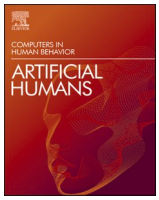
Computers in Human Behavior: Artificial Humans 2 (2024) 100058

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



The effect of source disclosure on evaluation of AI-generated messages

Sue Lim *, Ralf Schmälzle

Department of Communication, Michigan State University, 404 Wilson Rd. East Lansing, MI, 48824, USA

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Large language model (LLM)
Health communication
Source disclosure
Vaping prevention
Mixed effects modeling

ABSTRACT

Advancements in artificial intelligence (AI) over the last decade demonstrate that machines can exhibit communicative behavior and influence how humans think, feel, and behave. In fact, the recent development of ChatGPT has shown that large language models (LLMs) can be leveraged to generate high-quality communication content at scale and across domains, suggesting that they will be increasingly used in practice. However, many questions remain about how knowing the source of the messages influences recipients' evaluation of and preference for AI-generated messages compared to human-generated messages. This paper investigated this topic in the context of vaping prevention messaging. In Study 1, which was pre-registered, we examined the influence of source disclosure on young adults' evaluation of AI-generated health prevention messages compared to human-generated messages. We found that source disclosure (i.e., labeling the source of a message as AI vs. human) significantly impacted the evaluation of the messages but did not significantly alter message rankings. In a follow-up study (Study 2), we examined how the influence of source disclosure may vary by the adults' negative attitudes towards AI. We found a significant moderating effect of negative attitudes towards AI on message evaluation, but not for message selection. However, source disclosure decreased the preference for AI-generated messages for those with moderate levels (statistically significant) and high levels (directional) of negative attitudes towards AI. Overall, the results of this series of studies showed a slight bias against AI-generated messages once the source was disclosed, adding to the emerging area of study that lies at the intersection of AI and communication.

1. Introduction

"Imagine a world where persuasive content is crafted so masterfully that it becomes nearly indistinguishable from human creation, yet is generated by machines at the click of a button. This groundbreaking study unveils the potential of leveraging large language models (LLMs) to generate compelling messages, and puts it to the ultimate test: can they outperform human-crafted tweets in captivating the minds of their audience?" (Generated by GPT4-powered ChatGPT).

Recent technological breakthroughs in neural network modeling have ushered in an era of artificial intelligence (AI), and new AI-based systems such as OpenAI's ChatGPT are gaining rapid adoption. Within this context, the term AI generally refers to a field of study that aims to understand and build intelligent machines (Luger, 2005; Mitchell, 2019; Russell & Norvig, 2021). The precise and specific definition of intelligence differs based on the approach taken by the researchers, but a common theme is that machines can exhibit cognitive capacities such as intelligence, language, knowledge, and reasoning, which had traditionally been limited to human brains. AI technologies like ChatGPT, or similar systems (e.g., Google's Bard/Gemini, Meta's Llama) are driven by large language models (LLMs), a specific kind of transformer-based neural networks trained on massive amounts of text. Importantly, these LLMs can not only process and categorize text, but they can also be used to generate text that mimics the flow of natural human language (Bubeck et al., 2023; Hirschberg & Manning, 2015; Wei et al., 2022).

As the above content from ChatGPT shows, LLMs have advanced to the point where even with minimum instructions, they can generate high-quality creative and informative content. This has opened ample opportunities for health researchers and practitioners to leverage LLMs to augment their work. For instance, within health communication, researchers have found that messages generated by LLMs were clear and informative, and exhibited argument strength (Karinshak et al., 2023; Lim & Schmälzle, 2023; Schmälzle & Wilcox, 2022; Zhou et al., 2023). As LLMs continue to expand on these capabilities (Bubeck et al., 2023), we can expect to see LLMs being used as tools for generating persuasive

* Corresponding author.
E-mail addresses: limsue@msu.edu (S. Lim), schmaelz@msu.edu (R. Schmälzle).

https://doi.org/10.1016/j.chbah.2024.100058
Received 13 November 2023; Received in revised form 19 February 2024; Accepted 20 February 2024
Available online 21 February 2024

health messages. However, the rise of AI-generated content in the public communication environment raises the pressing question of how people react to AI as message creators.

Though this is a relatively novel area of study, there are two relevant bodies of literature that we can draw from: interdisciplinary research about the general sentiment of hesitancy towards novel technologies and source effects research within communication research. It is well-documented that new technologies are often met with skepticism. Studies suggest a general sentiment of hesitancy (e.g., von Eschenbach, 2021) and mild to moderate aversion (e.g., Castelo & Ward, 2021; Jussupow et al., 2020) towards AI and computer algorithms more broadly. Also, when told that AI was involved in the creation of communicative content, there was some reporting of preference against or lower evaluation of that content (e.g., Airbnb profile writing; Jakesch et al., 2019; email writing; Liu et al., 2022; generated paintings; Ragot et al., 2020; music creation; Shank et al., 2023; translation of written content; Asscher & Glikson, 2023). Within health contexts especially, some studies show that people tend to prefer human practitioners over AI-based technologies like chatbots when receiving consultation about health conditions (e.g., Miles et al., 2021), citing lack of personalization and incompetence in addressing individual needs as some of the reasons for hesitancy (Longoni et al., 2019).

Second, source effects have been studied extensively in persuasion and communication (Boster & Carpenter, 2021; Hovland et al., 1953). A plethora of literature has examined the influence of various aspects of the source, such as credibility, trustworthiness, and similarity, on people's attitudes and behavior (O'Keefe, 2015; Pornpitakpan, 2004; Wilson & Sherrell, 1993), as well as how source factors influence the processing and evaluation of messages supposedly crafted by specific sources (e.g., Bettinghaus et al., 1970; Chaiken, 1990). With the advancement of technology, research expanded to studying source effects in online settings, including how declared sources of online messages effected consumer behavior or health-related cognitions (Ismagilova et al., 2020; Ma & Atkin, 2017; Van Der Heide & Lim, 2016). In addition, some of the most well-known communication theories have examined cognitive mechanisms of source effects (ELM; Petty & Cacioppo, 1986; HSM; Chen & Chaiken, 1999). Speaking broadly, the results from these studies show that people's thoughts about the source of the message shape how they evaluate the communication content from the source. However, given the relative novelty of AI-related research, the role of AI as a source of communicative content has not been widely examined. Thus, it is not clear to what extent previous source effects research, which featured diverse features of human sources, can be generalized to the emerging AI-communication paradigm. Especially since there is already evidence that LLMs can augment health campaign practice, it is important to investigate how people respond health campaign messages authored by AI sources.

In addition to examining the role of AI as a new kind of message source, it will also be critical to identify potential moderators of such source effects. Much like how other individual differences can shape source effects in other domains (e.g. Bettinghaus et al., 1970), certain attitudes or personality traits could influence how people respond to AI-generated content. One apparent moderator are general attitudes towards AI. Indeed, there is already evidence that attitudes towards technologies (including AI) are associated with personality traits, demographics, and technology adoption (Kaya et al., 2024; Kwak et al., 2022; Tubaishat, 2014). Thus, if attitudes towards AI reflect people's level of openness and trust towards AI, then they could influence their perceptions of and responses to AI- generated health prevention content.

Based on these considerations, we conducted two experimental studies that shed light on the influence of source disclosure on the evaluation of prevention messages. For the first study (study 1), we examine how source disclosure influences young adults' evaluation of (in terms of effects perception) and preference for (in terms of ranking) prevention messages generated by a LLM compared to humans. Then, a follow-up study (study 2) investigates how the influence of source disclosure varies on the basis of people's general attitudes toward AI. The findings from our studies has the potential to augment the existing source effects literature by highlighting how people's awareness of LLM's role in message generation influences their evaluation of the messages.

2. Study 1

The goal of our first study was to examine whether source disclosure influenced people's evaluations of AI-generated messages as well as their preference for AI as the source of health information.

2.1. Hypotheses and Study Design

The current study examined how human participants respond to persuasive messages that were either generated by AI vs. humans by either adding accurate source labels to the messages (source disclosed) or not adding any labels (source not disclosed). As mentioned above, prior work from the earliest days of persuasion research as well as more recent work from computer-mediated communication document that source factors affect how people evaluate messages. Moreover, although the number of studies is still very low, a few empirical studies have examined how AI as the source changes people's evaluation of content. For example, Ragot et al. (2020) found that when people perceived AI as the creator of artwork, they evaluated the art more negatively (compared to human-generated artwork) in terms of beauty, novelty, or meaningfulness.

