024). Surveys indicate that many individuals express concerns about AI's role in healthcare decision-making, with general preferences for human involvement in health communication (Monteith et al., 2024). There may be some instances in which interventions can increase trust in AI-generated health information (Isaac et al., 2024), but it is not clear how context-bound such findings are.

To address these questions, we conducted a randomized experiment with 1,500 participants examining how AI disclosure affects perceptions of public health messages on social media. Participants were randomly assigned to one of five conditions: no AI disclosure, explicit denial of AI usage, disclosure of AI editing, disclosure of AI generation, or late disclosure of AI generation. In each condition, participants viewed four public health messages on different, relatively low-salience topics (antibiotic use, strep throat detection, melanoma prevention, and hearing loss) and evaluated message credibility, source credibility (trustworthiness and expertise), and organizational transparency.

Our findings demonstrate that AI disclosure significantly reduces perceptions of message credibility and both dimensions of source credibility (trustworthiness and expertise). The timing of disclosure also matters significantly—learning that the message was AI-generated but that the source had not initially disclosed it caused greater harm to source credibility than early disclosure, and was also associated with lower perceived transparency. On the other hand, relative to late disclosure, no mention of AI use, and active denial of AI use, early disclosure

harmed learning the information in the health messages. These results offer new empirical evidence on the tradeoffs involved in both using and being transparent about AI usage in health communication.

Source Credibility and AI-Generated Content

The integration of artificial intelligence (AI) into health communication presents opportunities for personalized information delivery and misinformation management along with increased efficiency for communicators. However, research consistently demonstrates that AI's effectiveness in health contexts depends on its perceived credibility, which it does not consistently have. Experimental evidence suggests that when audiences know content is AI-generated, their credibility assessments often differ from those applied to human-generated content. Lim and Schmälzle (2024) found that disclosing AI authorship significantly impacted audience evaluations of prevention messages, with human-generated messages generally receiving more favorable evaluations than AI-generated counterparts. A common pattern in research is that the content of AI-generated messages is not off-putting and may even be preferred; instead, it is the knowledge of AI involvement that harms perceptions (Karinshak et al., 2023). This AI aversion may be pronounced in domains like health, where personal stakes are high.

The concept of the "machine heuristic" may help explain these reactions to AI (Sundar & Kim, 2019). When people encounter AI systems, they apply preexisting beliefs about machines—that they lack human judgment, empathy, or moral agency—which can undermine perceived credibility regardless of content quality (Jia et al., 2024). This effect is particularly relevant in health contexts, where communications often require empathetic understanding and careful judgment that many believe only humans can provide (Isaac et al., 2024; Monteith et al.,

2024). A potential upside of the machine heuristic is that on issues for which there is public controversy, AI might be seen as less biased (Chung et al., 2023). Supporting the machine heuristic explanation is the finding that endorsement of the heuristic predicts reactions to AI disclosure (Wischnewski & Krämer, 2022).

Transparency and Disclosure Effects

Given the skepticism from ordinary people, transparency in AI-generated health content may pose an ethical challenge to health communicators. Although transparency is a generally recommended approach for anything that may concern one's audience, research shows that disclosure of AI use may significantly undermine the message and its sponsor. For instance, past research finds that news articles labeled as AI-generated are perceived as significantly less trustworthy than identical unlabeled content, though these articles were not evaluated as less accurate or fair (Toff & Simon, 2024). Related work has found loss of trust towards the source of AI-generated content even when the evaluation of the content itself was not harmed by disclosure (Rae, 2024). This loss of credibility at the source level could be particularly problematic in the health context where there are a finite set of public health organizations who try to leverage their credibility and reach to promote health behaviors.

Based on this past research regarding source credibility and audience perceptions of AI-generated content, we propose the following hypotheses:

H1: Disclosure of AI usage will reduce perceptions of message credibility.

H2a: Disclosure of AI usage will reduce perceptions of source trustworthiness.

H2b: Disclosure of AI usage will reduce perceptions of source expertise.

Altay and Gilardi (2024) observed that labeling content as AI-generated consistently decreased audience trust and willingness to share that content—even when the information was

factually accurate or actually human-authored but mislabeled. Importantly, they identified that this negative effect stems largely from audience assumptions about what AI involvement means, with most assuming complete automation rather than human-AI collaboration. When provided with definitions clarifying limited AI involvement (such as improving clarity or helping draft content), negative effects disappeared. Providing explanations such as the AI relying on human expert knowledge can also mitigate these downsides (Pareek et al., 2024). This suggests that how AI involvement is framed and contextualized may meaningfully alter credibility perceptions. The extent to which doing so is practical in real-world contexts is unclear, however, given how much audience attention is already at a premium.