Specifically in the context of health messaging, Karinshak et al. (2023) examined how source disclosure impact people's ratings of health campaigns messages. They conducted a set of three exploratory studies that used GPT3 to generate high-quality vaccination promotion messages. The third study, which manipulated source labels, found that prevention messages generated by GPT3 were rated higher in terms of perceived message effectiveness compared to those written by CDC when none of the messages were labeled. However, messages labeled as AI-generated were rated lower in terms of argument strength and perceived message effectiveness compared to those labeled as created by CDC or those not labeled at all. These results underscore the promise of examining source effects in the context of AI-generated health communication content; however, given the novelty of the topic and the scarcity of existing evidence, many questions remain open. Together, based on preliminary evidence suggesting a bias against AI-generated content, we posed the following hypothesis:

Hypothesis 1. (H1): People who know the source of the messages will rate AI-generated messages lower and human-generated tweets higher than those who did not know the source.

In addition, there is evidence that people feel averse to certain tasks being completed by machines, especially if those tasks are considered as subjective tasks that require more flexible and nuanced understanding of the situation. For instance, Castelo et al. (2019) found that people clicked more on advertisements that showed a human advice-provider than an algorithm advice-provider when the task was subjective (dating advice). This was not the case for more a objective task (financial advice). Similarly, Newman et al. (2020) demonstrated that people may perceive decisions made by algorithms as less fair than the same decisions made by humans in the contexts of promotions and layoffs, and this perception did not change even after increased transparency about the factors that led to those decisions. Claudy et al. (2022) found that people prefer human-based decisions about resource allocation even though they recognized AI as having greater capability to make impartial decisions. Based on such results about people's behavioral preferences, we also wanted to examine the influence of source disclosure on a measure of ranked preference (as opposed to, or in addition to the more conventional rating measures, see H1). Therefore, we asked participants to also perform a ranking of the messages and posited the following:

H2. Those who know the source will prefer human-generated tweets vs. AI-generated prevention messages.

Compared to Karinshak et al. (2023), one of the few existing works that examine the effect of source disclosure on the evaluation of AI-generated health messages, we added key elements of innovation. First, our comparison of human-generated messages were tweets. Tweets are a relevant type of messaging since much discussion about vaping occurs via social media platforms such as Twitter (Lyu et al., 2021; Wang et al., 2023). Moreover, social media are a core part of today's public communication environment and are arguably a place that will be massively affected by the influx of AI-generated content.

Second, we used vaping prevention as the context of the messages. The use of e-cigarettes (or vaping) has become a significant public health concern in the last decade, especially because of the high prevalence of e-cigarette use among youth (<18 years of age) and young adults (18–24 years of age). About 20% of high school and 5% of middle school students reported vaping in 2020 (Wang et al., 2021); it was also estimated that about 15% of young adults were using e-cigarettes in 2020 (Boakye et al., 2022). Moreover, much of smoking and vaping-related marketing leverages the power of social media - or its capacity in disseminating information and ideas at a rapid speed through networks of people following one another (Nahon & Hemsley, 2013) - to influence audiences and promote tobacco products (Allem et al., 2017; Clark et al., 2016; Collins et al., 2019). To combat the detrimental effects of vaping, health researchers and professionals have invested significant efforts into developing and testing effective campaign messages (Boynton et al., 2023; Liu & Yang, 2020; Noar et al., 2020; Villanti et al., 2021), leading to guidelines for best practices (e.g., Vaping Prevention Resource, 2023). These efforts could be further augmented by the capabilities of LLMs in generating effective health messages (Karinshak et al., 2023; Lim & Schmälzle, 2023). Thus, we selected vaping prevention as a health context to examine the evaluation of messages coming from AI as the message source.[1]

2.2. Method

We pre-registered our hypotheses and procedures at as.predicted.[2] The local review board approved the study.

2.2.1. Participants

A total of 151 young adult participants (18–24 years old) were recruited from two platforms: the University study pool (97 participants) and Prolific (Palan & Schitter, 2018) platform (54 participants). We specifically selected the young adult age group because of the prevalence of vaping in this age demographic (Boakye et al., 2022). Those recruited from the university study pool were college students in lower level lecture courses, and they received course credit for their participation. In addition, we used Prolific's screener questions to only recruit those between 18 and 24 years old living in the US to match the general breakdown of age groups in literature. The participants were compensated $2.80 for their participation. We discarded the data from nine participants who did not complete the study or who completed the study in under 5 min, leaving 142 young adults (m_{age} = 20.78, sd_{age} = 1.78; 59% women) in the final sample (See Supplementary Materials A in Appendix A for additional details about the sample).

2.2.2. Experimental messages: human- and AI-generated

We relied on previously published procedures to generate messages via a LLM, collect human-generated messages, and select 30 total messages (15 AI, 15 human) for the experiment (Lim & Schmälzle, 2023). For details, see Supplementary Materials B in Appendix A. For the sake of relevance and length, we briefly outline the process here.

To collect human-generated messages, we scraped vaping prevention tweets with hashtags #dontvape, #novaping, #quitvaping, #stopvaping, #vapingkills, and #vapingprevention using the snscrape package (Snscrape, 2021) in Python. After cleaning the tweets, we randomly selected 15 tweets that had been retweeted at least once for the experiment.

For AI message generation and selection, we generated 500 total vaping prevention messages using the Bloom LLM, and then randomly selected a subset of 15 messages. Bloom is the largest open-source multilingual language model available (Scao et al., 2022). As mentioned in previous sections, Bloom, like GPT3, is powered by the transformer neural network, the most advanced ANN system currently available (Tunstall et al., 2022). Pre-trained with 1.5 TB of pre-processed text from 45 natural and 12 programming languages, Bloom allows for text generation using prompting (inputting the beginning part of the text and the language model completes the text) and a set of statistical parameters. We chose Bloom because of its free cost, full transparency of the training process and training data, and the ability to use it on a local machine via Jupiter notebooks or Google Colab without a special computing system called graphic processing unit (GPU), often required to run large computational tasks.

2.2.3. Experimental procedure and conditions

The experiment was conducted online via Qualtrics. Once participants consented to the study, the young adult participants were randomly assigned to one of two groups: control and treatment ($n_{control}$ = 72, $n_{treatment}$ = 70). Then the survey asked the participants to rate each message on four perceived message effectiveness items and rank the 30 messages (15 AI-generated vs. 15 tweets). The order of the two activities was randomized to control for order effects. The participants in the treatment condition read messages with source labels (e.g., "AI-Generated Message: Nicotine in vapes …", "Human-Generated Tweet: Nicotine in vapes can …") while those in the control condition were not provided the source labels. The source labels were true - no deception was used. Upon completing the main experiment, participants completed demographic questions and were debriefed about the study's purpose (see Fig. 1 for the full conceptual figure).

2.2.4. Measures

Study 1 included two main measures, effects perception and ranking, as well as demographic questions including age.

Effects Perception: Within health campaigns research, one of the most used message evaluation metrics is perceived message effectiveness (PME). According to Baig et al. (2019), the PME measure tends to cover two major constructs, message perceptions and effects perception. Message perceptions refer to the extent the messages seem credible and understandable, while effects perception refers to how the message promotes self-efficacy and behavioral intention. Baig et al. (2019) developed an effects perception scale that focused on examining the extent the message does what it is intended. Existing research showed that effects perception was highly associated with health campaign outcomes such as risk beliefs, attitudes, and behavioral intentions (Grummon et al., 2022; Noar et al., 2020; Rohde et al., 2021), meanwhile in some cases message perceptions did not have significant associations with these outcomes.

Thus, we adopted Baig et al.'s (2019) effects perception ratings in the context of vaping as people's measure of the perceived effectiveness of the messages. The measure included the following four survey items: "This message discourages me from wanting to vape," "This message makes me concerned about the health effects of vaping," "This message

[1] Going forward, one could also determine whether the specific health topic matters. For instance, based on psychometric models of risk perception (Slovic, 1987), one could predict that certain critical topics could be particularly prone to AI-source effects. However, we opted to start with a straightforward and widely applicable, current health topic that was also relevant for our participants.

[2] The link to the pre-registration is here: https://aspredicted.org/uh4vk.pdf.