Since there are potential payoffs to more nuanced messaging regarding the extent of AI use, we propose a research question about whether describing content as “edited” by AI may affect audience reactions in comparison to the more common “generated” phrasing. The advantage to a subtle variation like this is that it is realistic compared to some past research interventions, which are comparatively heavy-handed in trying to explain exactly how AI is used. Whether a lighter touch has an effect will be useful information to professional communicators.

RQ1: Does the type of disclosure ("generated" vs. "edited") affect credibility perceptions?

Disclosure Timing, Format, and Transparency Perceptions

The timing of disclosure warrants careful consideration. Early disclosure, provided before or at the same time as audiences encounter the message, proactively establishes transparency of the source. This may be particularly helpful when organizational reputation is already in question (Auger, 2014). If there is a mild taboo against AI usage, such a disclosure is a costly signal and could therefore show uncommon commitment to transparency. On the other hand, organizations

may not want to disclose or obfuscate the disclosure to mitigate audience biases against AI (Luo et al., 2019). The rationale is that audiences may evaluate the message itself more favorably if their judgment is not clouded by preconceived notions about AI's capabilities or trustworthiness (Jakesch et al., 2019). Communicators may decide it best not to attribute their messages to AI as long as the message is of the same quality as would be expected of a human. Such decisions would be especially unsurprising if messages are created with some human involvement, perhaps leading to a sense that to label them "AI-generated" is itself not completely accurate.

The gamble with non-disclosure is that undisclosed AI involvement may eventually be uncovered independently – whether through increasingly sophisticated detection methods, whistleblowers, or even casual user observation. This is particularly salient in high-stakes domains like health communication. This is what Toff and Simon (2024) call the "dilemma" of AI disclosure. Although transparency is normatively desirable and builds trust in principle, increasing evidence shows that audiences negatively react to AI-generated content (Wang & Huang, 2024). On the other hand, failing to disclose and risking later exposure could conceivably carry even greater credibility costs. The perception of intentional concealment, especially if revealed by a third party, could damage organizational reputation, outweighing any potential short-term gains from initially withholding disclosure, but data is lacking on this scenario. In essence, while early disclosure might present immediate credibility headwinds, the risk of *late* disclosure triggering accusations of deception and undermining long-term trust is a potentially greater threat. Furthermore, it is possible that proactive disclosure enhances the perceived transparency of the communicator, which might be a worthwhile tradeoff even if other perceptions are harmed.

With this in mind, we pose the following hypotheses and research questions:

H3a: Early disclosure of AI usage will increase perceptions of organizational transparency.

H3b: Late disclosure of AI usage will harm perceptions of organizational transparency.

RQ2: Does late disclosure incur a greater penalty for source credibility compared to early disclosure?

Processing AI-Attributed Messages

Beyond source credibility, understanding the psychological mechanisms governing audience responses to AI-generated health communication is important to predict other communication outcomes, like knowledge acquisition and information seeking, which are common goals of health communication campaigns. The disclosure of AI involvement may change how people process and engage with health information, influencing not just whether they believe it, but also whether they learn from it and seek further information.

Applying expectancy violation theory to this setting would suggest that AI disclosure might itself violate audience expectations, influencing cognitive engagement and subsequent knowledge acquisition (Burgoon et al., 1989). This lower expectancy could, in turn, reduce motivation to deeply process the message content, potentially hindering knowledge gain and diminishing subsequent information seeking. On the other hand, a high quality message attributed to AI might be received as a positive violation with regard to the technology with concomitant positive outcomes (Burgoon et al., 2016; Hong et al., 2021). In this way, there might be closer attention to the particulars of the message and therefore better learning. If users are primarily focused on identifying flaws or biases (Ou et al., 2024), this critical processing may detract from actual learning and retention of the health information itself. Alternatively, if heightened scrutiny leads to deeper engagement with the message content to assess its validity, and if the AI-generated content is of sufficient quality, the audience may learn more. Similarly, the perceived absence of human intent could influence information seeking. If users perceive AI

as a neutral information provider devoid of persuasive intent, they may be more inclined to seek further information to form their own informed opinions. Conversely, if users distrust AI or perceive it as manipulative, disclosure might decrease information seeking as they dismiss the source entirely.

Revisiting the machine heuristic (Yang & Sundar, 2024), if users apply an algorithmic aversion heuristic upon AI disclosure, they might disengage from the content, assuming it lacks the competence necessary for health contexts. This aversion could manifest as reduced cognitive effort devoted to processing the message, limiting knowledge gain. Furthermore, aversion may decrease motivation for further information seeking, as the source (AI) is already perceived as untrustworthy or unhelpful. If appreciation heuristics are instead activated – where AI is perceived as objective and efficient – disclosure might not impede knowledge gain or information seeking, or may even enhance it if users believe AI offers superior access to high-quality and unbiased information.

Besides imparting knowledge directly, a common goal of public health messages are calls for the audience to get more information from a trusted source, like the sponsor's website or a physician. Concerns about AI's ability to understand individual circumstances can generate hesitancy toward acting on AI-generated health advice, particularly for critical healthcare decisions (Longoni et al., 2019). The perceived lack of "human touch" in AI-generated messages may diminish their motivational impact, affecting adherence to recommended health behaviors. Despite these potential barriers, research shows mixed patterns of engagement with AI-generated content. However, in emotionally sensitive contexts, AI's perceived lack of empathy presents a significant drawback (Nadarzynski et al., 2021). Authenticity is crucial for online users, and research shows that AI-authored emotional communication may reduce positive word-of-mouth

and erode user loyalty (Kirk & Givi, 2025). This suggests that engagement with AI-generated health messages may be qualitatively different and potentially lower than engagement with human-created counterparts, influencing both knowledge acquisition and subsequent information-seeking behavior.

These several pathways through which message processing can be changed by AI disclosure directly raise questions about the impact on cognitive communication outcomes. These considerations lead to the following research questions:

RQ3: Does disclosure of AI usage affect learning from messages?

RQ4: Does disclosure of AI usage affect knowledge seeking from messages?

Denial vs. Non-disclosure

The so-called replicant effect (Jakesch et al., 2019) describes a phenomenon where people exhibit reduced trust toward content they believe may have been generated by AI in mixed-source environments. The name references the science fiction film *Blade Runner*, where "replicants" are synthetic beings that imitate humans. The insight is that as more media content becomes AI-generated, audiences may increasingly be skeptical of anything that is not affirmatively identified as the work of humans only. In the original research, this effect is especially pronounced in environments where both human and AI-created content coexist in a way that is readily disclosed to participants. With the surge in recent innovation and diffusion of AI products, it may be the case that media consumers are beginning to have the mindset that any content they encounter is possibly created by AI.

From a strategic communications perspective, organizations may consider explicitly denying AI involvement as a tactic to enhance trust and reassure suspicious audiences. Through the lens of warranting theory, such denials could be viewed as an attempt to establish

authenticity and human authorship as "warrants" of message credibility (Hancock et al., 2020). However, such denials may also activate increased scrutiny. When an organization specifically states that no AI was involved in creating a message, it potentially invites audiences to actively evaluate the message for signs of AI involvement (Buchanan & Hickman, 2024).

From an information processing and persuasion standpoint, labeling content as solely human-authored may serve as a credibility heuristic (Dehnert & Mongeau, 2022). In real-world strategic communication contexts in which audiences are likely to be exerting relatively low effort when engaging with owned or paid media, these signals could be valuable if noticed. On the other hand, especially if audiences are not assuming their media environment may include undisclosed AI use, such denials may backfire by distracting attention from the substantive message or even raising suspicions. Given the existing research, we pose the research question:

RQ5: Do message effects differ depending on whether AI usage is explicitly denied compared to when it is not mentioned at all?

Methods

Participants and Design

This study, employing a between-subjects experimental design with a repeated-measures component, recruited 1500 participants from Prolific, targeting a representative sample of US adults. Data collection occurred in December 2024. Participants, compensated $2 for their time, were assigned to one of five AI disclosure conditions: no AI disclosure, explicit denial of AI usage, AI editing disclosure, AI generation disclosure, or late disclosure of AI generation. For generalizability, a between-subjects manipulation also randomized participants to view messages embedded on Facebook, Instagram, or LinkedIn. Each participant viewed four health messages on distinct topics—melanoma detection, strep throat detection, safe antibiotic use, and hearing

loss—presented with consistent AI disclosure conditions. All of the aforementioned hypotheses and research questions were pre-registered before data collection at the Open Science Framework. The anonymized pre-registration is available at https://osf.io/e5c2f/?view_only=aa7d361a425640f8aa9fd378e127b45d.