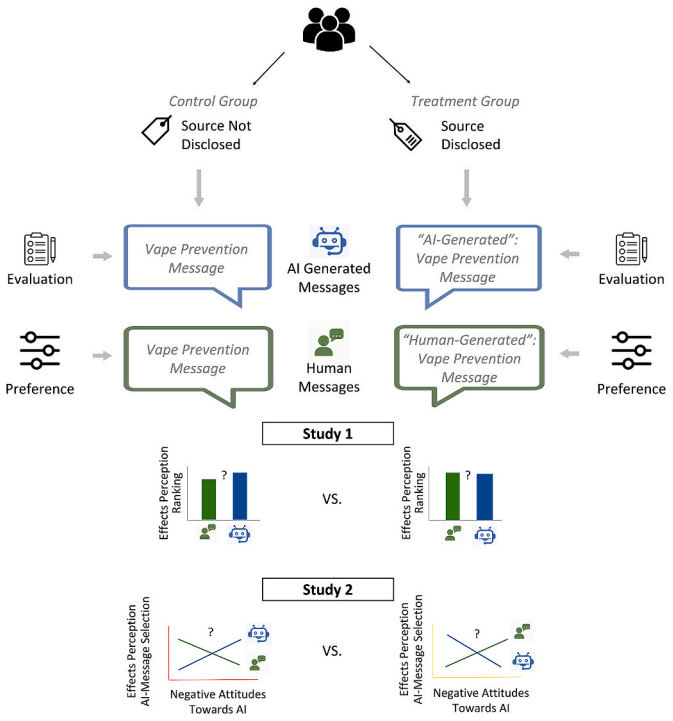


Fig. 1. Conceptual diagram of study design.

makes vaping seem unpleasant to me," and "This message makes vaping seem less appealing to me." Participants rated each item on a likert scale from 1 (Strongly disagree) to 5 (Strongly agree). Averaged across the 30 messages, the three items showed acceptable level of reliability (α = 0.76). The mean and standard deviation values by experimental condition are provided in the results section (Table 1).

Ranking: Second, for the ranking activity, we asked participants to rank the 30 messages from the best (1) to the worst (30) message by dragging each message to its rank. In addition to effects perception ratings, rankings have also been used in existing research to gather information about preference. Unlike ratings, rankings ask participants to order the messages from the best to the worst, using whatever criteria provided by the researcher and/or determined by the participants (Ali & Ronaldson, 2012). Rankings have been used extensively in the social sciences to gather data about constructs such as values (Abalo et al., 2007; Alwin & Krosnick, 1985), and attribute preferences (Lagerkvist, 2013). Within health communication, ranking measurement was used to examine people's preferences, including preferred health promotion icons (Prasetyo et al., 2021) and factors that influence demand for vaccinations (Ozawa et al., 2017).

2.2.5. Data analysis

All analyses were conducted in R. To examine H1, the responses for the three items of the EP scale were averaged into a composite EP score for each participant; the last item about the appeal of vaping was excluded from the analysis to keep consistent with the results from Baig et al. (2019). Then we fitted a linear mixed effects (LME) model via lmerTest R package (Kuznetsova et al., 2017) and lme4 R package (Bates et al., 2014) and allowed for the intercept and the effect of the AI-generated and human-generated messages to vary by the participant. For testing H2, we first subtracted the mean ranks for the human messages from the mean ranks of the AI messages (AI - Human). Thus, if the human-generated messages were on average ranked higher than AI-generated messages, then this difference value would be negative, and vice versa. Using the stats package (Chambers et al., 1992), we conducted the Wilcoxon Rank Sum Test, the non-parametric alternative to a two-sample ANOVA. We used the alpha level of $\alpha = 0.05$ to test for significance for both LME modeling and Wilcoxon Rank Test.

In addition, we conducted a supplementary computational analysis. The purpose of this was to extract and compare various textual features of the AI-generated messages and human-generated tweets, showing that the two groups of messages could be adequately compared. The textual methods we used included semantic analysis, n-gram analysis, topic modeling, sentiment analysis, and assessment of readability metrics. These analyses were carried out using Python and R packages including spacy, textacy, vader, topicmodels, and the sentence-transformers (DeWilde, 2020; Grün & Hornik, 2011; Honnibal et al., 2020; Hutto & Gilbert, 2014; Reimers & Gurevych, 2019). For all computational analysis of tweets, we removed the hashtags used to scrape the tweets. We also removed the prompts from the AI-generated messages for all analyses except semantic analysis. See Supplementary Materials C in Appendix A for the results of the supplementary analysis[3].

2.2.6. Deviation from pre-registration

While the main ideas from the pre-registration remained the same, we altered some of the details of the pre-registration. First, the pre-registration only included the data collection plan for the University sample. We decided to gather additional data from Prolific to make the results more generalizable to young adults beyond the University sample. Second, we decided to aggregate only the first three out of the four items for the EP measure to be more consistent with the existing literature (Baig et al., 2019). Finally, the statistical analysis methods were altered: For EP ratings, we chose to fit LME model instead of the originally registered mixed ANOVA to better account for the influence of individual differences in effectiveness perceptions; for the rank data, we used the Wilcoxon test, which is a two-sample extension of the Kruskal-Wallis test.

2.3. Results

First, we present the results from LME model, which tested the influence of source disclosure on message ratings (see Table 1 for mean effect perception information). We found that the effect of the source disclosure differed depending on the message source ($b = 0.10$, $SE = 0.046$, $t[140] = 2.18$; $p = 0.031$; see Table 2 and Fig. 2). A deeper inspection of the model showed that when the source was not disclosed,

Table 1
Mean and standard deviation of observed effects perception scores.

Mean (Standard Deviation)			
AI		*Human*	
Not Disclosed	*Disclosed*	*Not Disclosed*	*Disclosed*
4.31 (0.86)	4.23 (0.77)	4.18 (0.99)	4.21 (0.79)

Table 2
Influence of source disclosure on effects perception (LME model[a]).

	Estimate	*Standard Error*	*t*	*p-value*
Intercept	4.31	0.066	65.27	**<0.001**
Experimental Group: Disclosed (vs. Not Disclosed)	−0.077	0.094	−0.82	0.42
Message Source: Human (vs.AI)	−0.12	0.033	−3.83	**<0.001**
Experimental Group: Disclosed (vs. Not Disclosed) x Message Source: Human (vs. AI)	0.10	0.046	2.18	**0.031**

[a] Conditional $R^2 = .44$; Marginal $R^2 = .003$; ICC = .44.

[3] The anonymized data files and code files are available here: https://github.com/nomcomm/Evaluation_Vaping_Messages.git

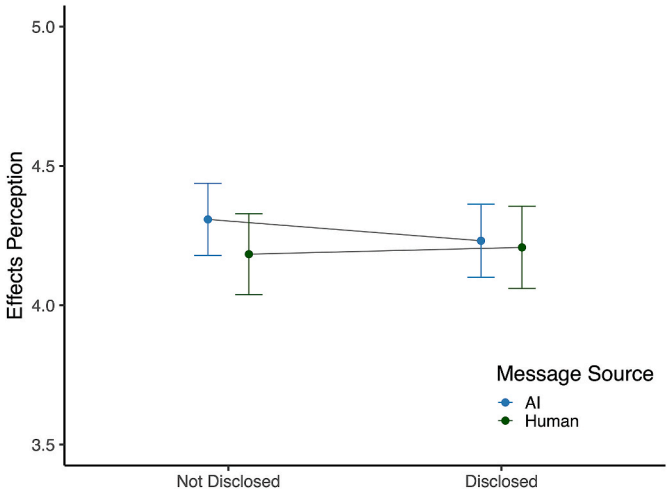


Fig. 2. Predicted effects perception scores by experimental condition (lme model).

AI-generated messages were perceived as more effective compared to human-generated messages (*difference* = 0.13, *SE* = 0.033, *z* = 3.83; *p* = <0.001; see Table 3). When the source was disclosed, EP ratings of AI-generated messages directionally decreased (see Table 1 and Fig. 2), leading to no differences in perceived effectiveness of AI generated and human-generated messages (*difference* = 0.024, *SE* = 0.033, *z* = 72; *p* = 0.47; see Table 3). This alludes to a slight bias in people's evaluation of messages based on the source of the message. Thus, H1 was partially supported.

To test H2, we compared the difference in the mean ranks of AI and human-generated messages (AI mean rank - Human mean rank) using the Wilcoxon Rank Sum Test. For the rank activity, the lower quantitative value represented a higher relative quality rank, with 1 representing the best message. Thus, the smaller differences in rank suggested a lower quantitative value for AI mean rank, hence a higher preference for AI-generated messages. We found that the median difference in rank for participants who knew the source, *Mdn* = −0.6, was slightly higher than the median difference in rank for participants who did not know the source, *Mdn* = −0.87, though this difference was not statistically significant (*W* = 2652.5, *p* = 0.59). These results suggested that the exposure to message source did not influence people's preference for the message source, measured by people's ranking of the 30 messages from the best (rank = 1) to the worst (see Fig. 3). Thus, our H2 was not supported.

2.4. Study 1 discussion

Study 1 examined how disclosing the source of a message as coming from an AI (vs. humans) influenced the evaluations of the messages and the preferences for the message source. Our H1 was partially supported -

Table 3
Pairwise comparisons of predicted EP ratings (LME model).

	Difference	*Standard Error*	*z*	*p-value*
Disclosed vs. Not Disclosed Conditions[a]				
AI-Generated Messages	−0.077	0.094	−0.82	0.42
Human-Generated Messages	0.024	0.11	0.23	0.82
AI vs. Human-Generated Messages**				
Not Disclosed	0.13	0.033	3.83	**<0.001**
Disclosed	0.024	0.033	0.72	0.47

[a] Difference Calculation: (EP ratings for disclosed - EP ratings for not disclosed) **Difference Calculation: (EP ratings for AI-gen. - EP ratings for human-gen. messages).

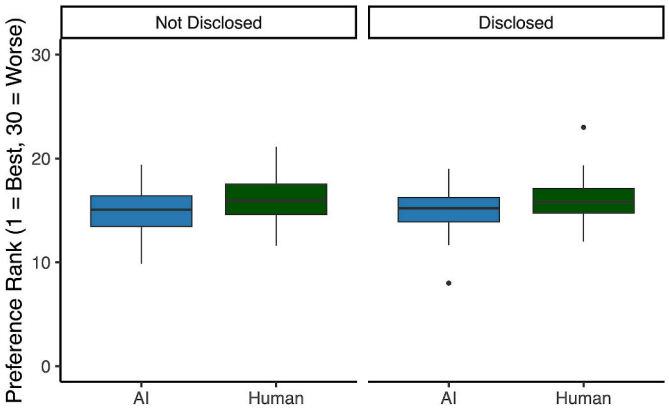


Fig. 3. Mean rank difference scores by experimental condition.

source disclosure slightly altered people's perception of the effectiveness of AI-generated and human-generated messages. Specifically, people who did not know the source perceived AI-generated messages as more effective than human-generated messages, whereas the perceived message effectiveness of the two groups did not differ for those who knew the source. These results suggest that source disclosure induced slight bias in people's evaluation of the messages. However, it is worth noting that both the AI and human-generated messages were rated as highly effective (>4.0 for the EP scale) across both conditions. This could be attributed to the specific population we were interested in (i.e., the messages overall seemed effective to the young adult population, whom, as noted in the introduction, were the demographic with the highest rate of vaping in the US). Thus, we thought it would be important to replicate this finding with a broader adult demographic for a follow-up study (Study 2).

Furthermore, our H2, which addressed the ranking task that required participants to make an active selection to express their preferences about messages, was not supported. This could have occurred for many reasons, one of which is that ranking all 30 messages may have required too much cognitive effort. Another reason could be that other factors, such as attitudes toward AI in general, could moderate the influence of source disclosure on preference.

To further inspect source effects of AI-generated messages, we conducted a follow-up study (Study 2), examining individual differences that could boost or buffer the effects of source disclosure. For instance, participants could vary in their attitudes about the use of and general sentiment towards AI, which in turn could influence their judgments of AI-generated content. Thus, we examined attitudes towards AI as a potential factor in Study 2.

3. Study 2

Study 2 replicated the source disclosure manipulation from Study 1 with a few modifications. First, we assessed people's preference for messages via message selection (top 5 out of 30) rather than the ranking task to decrease the participants' cognitive burden of comparing all 30 messages. Next, we expanded our population of interest to all adults. Finally, we examined how the influence of source disclosure on the evaluation and selection of AI-generated messages varied by the level of negative attitudes towards AI.

3.1. Negative attitudes towards AI as moderator

Within communication, persuasion, and social psychology more broadly, attitudes play a central role in shaping evaluative reactions to messages or stimuli (Thurstone, 1931; Petty & Cacioppo, 1986; Lindzey & Aronson, 1985, and the role of an evaluative dimension (good/bad, positive/negative) is widely recognized as a fundamental organizing

force of the human conceptual system (Osgood et al., 1955). Eagly and Chaiken (1993) define attitudes as *"a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor"* (pg. 1). Accordingly, attitudes towards a source can influence how people view the message (Tannenbaum, 1956). Adapting this line of thought to the present context, we may surmise that AI-related attitudes could lead to a more or less favorable response towards messages if the messages are explicitly designated as coming from an AI source (i.e., source-labeled or source-disclosed). In line with this, some evidence from existing studies examining related topics supports this reasoning. For instance, recent studies examined how source factors such as source credibility and expertise influences people's perceptions of the message or the object of interest (e.g., attitudes toward brand; Buda & Zhang, 2000; celebrity endorsement; Roy et al., 2013; evaluation of the product they found online; Smith, De Houwer, & Nosek, 2013; online recommendation credibility; Luo et al., 2013). Though not much empirical evidence specifically tested the moderating effect of negative attitudes towards AI, we apply the findings from the existing work about attitudes in general and message effects presented above to posit the following hypotheses:

H3. Negative attitude toward AI will moderate the influence of source disclosure on the evaluation of prevention messages.

In addition to message evaluation, the effect of source disclosure on preference of the source of the message can also vary by their attitudes towards the source. Arendt et al. (2019), for example, found that attitudes towards television stations predicted which news headline was selected. Though not directly about attitudes towards the source, Sülflow et al. (2019) and Winter and Krämer (2012) demonstrated that people's preference for online content varied by source-related cues. These findings, in conjunction with evidence presented above, suggests that the influence of disclosing the actual source of the message (AI vs. human) on message selection will differ by people's general attitudes towards AI:

H4. Negative attitude toward AI will moderate the influence of source disclosure on the preference for AI as the message source.

3.2. Method

3.2.1. Participants

The local review board approved the study. Like study 1, participants for study 2 were recruited from the university recruitment pool (86 participants) as well as Prolific platform (130 participants). Those recruited from the university recruitment pool were college students enrolled in a large lecture and received course credit for participating in the study. In addition, we expanded the age range of potential participants from Prolific platform (Palan & Schitter, 2018) to all adults (18 or older).[4] The participants were paid $2.83 as compensation for participating in the study. We discarded the data from 33 participants who did not complete the study, completed the study in under 5 min, or failed to pass the attention check questions, leaving 183 participants (m_{age} = 33.83, sd_{age} = 14.42; 56% women) in the final dataset (see Supplementary Materials A for further details on participant selection and samples).

3.2.2. Experimental procedure, measures, and data analysis

The same 30 messages from Study 1 were tested in the main study. The experiment followed the same procedure as study 1 ($n_{control}$ = 94, $n_{treatment}$ = 89) with a few modifications. First, we asked the first three items of the Effects Perception scale (Baig et al., 2019). Averaged across the 30 messages, the items showed good level of reliability (α = 0.85). The mean and standard deviation values by experimental condition are provided in the results section (Table 4). In addition, instead of ranking the messages, we asked participants to select the 5 best messages from the pool instead of having them rank all messages. This was done because, in campaign practice, the best-in-show messages are chosen from a larger pool of candidates. Moreover, having participants and all messages is rather taxing and we expected better compliance with a more focused task. Upon completing the main experiment, participants answered background and demographics questions that included questions about their attitudes towards AI.

Negative Attitude Towards AI: We adopted the subscale from Schepman and Rodway's (2023) general attitudes towards AI scale (GAAIS). A major part of the measure is based on the concept of trust in the capabilities and the uses of AI. The paper showed that GAAIS was associated with psychological features such as the Big Five personality, showing that it can be used to represent various individual differences that could exist when processing messages generated by AI. For example, Bellaiche et al. (2023) examined the association between attitudes towards AI and people's judgments of art labeled as AI-created or human-created. In this study, we adopted the negative attitudes towards AI subscale, which included people's concerns about and negative sentiment towards AI, as a moderator. The negative attitude toward AI scale asked people to rate 8 items related to negative attitudes (e.g., "I shiver with discomfort when I think about future uses of Artificial Intelligence") from a scale of 1 (strongly disagree) to 5 (strongly agree) (Schepman & Rodway, 2023). The overall mean and standard deviation were 3.04 and 0.80, and the scale showed good reliability (α = 0.84). The demographics questions stayed the same as in study 1.

All analyses were conducted in R. First, we calculated the average score for negative attitudes towards AI for each participant. To examine how the influence of source disclosure on EP of AI vs. human-generated messages differed by the extent of negative attitude (H3), we fitted a LME model, allowing for the intercept and the main effect of message source to vary by participant to consider individual differences in people's perception of the messages. Then, to examine how the effect of source disclosure on source preference differed by negative attitude (H4), we first calculated how many AI-generated messages were selected (out of 5), and then fitted a Poisson regression model.