Procedure

The online experiment, hosted on Qualtrics, commenced with participants providing informed consent, followed by random assignment to AI disclosure and platform conditions. Participants then viewed each of the four health messages sequentially, responding to measures of message credibility, knowledge gain, and knowledge-seeking intention after each. Participants in the late disclosure condition received the AI disclosure after reviewing all messages. All participants subsequently completed source credibility (CDC perceptions) and organizational transparency measures.

Materials

The messages themselves were primarily visual with an added caption and attributed to the Centers of Disease Control (CDC), the major public health authority in the United States. Each message was sourced from CDC's Instagram account; for ecological validity, it was decided that the most realistic messages would be those that one already knows are deemed of acceptable quality by the CDC. Furthermore, this means that the messages were not actually generated by AI, to the knowledge of the research team. An operating assumption of the design is that a competent public health communicator would not use messages that are obviously AI-generated insofar as this would only be apparent due to quality issues. This design decision also allows a focus on the effects of disclosure rather than testing the capabilities of generative AI technology that is undergoing rapid improvements which render findings related to them almost

outdated upon completion. The stimuli are included with the anonymized open materials at https://osf.io/83fbq/files/osfstorage?view_only=60e1e4e4aaf74eb9ac5767466f0ed645.

The AI disclosure conditions were manipulated using a visual badge as well as wording integrated into the captions. The AI editing disclosure read "This message was edited by AI," while the AI generation disclosure stated "This message was generated by AI." The explicit denial of AI usage condition contained a statement denying AI involvement. The badge visuals were based on the Content Credentials framework, a new initiative spearheaded by Adobe, Microsoft, and other companies in the digital media industry aiming to label the provenance of images on the web. Although not yet widespread, using their designs ensured that the badging in the stimuli is realistic. The no AI disclosure condition contained no AI-related statements. Finally, in the late disclosure condition participants saw a message in the survey interface after they finished viewing all four messages. It stated that recent news reports found that the CDC had been using generative AI to make social media messages like the ones the participant had just seen but had decided not to disclose that usage. The goal of this condition is to explore the potential consequences of trying — but failing — to withhold transparency about generative AI use.

Measures

Message Credibility was measured with a 3-item, 7-point scale (Appelman & Sundar, 2016) assessing accuracy, authenticity, and believability, averaged into a composite score (M = 6.12, SD = 0.83). *Source credibility* of the CDC was assessed using an 8-item, 7-point scale (McCroskey & Teven, 1999) with subscales for *trustworthiness* (M = 5.51, SD = 1.43) and *expertise* (M = 5.93, SD = 1.13). *Perceived transparency* of the CDC was measured using a 4-item, 7-point scale (M = 4.84, SD = 1.38). The response choices were anchored at 1 (“Describes

very poorly") and 7 ("Describes very well"). *Knowledge gain* was measured using two multiple-choice questions per message, scored for correctness, targeting information that was manifest in the image (M = 6.32 out of 8, SD = 1.04). *Knowledge seeking* was assessed by tracking clicks on an information link following each message (M = 0.31 clicks out of 4 opportunities, SD = 0.87).

Data Analysis

Hypotheses and research questions were tested using multilevel and OLS regression models in R, with a significance threshold of $p < .05$ and one-tailed tests for hypotheses and two-tailed tests reported for research questions. Multilevel models, estimated with the "lme4" package (Bates et al., 2015), were used for message-level analyses. OLS regression was used for participant-level analyses. Linear hypothesis tests were conducted using the "marginaleffects" package (Arel-Bundock et al., 2024) to compare specific coefficients and *p* values for multilevel models were calculated using the "jtools" package with Satterthwaite degrees of freedom (Long, 2024).

Results

Table 1 presents the multilevel models examining message credibility. Hypothesis 1, predicting reduced message credibility with AI disclosure, was supported. Model 1 in Table 1 shows a significant negative effect of early AI disclosure on message credibility, $b = -0.18$, $SE = 0.04$, $t(1498) = -4.13$, $p < .001$. Research Question 1 concerned potential differences between describing the message as "generated" or "edited" by AI. As shown in Model 2 of Table 1, the coefficients for these variations are very similar to one another and "edited" is associated with a slightly worse credibility penalty than "generated." A test of the difference between these two coefficients is statistically insignificant ($p = .51$). In contrast, Research Question 5 explored

differences between active and passive non-disclosure, revealing both forms significantly *increased* message credibility relative to disclosure, as shown in Model 2 of Table 1. Specifically, passive non-disclosure had a coefficient of $b = 0.16$, $SE = 0.05$, $t(1497) = 3.28$, $p = < .001$ and active non-disclosure had a coefficient of $b = 0.23$, $SE = 0.06$, $t(1497) = 3.86$, $p < .001$. Although the estimated credibility benefit for active denial of AI usage was larger than when AI was not mentioned at all, a test of the difference between these coefficients was not statistically significant (p value for the test of the difference = .23).