3.3. Results

See Table 4 for the descriptive statistics of the EP measure with a broader adult sample. For the EP measure, we found a significant three-way interaction of source disclosure, message source (AI vs. Human), and the extent of having a negative attitude toward AI ($b = -0.14$, $SE = 0.058$, $t[179] = -2.39$; $p = 0.018$; see Table 5). Deeper inspection of the model through pairwise comparisons of EP ratings by negative attitudes towards AI showed that across levels of negative attitudes towards AI, knowing the source did not change people's perception of the effectiveness of AI-generated messages (see Table 6). Interestingly, among those who knew the source, negative attitudes towards AI was associated with greater perceived effectiveness of AI-generated messages compared to human-generated messages (see Fig. 4). Specifically, those with moderate to high levels of negative attitudes towards AI perceived AI-generated messages as more effective than human-generated messages (*difference* = 0.10, $SE = 0.033$, $z = 3.068$; $p = 0.0022$ for moderate level and *difference* = 0.13, $SE = 0.046$, $z = 2.88$; $p = 0.0039$ for high

Table 4
Mean and standard deviation of observed EP scores.

Mean (Standard Deviation)			
AI		*Human*	
Not Disclosed	Disclosed	Not Disclosed	Disclosed
4.08 (1.06)	4.10 (0.93)	3.92 (1.12)	4.0 (0.99)

[4] The participant pool was restricted to those who resided in Michigan to align with the requirements of the grant funding this study.

Table 5
Effects of source disclosure on EP, attitudes towards AI as moderator (LME model).

Term	*Estimate*	*Standard Error*	*t*	*p-value*
Intercept	3.32	0.30	10.97	**<0.001**
Experimental Group: Disclosed (vs. Not Disclosed)	0.61	0.45	1.34	0.18
Message Source: Human (vs. AI)	−0.46	0.12	−3.75	**<0.001**
Negative Attitude Towards AI	0.25	0.097	2.61	**0.0099**
Experimental Group: Disclosed (vs. Not Disclosed) x Message Source: Human (vs. AI)	0.48	0.18	2.60	**0.010**
Experimental Group: Disclosed (vs. Not Disclosed) x Negative Attitude Towards AI	−0.20	0.14	−1.36	0.18
Message Source: Human (vs. AI) x Negative Attitude Towards AI	0.099	0.039	2.52	**0.013**
Experimental Group: Disclosed (vs. Not Disclosed) x Message Source: Human (vs. AI) x Negative Attitude Towards AI	−0.14	0.058	−2.39	**0.018**

*Conditional R^2 = .57; Marginal R^2 = .036; ICC = .56.

Table 6
Pairwise comparisons of predicted EP ratings by negative attitudes towards AI.

	Difference	*Standard Error*	*z*	*p-value*
Disclosed vs. Not Disclosed Conditions[a]				
AI-Generated Messages				
Negative Attitudes Towards AI = 2.24[+]	0.17	0.16	1.037	0.30
Negative Attitudes Towards AI = 3.04	0.013	0.12	0.11	0.91
Negative Attitudes Towards AI = 3.84	−0.14	0.16	-0.89	0.38
Human-Generated Messages				
Negative Attitudes Towards AI = 2.24	0.34	0.17	2.03	**0.042**
Negative Attitudes Towards AI = 3.04	0.068	0.12	0.58	0.56
Negative Attitudes Towards AI = 3.84	−0.20	0.16	−1.22	0.22
AI vs. Human-Generated Messages[b]				
Not Disclosed				
Negative Attitudes Towards AI = 2.24	0.24	0.044	5.38	**<0.001**
Negative Attitudes Towards AI = 3.04	0.16	0.032	4.85	**<0.001**
Negative Attitudes Towards AI = 3.84	0.078	0.046	1.68	0.093
Disclosed				
Negative Attitudes Towards AI = 2.24	0.070	0.049	1.43	0.15
Negative Attitudes Towards AI = 3.04	0.10	0.033	3.068	**0.0022**
Negative Attitudes Towards AI = 3.84	0.13	0.046	2.88	**0.0039**

[+]3.04 = overall mean; 2.24 = mean - 1 standard deviation; 3.84 = mean + 1 standard deviation.
[a] Difference Calculation: (EP ratings for disclosed - EP ratings for not disclosed).
[b] Difference Calculation: (EP ratings for AI-gen. - EP ratings for human-gen. messages).

level of negative attitudes towards AI). The opposite pattern was observed for those who did not know the source.

Table 7 shows the results for messages selection. There was no moderation effect of negative attitudes toward AI ($b = -0.042$, $SE = 0.12$, $p = 0.73$), and H4 was not supported. However, a deeper inspection of the results showed that those who knew the source selected less number of AI-generated messages compared to those who did not know

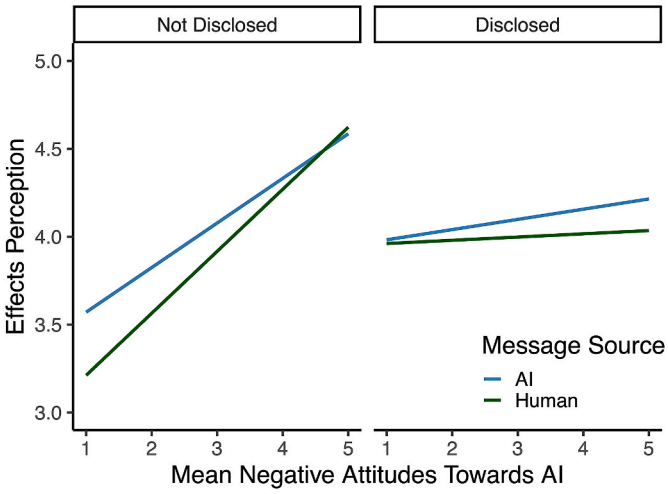


Fig. 4. Predicted EP ratings by levels of negative attitudes towards AI

Table 7
Attitudes towards AI and selection of AI-Generated messages.

	Estimate[a]	*Standard Error*	*z*	*p-value*
Intercept	0.74	0.25	2.97	**<0.003**
Experimental Group: Disclosed (vs. Not Disclosed)	−0.093	0.39	−0.24	0.81
Negative Attitude Towards AI	0.070	0.079	0.90	0.37
Experimental Group: Disclosed (vs. Not Disclosed) x Negative Attitude Towards AI	−0.042	0.12	−0.34	0.73

[a] Note: The coefficient estimations are log counts; R^2 = .057.

Table 8
Pairwise comparison of AI-Message selection by negative attitudes towards AI.

	Disclosed -Not Disclosed	*Standard Error*	*z*	*p-value*
Negative Attitudes Towards AI = 2.24[a]	−0.42	0.31	−1.33	0.18
Negative Attitudes Towards AI = 3.04	−0.51	0.23	−2.27	**0.023**
Negative Attitudes Towards AI = 3.84	−0.62	0.33	−1.88	0.060

[a] 3.04 = overall mean; 2.24 = mean - 1 standard deviation; 3.84 = mean + 1 standard deviation.

the source for those with moderate (i.e., mean) and slightly less number of AI-generated messages for those with high (i.e., 1 standard deviation above the mean) levels of negative attitudes towards AI (see Fig. 5 and Table 8). The results suggest that negative attitudes was not a significant moderator. Instead, high levels negative attitude towards AI was generally associated with decreased preference for AI-generated messages when the source was disclosed.

3.4. Study 2 discussion

Study 2 examined whether negative attitudes towards AI moderated the influence of source disclosure on the evaluation of and preference for AI-generated vs. human-generated messages. For EP ratings, having a negative attitude toward AI emerged as a significant moderator, supporting H3. While negative attitude toward AI emerged as a significant moderator (supporting H3), we also observed an unexpected pattern of results: when the source was disclosed, negative attitudes towards AI was associated with greater perceived effectiveness of AI-generated

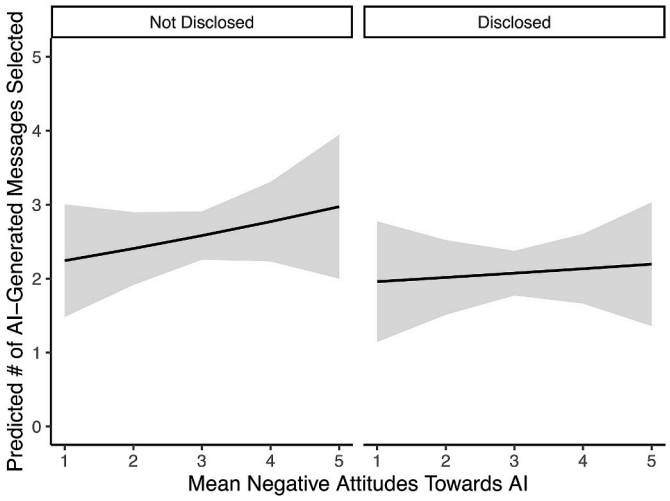


Fig. 5. Predicted number of AI-generated messages selected.

messages compared to human-generated messages. The opposite pattern was observed for those who did not know the source of the messages.