Table 2 displays regression models for source credibility. Hypotheses 2a and 2b, predicting reduced source trustworthiness and expertise with AI disclosure, were supported. Disclosure significantly reduced both trustworthiness, $b = -0.18$, $SE = 0.08$, $t(1498) = -2.40$, $p = .01$, and expertise, $b = -0.19$, $SE = 0.06$, $t(1498) = -3.09$, $p = < .001$. Research Question 2 examined the penalty of late disclosure on source credibility. As shown in Table 2, late disclosure incurred a significantly greater penalty to trustworthiness compared to early disclosure, difference = -0.50, $z = 5.14$, $p < .001$. H3 investigates perceptions of organizational transparency. Here we see a significant difference between the estimates for early as opposed to late disclosure ($z = 2.98$, $p = .002$), although we note that early disclosure does not clearly boost transparency perceptions relative to non-disclosure ($b = 0.11$, $t(1498) = 1.35$, $p = .18$).

Table 3 presents multilevel logistic regression models for knowledge outcomes. Research Question 3 explored the effect of AI disclosure on learning. Early AI disclosure significantly reduced knowledge gain, $b = -0.19$, $SE = 0.07$, $z = -2.72$, $p = .01$. Research Question 4, also shown in Table 3, examined the effect of early AI disclosure on knowledge seeking behavior; this effect was not significant, $b = -0.05$, $SE = 0.39$, $z = -0.14$, $p = .89$, meaning no difference in click-through rates by experimental condition.

Discussion

Our study presents an early empirical investigation into the complex trade-offs involved when health organizations use AI to generate content but must navigate disclosure decisions. The findings have implications for health communicators in an era where AI tools are increasingly accessible yet public skepticism remains high. Put together, our findings suggest that early disclosure will render messages meaningfully less effective but preserve institutional credibility. On the other hand, if AI usage is revealed later, there may be significant reputational harm. Our descriptive results suggest that for these messages with no obvious quality deficits that might be expected with earlier versions of generative AI tools, participants did not suspect AI use when it was not disclosed. This also means that if AI messages outperform human-generated ones (which previous research suggests is plausible; Lim & Schmälzle, 2023), the most effective short-term strategy is to not disclose if there is no concern about detection. That said, doing so would be ethically questionable and resolving the ethical question is outside the purview of this study. A more ethically sound approach would be to avoid usage of AI in health messaging unless other benefits of using it (e.g., efficiency, quality, personalization) making up for the clear drawbacks to the credibility and cognitive processing of those messages.

This study's design has several strengths worth mentioning. First is the large, diverse sample of US adults, giving strong statistical power for most tests which is clearly paid off with precise rejections of several null hypotheses. Those hypotheses and associated analyses were also preregistered, offering readers assurance that the motivations and procedures are not *post hoc*. Stimuli were carefully created to be realistic and varied across platforms to ensure results would not be predicated on the peculiarities of some interface. Although formal disclosures of AI usage are not yet common, we did use a newly-introduced standard design for disclosure that is

relatively subtle without being easy to miss. For the same reasons, we did not rely on a single message or health topic; the goal is to generalize across a large number of health communication situations.

As for the particulars of our results, consistent with our predictions in H1, H2a, and H2b, disclosing AI involvement in message creation significantly reduced perceived message credibility, source trustworthiness, and source expertise. These findings align with and extend previous research on the machine heuristic (Sundar & Kim, 2019; Yang & Sundar, 2024), confirming that when audiences learn health content involves AI, they apply preexisting beliefs about machines that undermine credibility judgments regardless of the actual message quality. This extends previous work (Karinshak et al., 2023; Lim & Schmälzle, 2024) that found similar patterns with different health topics. The consistency of these findings across multiple health domains (and other domains, like news; Wang & Huang, 2024) makes clear that the public harbors reservations about the outputs of these technologies. Importantly, our experiment used actual CDC messages of professional quality, showing that the negative effects stem not from the content itself but from the attribution to AI. Although some research has suggested that AI might be perceived as more objective or unbiased in controversial contexts (Chung et al., 2023), we found no evidence of such benefits in our relatively low-controversy health topics. This may indicate that the advantages of perceived AI objectivity only emerge in explicitly polarized contexts where human biases are more salient concerns than expertise or empathy.