From the perspective of the experimental manipulations and their outcomes, we can thus confirm interaction effects among these variables (source disclosure, message source, and attitude towards AI). However, from an information processing perspective (i.e., how people take in, process, and then evaluate the messages in the varying conditions), the processes remain somewhat opaque. One potential explanation could be that the level of negative attitudes towards AI could have influenced how closely participants examined the messages: those with greater levels of negative attitudes towards AI could have paid more attention to the messages compared to those with less negative attitudes towards AI. There's been some evidence from literature that suggests that people scrutinize information from less trustworthy sources than that from more trustworthy sources under certain circumstances (Priester & Petty, 1995, 2003). Greater scrutiny of the messages could have focused people's attention more to the message content rather than the fact that AI was the source of the message. Furthermore, even for the current AI-generated health messages, humans were still somewhat involved in the process (see Supplementary Material B in Appendix A for details of how the messages used for this study were crafted), and awareness of this could also obscure the distinction between AI- and human-generated messages. Taken together, it is possible that people's general attitudes towards AI did not play as large of a role as we expected in their evaluation of the generated messages.

Lastly, for message selection, we did not find any significant moderating effects; thus, our H4 was not supported. However, source disclosure significantly decreased the number of AI-generated messages selected for those with mean levels (statistically significant) and high levels (directional) of negative attitudes towards AI. These results provide support for people's preference against AI-generated messages when they perceive the source of the message as an AI. We discuss the theoretical and practical implications of our findings in the next section.

4. Overall discussion

4.1. Summary of the findings

Overall, our results provide novel insights into the effects of source disclosure on people's evaluations and preferences of the messages. We found that people perceived AI-generated as more effective than the human-generated messages when the source was masked, whereas there were no significant differences in the ratings after the source was disclosed (partially supporting H1 in Study 1). As hypothesized in study 2 (H3), negative attitudes towards AI moderated the effect of source disclosure on the evaluation of AI and human generated messages. Interestingly, for those who saw the source labels, negative attitudes towards AI was associated with greater perceived effectiveness of AI-generated messages compared to human-generated messages. This pattern was not observed for those who did not know the source.

Though the analyses of ranking and message selection tasks did not support our hypotheses (H2 in Study 1 and H4 in Study 2), they also revealed interesting effects: source disclosure decreased the number of AI-generated messages selected for those with moderate levels of negative attitudes towards AI. These results suggest a slight negative bias against AI-generated messages, aligning with previous studies that showed hesitation and slight negative bias against the communicative content when participants believed AI was involved in the process (Asscher & Glikson; Jakesch et al., 2019; Karinshak et al., 2023; Liu et al., 2022; Ragot et al., 2020; Shank et al., 2023).

4.2. Implications for source effects research

This paper contributes to the emerging area of study at the intersection of communication and AI by being one of the first papers to examine how knowing the source changes people's evaluation of and preference for AI-generated messages.

The source of a message has always been an integral part of theories and models of communication and persuasion, even going back to Aristotle's rhetoric theory (Murphy, 1981). Likewise, early models of social scientific communication research, such as Berlo's SMCR model (Berlo, 1960), Lasswell's model of communication (Lasswell, 1948), and even the Shannon-Weaver model of communication (Shannon, 1948) all included components about the source, or the creator and deliverer of the message. Since then, a plethora of studies have studied source effects, or how various characteristics of the source impact the way people receive, process, and subsequently make judgments about the message. These studies often manipulated certain aspects about the source (e.g., expert vs. nonexpert; Clark et al., 2012) and examined in which scenarios the various levels led to greater persuasive outcomes (e.g., when people had little information about a product, they relied on expert sources, but not necessarily when they had more information; Ratneshwar & Chaiken, 1991).

With the rise of AI-based technologies such as LLMs, source effects have once again come to the forefront of communication research, but the notion of "source" for AI-generated messages is quite complex. In particular, the message generation process for LLMs generally consists of the following steps: First, a human user feeds prompts, or intentionally crafted instructions or beginning parts of the message, to the LLM; second, the user adjusts the parameters, such as how many messages should be crafted and other factors; third, the LLM generates the messages according to step 1 and 2. Thus, in this process, it is actually a human who initiates the message creation sequence, whereas the LLM only completes the message generation command. It seems plausible to assume that people's knowledge of AI (and their perceptions of its expertise, trustworthiness, etc.), will impact their evaluations. For example, we may surmise that it would not only matter that a message was AI-generated, but also whom people believe to have started the process. In other words, if people think that the AI message generation was initiated by expert organizations, such as the Center for Disease Control (CDC), evaluation might differ compared to AI-generated messages initiated by general users of social media platforms, or even by agents from a foreign country. In sum, with AI, there is an intersection of source effects, perceptions of AI, and various social-cognitive inferences about creator, intent, and expertise. Going forward, it will thus be important to comprehensively study these topics.

Based on the present results, we can thus say that there are small but significant effects of source disclosure, consistent with a small preferential treatment for human-generated messages. The effects uncovered in the current research are small in terms of statistical effect sizes (e.g., Cohen, 1988). In particular, compared to relatively robust effects of

source manipulations in other domains, the effects in this study are orders of magnitude weaker (e.g., Wilson & Sherrell, 1993), and nowhere near the strength of source effects for doctors vs. novices (for health advice) or celebrities vs. nobodies (for consumer behavior, e.g., Atkin & Block, 1983; Rollins et al., 2021).

However, it would be hasty to conclude from these results that AI-related source effects are inconsequential. Rather, we note that the small effects could also have to do with the nature of our messages or the issue at hand, as the tweets are relatively short and information-limited messages, and the health issue of vaping is less severe compared to other issues, such as messages about cancer or major surgery. Thus, it may well be that AI-related source effects in other contexts could be much stronger (e.g., jury justifications in life-and-death-trials that were generated by human juries or AI-advisors). Lastly, we also hold that even nominally very small effects can have major consequences if aggregated over longer periods of time or multiple individuals, or if they affect variables that are difficult to impact (see Prentice & Miller, 1992). As an analogy, we can already observe that small, but consistent content preferences by humans can be picked up (and potentially amplified) by algorithms on social media platforms. Thus, even if AI-related source effects were only small, they would still matter if major platforms decided to add or not add AI-source-labels to AI-generated messages.

4.3. Implications for public health campaigns

The current results have interesting implications for research on health message generation and dissemination. With the advent of AI-language models, it has become extremely easy to generate high-quality health messages about any given topic. This potential can either be a blessing or a curse, depending on the source and their intent. For instance, if the CDC leveraged the power of AI for health message generation, this would be seen as largely beneficial; however, malicious actors could also leverage AI to spread fake news - or even just promote unhealthy products (e.g., cigarettes). Indeed, there are already commercial applications of AI-LLMs for copywriting purposes, and these could also be used to influence users towards unhealthy, risky, or other kinds of behaviors. Thus, more work is needed to explore how these aspects intersect with the topic of AI-*as*-message-source as well as the influence of source disclosure.

A related concern is about the factual truthfulness of health-related claims. It is well known that although LLMs are capable of generating persuasive messages, they are prone to hallucinations (Kaddour et al., 2023; Zhang et al., 2023). Although the creators of AI systems are investing large efforts to minimize such false generations, this is still an unsolved problem of the underlying technology. These false generations will affect the evaluation of AI systems (Marcus, 2018, 2020), particularly whether AIs are seen as knowledgeable, reliable, and trustworthy. In sum, while we can expect that AI-generated messages will increasingly find their way into real-world health campaigns, numerous questions persist about their accuracy and the intent of the humans generating the messages using AI. At this point in time, the dynamically evolving landscape of AI-language generation systems prevents any final answers to these questions. Rather, longitudinal research would be needed to assess how people think about AI sources, how they adapt to the increasing prevalence of AI content, and how their evaluations are influenced by contextual factors.

4.4. Limitations, future avenues, and ethical considerations

As with all research, several limitations that require future research and important ethical considerations are worth highlighting. One limitation is that this study used tweets as messages. It would be interesting to examine other kinds of health messages, such as longer flyers and posters (Cho, 2011). The decision to use tweets was made because we wanted to take into account user-generated messages and because tweets have become a rather widespread form of health communication content that also gets used by the CDC and other key health organizations. In addition, the topic of AI-generated health messages raises ethical questions. In particular, the regulatory framework around these topics is currently in flux, and discussions about mandatory labeling of AI-generated content have barely even begun. Furthermore, the allowed use cases for AI content generation are also debated. For instance, using AI to generate medical diagnoses is explicitly prohibited by the creators, but generating general health information falls within the range of acceptable use (BigScience, 2022). Finally, no manipulation checks were included in the two studies because the experimental manipulation consists of clearly altering the feature of the messages (adding source labels in the beginning of the message vs. not adding the source labels; see O'Keefe, 2003). Instead, we added the source label in the beginning of each message and took measures to account for the participants' attention in order to maximize the likelihood of people's exposure to our manipulation. Future research could implement eye-tracking or other measures to check people's exposure to the source labels and asses AI-related attitudes while avoiding priming them about the purpose of the study.

5. Summary and conclusion

Taken together, we examined the influence of source disclosure on evaluations of AI-generated messages. We found that source disclosure (i.e., labeling the source of a message as AI vs. human) significantly impacted the evaluation of the messages, albeit the effects were of relatively small magnitude, but did not significantly alter message rankings. Moreover, in study 2 we found a significant moderating effect of negative attitudes toward AI on message evaluation. Our results show that at the point when we conducted our research, humans appear to exhibit a small preference for human-generated content if they know the source, but AI-generated messages are evaluated as equally good, if not better, if the source stays unknown. These results highlight the role of source factors for communication, and they have implications for the potential labeling of AI-generated content in the context of health promotion efforts.

Funding statement

This work was supported in part through Michigan State University's Institute for Cyber-Enabled Research Cloud Computing Fellowship, with computational resources and services provided by Information Technology Services and the Office of Research and Innovation at Michigan State University. This research was additionally supported with funding from the Charles J. Strosacker Graduate Research Fund for Health and Risk Communication in the Michigan State University College of Communication Arts and Sciences.

Declaration of competing interest

Both authors declare that there's no financial/personal interest or belief that could affect their objectivity. They do not have any potential competing interests.

CRediT authorship contribution statement

Sue Lim: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Ralf Schmälzle:** Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.chbah.2024.100058.

References

Abalo, J., Varela, J., & Manzano, V. (2007). Importance values for importance–performance analysis: A formula for spreading out values derived from preference rankings. *Journal of Business Research, 60*(2), 115–121. https://doi.org/10.1016/j.jbusres.2006.10.009

Ali, S., & Ronaldson, S. (2012). Ordinal preference elicitation methods in health economics and health services research: Using discrete choice experiments and ranking methods. *British Medical Bulletin, 103*(1), 21–44. https://doi.org/10.1093/bmb/lds020

Allem, J. P., Escobedo, P., Chu, K. H., Soto, D. W., Cruz, T. B., & Unger, J. B. (2017). Campaigns and counter campaigns: Reactions on twitter to e-cigarette education. *Tobacco Control, 26*(2), 226–229. https://doi.org/10.1136/tobaccocontrol-2015-052757

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49*(4), 535–552. https://doi.org/10.1086/268949

Arendt, F., Northup, T., & Camaj, L. (2019). Selective exposure and news media brands: Implicit and explicit attitudes as predictors of news choice. *Media Psychology, 22*(3), 526–543. https://doi.org/10.1080/15213269.2017.1338963

Asscher, O., & Glikson, E. (2023). Human evaluations of machine translation in an ethically charged situation. *New Media & Society, 25*(5), 1087–1107. https://doi.org/10.1177/14614448211018833

Atkin, C., & Block, M. (1983). Effectiveness of celebrity endorsers. *Journal of Advertising Research, 23*(1), 57–61.

Baig, S. A., Noar, S. M., Gottfredson, N. C., Boynton, M. H., Ribisl, K. M., & Brewer, N. T. (2019). UNC perceived message effectiveness: Validation of a brief scale. *Annals of Behavioral Medicine, 53*(8), 732–742. https://doi.org/10.1093/abm/kay080

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv*. https://doi.org/10.48550/arXiv.1406.5823

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., … Seli, P. (2023). Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research, 8*(1), 1–22. https://doi.org/10.1186/s41235-023-00499-6

Berlo, D. (1960). *The process of communication*. Holt, Rinehart, and Winston.

Bettinghaus, E. P., Miller, G., & Steinfatt, T. (1970). Source evaluation, syllogistic content, and judgments of logical validity by high- and low-dogmatic persons. *Journal of Personality and Social Psychology, 4*, 614–621. https://doi.org/10.1037/h0029864

BigScience. (2022). *BigScience RAIL license v1.0*. https://huggingface.co/spaces/bigscience/license.

Boakye, E., Osuji, N., Erhabor, J., Obisesan, O., Osei, A. D., Mirbolouk, M., … Blaha, M. J. (2022). Assessment of patterns in e-cigarette use among adults in the US, 2017-2020. *JAMA Network Open, 5*(7), Article e2223266-e2223266. https://doi.org/10.1001/jamanetworkopen.2022.23266

Boster, F., & Carpenter, C. (2021). *Critical questions in persuasion research*. Cognella Publishing.

Boynton, M. H., Sanzo, N., Brothers, W., Kresovich, A., Sutfin, E. L., Sheeran, P., & Noar, S. M. (2023). Perceived effectiveness of objective elements of vaping prevention messages among adolescents. *Tobacco Control, 32*(e2), e228–e235. https://doi.org/10.1136/tobaccocontrol-2021-057151

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712.

Buda, R., & Zhang, Y. (2000). Consumer product evaluation: The interactive effect of message framing, presentation order, and source credibility. *Journal of Product and Brand Management, 9*(4), 229–242. https://doi.org/10.1108/10610420010344022

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Castelo, N., & Ward, A. F. (2021). Conservatism predicts aversion to consequential Artificial Intelligence. *PLoS One, 16*(12), Article e0261467. https://doi.org/10.1371/journal.pone.0261467

Chaiken, S. (1990). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*, 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chambers, J. M., Freeny, A., & Heiberger, R. M. (1992). Analysis of variance; designed experiments. In T. J. Hastie (Ed.), *Statistical models in S* (pp. 145–193). Routledge.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken, & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). The Guilford Press.

Cho, H. (Ed.). (2011). *Health communication message design: Theory and practice*. Sage Publications.

Clark, E. M., Jones, C. A., Williams, J. R., Kurti, A. N., Norotsky, M. C., Danforth, C. M., & Dodds, P. S. (2016). Vaporous marketing: Uncovering pervasive electronic cigarette advertisements on Twitter. *PLoS One, 11*(7), Article e0157304. https://doi.org/10.1371/journal.pone.0157304

Clark, J. K., Wegener, D. T., Habashi, M. M., & Evans, A. T. (2012). Source expertise and persuasion: The effects of perceived opposition or support on message scrutiny. *Personality and Social Psychology Bulletin, 38*(1), 90–100. https://doi.org/10.1177/0146167211420733

Claudy, M. C., Aquino, K., & Graso, M. (2022). Artificial intelligence can't be charmed: The effects of impartiality on laypeople's algorithmic preferences. *Frontiers in Psychology, 13*, 898027. https://doi.org/10.3389/fpsyg.2022.898027

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Collins, L., Glasser, A. M., Abudayyeh, H., Pearson, J. L., & Villanti, A. C. (2019). E-cigarette marketing and communication: How e-cigarette companies market e-cigarettes and the public engages with e-cigarette information. *Nicotine & Tobacco Research, 21*(1), 14–24. https://doi.org/10.1093/ntr/ntx284

DeWilde, B. (2020). Textacy: NLP, before and after spaCy. https://github.com/chartbeat-labs/textacy. (Accessed 10 September 2022).

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Brace: Harcourt. Jovanovich.

Grummon, A. H., Hall, M. G., Mitchell, C. G., Pulido, M., Sheldon, J. M., Noar, S. M., … Brewer, N. T. (2022). Reactions to messages about smoking, vaping and COVID-19: Two national experiments. *Tobacco Control, 31*(3), 402–410. https://doi.org/10.1136/tobaccocontrol-2020-055956

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*, 1–30. https://doi.org/10.18637/jss.v040.i13

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python*. https://doi.org/10.5281/zenodo.1212303

Hovland, C. I., Janis, I. L., & Kelley, H. K. (1953). *Communication and persuasion*. New Haven, CT: Yale University Press.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. *Journal of Retailing and Consumer Services, 53*, 101736. https://doi.org/10.1016/j.jretconser.2019.01.005

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 *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13).

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *ECIS 2020 Proceedings*.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and applications of large language models*. https://doi.org/10.48550/arXiv.2307.10169. arXiv.

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. In *Proceedings of the ACM on human-computer interaction (CSCW)*.

Kaya, F., Aydin, F., Schepman, A., Rodway, P., Yetişensoy, O., & Demir Kaya, M. (2024). The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. *International Journal of Human-Computer Interaction, 40*(2), 497–514. https://doi.org/10.1080/10447318.2022.2151730

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Kwak, Y., Ahn, J. W., & Seo, Y. H. (2022). Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. *BMC Nursing, 21*(1), 1–8. https://doi.org/10.1186/s12912-022-01048-0

Lagerkvist, C. J. (2013). Consumer preferences for food labelling attributes: Comparing direct ranking and best–worst scaling for measurement of attribute importance, preference intensity and attribute dominance. *Food Quality and Preference, 29*(2), 77–88. https://doi.org/10.1016/j.foodqual.2013.02.005

Lasswell, H. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The Communication of ideas* (pp. 37–51). New York: Institute for Religious and Social Studies.