Our results regarding timing effects are particularly noteworthy. Supporting H3a and H3b, early disclosure increased perceptions of organizational transparency compared to late disclosure, which significantly harmed transparency perceptions. More striking was our finding addressing RQ2 – late disclosure caused substantially greater damage to source trustworthiness

than early disclosure. In fact, early disclosure had no effect on *source* credibility, only at the message level. This makes rather plain the tradeoffs involved; communicators either undermine the message via transparency or undermine themselves via obfuscation. These findings show that when organizations proactively disclose potentially negative information like AI involvement, they signal honesty that may mitigate credibility losses. In contrast, when disclosure appears forced or reactive, audiences appear to perceive intentional deception, amplifying the negative effects beyond those of the AI attribution itself.

Regarding RQ1, we found no significant difference between framing AI involvement as "editing" versus "generating" content. In fact, the estimates are in the direction of “edited” being worse for credibility of both the message and the source, raising a question about how these terms are understood by participants. This suggests that audiences may not make fine distinctions about the degree of AI involvement once it has been disclosed. This contrasts with Altay and Gilardi's (2024) finding that clarifying limited AI involvement mitigated negative effects. However, their interventions were more extensive, providing detailed explanations of the AI's role while this study opted for a perhaps more realistic and subtle intervention. This has important practical implications, suggesting that minor wording changes may not be enough – organizations might need more substantive explanations of AI's role if they hope to mitigate negative reactions. In the context of social media campaigns, such explanations may simply not be practical as they would quite possibly overwhelm the information content of the message.

Our findings on learning outcomes show a downside to AI disclosure that has not yet been explored in a literature primarily focused on credibility. In response to RQ3, we found that early AI disclosure significantly reduced knowledge gain from the messages. This suggests that when audiences know content is AI-generated, they may process it less deeply or allocate

attention differently, hindering actual learning despite the content's quality. We considered that perhaps the disclosure badge itself is a distraction; however, the active denial badge looks equivalent and was associated with an equivalent amount of learning as the badge-free message. This finding relates to expectancy violations theory (Burgoon et al., 1989, 2016) in AI contexts, suggesting that disclosure creates a negative expectancy violation that reduces cognitive engagement. Regardless of the exact cognitive cause, our results estimate that those who saw an AI disclosure knew 17% less of the content of the messages, which is hardly trivial. This suggests health organizations face a genuine trade-off between maximizing immediate message effectiveness and maintaining long-term source credibility if they insist on using AI in their creative process.

We found no significant effect of AI disclosure on information-seeking behavior (RQ4). This contrasts with some prior work suggesting that source perceptions affect willingness to engage further (Longoni et al., 2019). One explanation may be that our click-through measure was relatively low-cost and encountered during a paid study, potentially masking differences that might emerge in more naturalistic settings where attention is scarcer. Alternatively, this null finding might indicate that while AI disclosure affects initial message processing, it may not extend to subsequent information-seeking decisions once exposure has occurred.

Our test of RQ5 found no significant difference between explicitly denying AI involvement and simply not mentioning it, though both approaches yielded higher message credibility than disclosure. This provides partial support for the replicant effect (Jakesch et al., 2019), as content explicitly labeled as human-created was evaluated more positively, but the act of denying AI use did not provide additional benefits beyond passive non-disclosure. This suggests that in the current media environment, audiences may not be actively suspicious about

AI involvement in health messages unless prompted. The lack of added benefit from explicit denial is noteworthy from a practical perspective, as it implies organizations need not go out of their way to emphasize human authorship—simply avoiding mention of AI may be sufficient to prevent immediate credibility penalties. Given the estimate associated with active denial was higher, though, we do not want to make a strong conclusion that the two strategies are equivalent.