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

Lindzey, G., & Aronson, E. (1985). *The handbook of social psychology* (3rd ed.). New York: Random House.

Liu, S., & Yang, J. Z. (2020). Incorporating message framing into narrative persuasion to curb e-cigarette use among college students. *Risk Analysis, 40*(8), 1677–1690. https://doi.org/10.1111/risa.13502

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–13).

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

Luger, G. F. (2005). *Artificial Intelligence: Structures and strategies for complex problem solving*. Pearson Education.

Luo, C., Luo, X. R., Schatzberg, L., & Sia, C. L. (2013). Impact of informational factors on online recommendation credibility: The moderating role of source credibility. *Decision Support Systems, 56*. https://doi.org/10.1016/j.dss.2013.05.005, 92-10.

Lyu, J. C., Luli, G. K., & Ling, P. M. (2021). Vaping discussion in the COVID-19 pandemic: An observational study using Twitter data. *PLoS One, 16*(12), Article e0260290. https://doi.org/10.1371/journal.pone.0260290

Ma, T. J., & Atkin, D. (2017). User generated content and credibility evaluation of online health information: A meta analytic study. *Telematics and Informatics, 34*(5), 472–486. https://doi.org/10.1016/j.tele.2016.09.009

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv. https://doi.org/10.48550/arXiv.1801.00631.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv*. https://doi.org/10.48550/arXiv.2002.06177

Miles, O., West, R., & Nadarzynski, T. (2021). Health chatbots acceptability moderated by perceived stigma and severity: A cross-sectional survey. *Digital Health, 7*. https://doi.org/10.1177/20552076211063012, 20552076211063012.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. London: Penguin UK.

Murphy, J. J. (1981). *Rhetoric in the middle ages: A history of rhetorical theory from saint augustine to the renaissance*. Berkeley, CA: University of California Press.

Nahon, K., & Hemsley, J. (2013). *Going viral*. Polity Publishers.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Noar, S. M., Rohde, J. A., Prentice-Dunn, H., Kresovich, A., Hall, M. G., & Brewer, N. T. (2020). Evaluating the actual and perceived effectiveness of e-cigarette prevention advertisements among adolescents. *Addictive Behaviors, 109*, 106473. https://doi.org/10.1016/j.addbeh.2020.106473

O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory, 13*(3), 251–274. https://doi.org/10.1111/j.1468-2885.2003.tb00292.x

O'Keefe, D. J. (2015). *Persuasion: Theory and research*. Sage Publications.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1955). *The measurement of meaning*. University of Illinois Press.

Ozawa, S., Wonodi, C., Babalola, O., Ismail, T., & Bridges, J. (2017). Using best-worst scaling to rank factors affecting vaccination demand in northern Nigeria. *Vaccine, 35*(47), 6429–6437. https://doi.org/10.1016/j.vaccine.2017.09.079

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* (pp. 1–24). New York: Springer.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Prasetyo, Y. T., Dewi, R. S., Balatbat, N. M., Antonio, M. L. B., Chuenyindee, T., Perwira Redi, A. A. N., … Kurata, Y. B. (2021). The evaluation of preference and perceived auality of health communication icons associated with COVID-19 prevention measures. *Healthcare, 9*(9). https://doi.org/10.3390/healthcare9091115

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

Priester, J. R., & Petty, R. E. (1995). Source attribution and persuasion: Perceived honesty as a determinant of message scrutiny. *Personality and Social Psychology Bulletin, 21*, 637–654. https://doi.org/10.1177/0146167295216010

Priester, J. R., & Petty, R. E. (2003). The influence of spokesperson trustworthiness on message elaboration, attitude strength, and advertising effectiveness. *Journal of Consumer Psychology, 13*(4), 408–421. https://doi.org/10.1207/S15327663JCP1304_08

Ragot, M., Martin, N., & Cojean, S. (2020). In *AI-generated vs. human artworks. a perception bias towards artificial intelligence? Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–10). https://doi.org/10.1145/3334480.3382892

Ratneshwar, S., & Chaiken, S. (1991). Comprehension's role in persuasion: The case of its moderating effect on the persuasive impact of source cues. *Journal of Consumer Research, 18*(1), 52–62. https://doi.org/10.1086/209240

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv. https://doi.org/10.48550/arXiv.1908.10084.

Rohde, J. A., Noar, S. M., Prentice-Dunn, H., Kresovich, A., & Hall, M. G. (2021). Comparison of message and effects perceptions for the Real Cost e-cigarette prevention ads. *Health Communication, 36*(10), 1222–1230. https://doi.org/10.1080/10410236.2020.1749353

Rollins, B., Huh, J., Bhutada, N., & Perri, M. (2021). Effects of endorser type and testimonials in direct-to-consumer prescription drug advertising (DTCA). *International Journal of Pharmaceutical and Healthcare Marketing, 15*(1), 1–17. https://doi.org/10.1108/IJPHM-06-2019-0042

Roy, S., Jain, V., & Rana, P. (2013). The moderating role of consumer personality and source credibility in celebrity endorsements. *Asia-Pacific Journal of Business Administration, 5*(1), 72–88. https://doi.org/10.1108/17574321311304549

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Hoboken: Prentice Hall.

Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv. https://doi.org/10.48550/arXiv.2211.05100.

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

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

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied, 29*(3), 676. https://doi.org/10.1037/xap0000447

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slovic, P. (1987). Perception of risk. *Science, 236*(4799), 280–285. https://doi.org/10.1126/science.3563507

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin, 39*(2), 193–205. https://doi.org/10.1177/0146167212472374

snscrape: A social networking service scraper in Python. (2021). Github. https://github.com/JustAnotherArchivist/snscrape.

Sülflow, M., Schäfer, S., & Winter, S. (2019). Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *New Media & Society, 21*(1), 168–190. https://doi.org/10.1177/1461444818791520

Tannenbaum, P. H. (1956). Initial attitude toward source and concept as factors in attitude change through communication. *Public Opinion Quarterly, 20*(2), 413–425. https://doi.org/10.1086/266638

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26*(3), 249. https://doi.org/10.1037/h0070363

Tubaishat, A. (2014). An investigation into the attitudes of nursing students toward technology. *Journal of Nursing Research, 22*(2), 119–125. https://doi.org/10.1097/jnr.0000000000000029

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural language processing with Transformers*. O'Reilly Media, Inc.

Van Der Heide, B., & Lim, Y. S. (2016). On the conditional cueing of credibility heuristics: The case of online influence. *Communication Research, 43*(5), 672–693. https://doi.org/10.1177/0093650214565915

Vaping Prevention Resource. (2023). *Vaping prevention communication: Evidence-based practices*. https://vapingprevention.org/what-works.

Villanti, A. C., LePine, S. E., West, J. C., Cruz, T. B., Stevens, E. M., Tetreault, H. J., & Mays, D. (2021). Identifying message content to reduce vaping: Results from online message testing trials in young adult tobacco users. *Addictive Behaviors, 115*, 106778. https://doi.org/10.1016/j.addbeh.2020.106778

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wang, T. W., Gentzke, A. S., Neff, L. J., Glidden, E. V., Jamal, A., Park-Lee, E., … Hacker, K. A. (2021). Characteristics of e-cigarette use behaviors among US youth, 2020. *JAMA Network Open, 4*(6), Article e2111336-e2111336. https://doi.org/10.1001/jamanetworkopen.2021.11336

Wang, Y., Xu, Y. A., Wu, J., Kim, H. M., Fetterman, J. L., Hong, T., & McLaughlin, M. L. (2023). Moralization of e-cigarette use and regulation: A mixed-method computational analysis of opinion polarization. *Health Communication, 38*(8), 1666–1676. https://doi.org/10.1080/10410236.2022.2027640

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., Le, Q, V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science, 21*, 101–112. https://doi.org/10.1007/BF02894421

Winter, S., & Krämer, N. C. (2012). Selecting science information in Web 2.0: How source cues, message sidedness, and need for cognition influence users' exposure to blog posts. *Journal of Computer-Mediated Communication, 18*(1), 80–96. https://doi.org/10.1111/j.1083-6101.2012.01596.x

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2309.01219

Zhou, S., Silvasstar, J., Clark, C., Salyers, A. J., Chavez, C., & Bull, S. S. (2023). An artificially intelligent, natural language processing chatbot designed to promote COVID-19 vaccination: A proof-of-concept pilot study. *Digital Health, 9*. https://doi.org/10.1177/20552076231155679, 20552076231155679.