Several limitations should be acknowledged. First, our experiment used actual CDC messages, meaning participants saw professionally crafted content that was labeled as AI-generated rather than actual AI-generated content. While this design choice prioritized ecological validity and controlled for message quality, future research could consider comparing actual AI-generated content against human-created alternatives. Second, our study focused on relatively low-stakes, non-controversial health topics. The effects might differ for more sensitive issues or those with greater perceived consequences. Furthermore, while we measured immediate effects, longitudinal research is needed to assess whether disclosure effects persist over time or whether they primarily influence initial reception. Research could also explore interventions that might mitigate negative AI perceptions, such as explanations of human oversight or demonstrations of AI capabilities. Additionally, our sample, while reasonably representative of the US population, may not capture attitudes among specific vulnerable populations who might have different levels of trust in health authorities or technology. Future work should examine whether demographic factors, prior experiences with AI, or health literacy moderate the effects observed here.

As AI tools become increasingly integrated into health communication workflows, organizations face difficult decisions about transparency and disclosure. Our findings demonstrate that although AI disclosure carries immediate costs to message and source

credibility as well as learning outcomes, concealing and later revealing AI use incurs significant reputational penalties. This study provides empirical evidence of the trade-offs involved in AI disclosure decisions, suggesting that health organizations must carefully weigh short-term message effectiveness against long-term credibility and relationship maintenance. As public familiarity with AI continues to increase, these relationships may change, so continued research on technology, trust, and health communication will be necessary.

References

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and python. *Journal of Statistical Software*, *111*, 1–32. https://doi.org/10.18637/jss.v111.i09

Auger, G. A. (2014). Trust me, trust me not: An experimental analysis of the effect of transparency on organizations. *Journal of Public Relations Research*, *26*(4), 325–343. https://doi.org/10.1080/1062726X.2014.908722

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Buchanan, J., & Hickman, W. (2024). Do people trust humans more than ChatGPT? *Journal of Behavioral and Experimental Economics*, *112*, 102239. https://doi.org/10.1016/j.socec.2024.102239

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies*, *91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Burgoon, J. K., Newton, D. A., Walther, J. B., & Baesler, E. J. (1989). Nonverbal expectancy violations and conversational involvement. *Journal of Nonverbal Behavior*, *13*(2), 97–119. https://doi.org/10.1007/BF00990793

Burke-Garcia, A., & Soskin Hicks, R. (2024). Scaling the idea of opinion leadership to address health misinformation: The case for "health communication AI." *Journal of Health Communication*, *29*(6), 396–399. https://doi.org/10.1080/10810730.2024.2357575

Chung, M., Moon, W.-K., & Jones-Jang, S. M. (2023). AI as an apolitical referee: Using alternative sources to decrease partisan biases in the processing of fact-checking messages. *Digital Journalism*. https://doi.org/10.1080/21670811.2023.2254820

Dehnert, M., & Mongeau, P. A. (2022). Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion. *Human Communication Research*, *48*(3), 386–403. https://doi.org/10.1093/hcr/hqac006

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Hong, J. W., Peng, Q., & Williams, D. (2021). Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society*, *23*(7), 1920–1935. https://doi.org/10.1177/1461444820925798

Isaac, M. S., Jen-Hui Wang, R., Napper, L. E., & Marsh, J. K. (2024). To err is human: Bias salience can help overcome resistance to medical AI. *Computers in Human Behavior*, *161*, 108402. https://doi.org/10.1016/j.chb.2024.108402

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. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300469

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100093. https://doi.org/10.1016/j.chbah.2024.100093

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. *Proc. ACM Hum.-Comput. Interact.*, *7*(CSCW1), 116:1-116:29. https://doi.org/10.1145/3579592

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. https://doi.org/10.1016/j.jbusres.2024.114984

Lim, S., & Schmälzle, R. (2023). Artificial intelligence for health message generation: An empirical study using a large language model (LLM) and prompt engineering. *Frontiers in Communication*, *8*, 1129082. https://doi.org/10.3389/fcomm.2023.1129082

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

Long, J. A. (2024). jtools: Analysis and presentation of social scientific data. *The Journal of Open Source Software*, *9*(101), 6610. https://doi.org/10.21105/joss.06610

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

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

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs*, *66*(1), 90–103. https://doi.org/10.1080/03637759909376464

Monteith, S., Glenn, T., Geddes, J. R., Achtyes, E. D., Whybrow, P. C., & Bauer, M. (2024). Differences between human and artificial/augmented intelligence in medicine. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100084. https://doi.org/10.1016/j.chbah.2024.100084

Nadarzynski, T., Puentes, V., Pawlak, I., Mendes, T., Montgomery, I., Bayley, J., Ridge, D., & Newman, C. (2021). Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: A qualitative analysis. *Sexual Health*, *18*(5), 385–393. https://doi.org/10.1071/SH21123

Nan, X., Thier, K., & Wang, Y. (2023). Health Misinformation: What it is, Why People Believe it, How to Counter it. *Annals of the International Communication Association*, *47*(4), 381–410. https://doi.org/10.1080/23808985.2023.2225489

Ou, M., Zheng, H., Zeng, Y., & Hansen, P. (2024). Trust it or not: Understanding users' motivations and strategies for assessing the credibility of AI-generated information. *New Media & Society*, 14614448241293154. https://doi.org/10.1177/14614448241293154

Pareek, S., van Berkel, N., Velloso, E., & Goncalves, J. (2024). Effect of explanation conceptualisations on trust in AI-assisted credibility assessment. *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW2), 383:1-383:31. https://doi.org/10.1145/3686922

Rae, I. (2024). The effects of perceived AI use on content perceptions. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3613904.3642076

Schmälzle, R., & Wilcox, S. (2022). Harnessing artificial intelligence for health message generation: The folic acid message engine. *Journal of Medical Internet Research*, *24*(1), e28858. https://doi.org/10.2196/28858

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Toff, B., & Simon, F. M. (2024). "Or They Could Just Not Use It?": The Dilemma of AI Disclosure for Audience Trust in News. *The International Journal of Press/Politics*, 19401612241308697. https://doi.org/10.1177/19401612241308697

Wang, S., & Huang, G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research*, *51*(7), 815–842. https://doi.org/10.1177/00936502241229794

Wasike, B. (2022). Memes, Memes, Everywhere, nor Any Meme to Trust: Examining the Credibility and Persuasiveness of COVID-19-Related Memes. *Journal of Computer-Mediated Communication*, *27*(2), zmab024. https://doi.org/10.1093/jcmc/zmab024

Wischnewski, M., & Krämer, N. (2022). Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news. In *HHAI2022: Augmenting Human Intellect* (pp. 184–198). IOS Press. https://doi.org/10.3233/FAIA220198

Yang, H., & Sundar, S. S. (2024). Machine heuristic: Concept explication and development of a measurement scale. *Journal of Computer-Mediated Communication*, *29*(6), zmae019. https://doi.org/10.1093/jcmc/zmae019

Table 1

Multilevel Models Predicting Message Credibility

Predictor	**Model 1: H1 *b* (*SE*)**	**Model 2: RQ1 *b* (*SE*)**	**Model 3: RQ5 *b* (*SE*)**
Intercept	6.19 (0.05)***	6.19 (0.05)***	6.01 (0.05)***
Disclosure (Early)	-0.18 (0.04)***		
Edited Disclosure		-0.20 (0.05)***	
Generated Disclosure		-0.16 (0.05)***	
Passive Nondisclosure			0.16 (0.05)***
Active Nondisclosure			0.23 (0.06)***
R^2	0.68	0.68	0.68
AIC	13066.43	13071.57	13070.78
BIC	13099.92	13111.76	13110.98

Note. Random intercepts for participant and message topic were included in all models. N = 6000 observations, 1500 participants. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 2

Regression Models Predicting Source Credibility (H2, RQ2)

Predictor	**Trustworthiness *b* (*SE*)**	**Trustworthiness *b* (*SE*)**	**Expertise *b* (*SE*)**
Intercept	5.62 (0.06)***	5.62 (0.06)***	6.05 (0.05)***
Disclosure (All)	-0.18 (0.08)*		-0.19 (0.06)***
Disclosure (Early)		-0.01 (0.08)	
Disclosure (Late)		-0.51 (0.10)***	
R^2	0.01	0.02	0.01
Adj. R^2	0.01	0.02	0.01

Note. The first row of coefficients represents H2. The second two rows represent RQ2. N = 1500 participants. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3

Multilevel Logistic Regression Models Predicting Knowledge Outcomes (RQ3, RQ4)

Predictor	**RQ3: Knowledge Gain *b* (*SE*)**	**RQ4: Knowledge Seeking *b* (*SE*)**
Intercept	2.01 (0.59)***	-8.83 (0.40)***
Disclosure (Early)	-0.19 (0.07)**	-0.05 (0.39)
Pseudo R^2	0.49	0.96
AIC	8148.38	2135.61
BIC	8177.95	2162.40

Note. Model for RQ3 includes random intercepts for participant, message topic, and question. Model for RQ4 includes random intercepts for participant and message topic. N = 12000 observations for RQ3 and 6000 observations for RQ4, 1500 participants. * $p < .05$, ** $p < .01$, *** $p < .001$.