The effect of source disclosure on evaluation of AI-generated messages: A two-part study

Sue Lim[a,1], Ralf Schmälzle[a]

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

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 people's 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 participants' negative attitudes towards AI. We found a significant moderating effect of negative attitudes towards AI on message evaluation, but not for message selection. However, for those with moderate levels of negative attitudes towards AI, source disclosure decreased the preference for AI-generated messages. 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.

Keywords:
Artificial Intelligence (AI), large language model (LLM), health communication, source disclosure, vaping prevention, mixed effects modeling

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 and 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, 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 and 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 and Schmälzle, 2023; Schmälzle and Wilcox, 2022; Zhou et al., 2023). As LLMs continue to expand in these capabilities (Bubeck et al., 2023), we can expect to see LLMs being used as tools for generating persuasive 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 (von Eschenbach, 2021) and mild

[1]Corresponding Author. Email: limsue@msu.edu

to moderate aversion (Castelo and 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 and 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 (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. For instance, 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 and Sherrell, 1993). With the advancement of technology, research also examined source effects in online settings (Ismagilova et al., 2020; Ma and Atkin, 2017). In addition, some of the most well-known theories within communication have examined cognitive mechanisms of source effects (ELM; Petty and Cacioppo (1986); HSM; Chen and 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. Since there's already been evidence that LLMs have the potential to be powerful tools in expanding health communication theory and augmenting health campaign practice, it is thus important to investigate how people's perception of AI influences people's evaluation of health campaign messages. Moreover, it will also be critical to identify potential moderators of such influence.

This paper presents two experimental studies that shed light on the influence of source disclosure on the evaluation of prevention messages (see Figure 1). For the first study (study 1), we conducted an experimental study examining how source disclosure influenced people's 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) inspected how the influence of source disclosure varied on the basis of people's general attitudes toward AI. The findings from our studies have the potential to augment source effects theory within mediated health communication 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. We selected vaping prevention as a health context to examine the evaluation of messages coming from AI source [1].

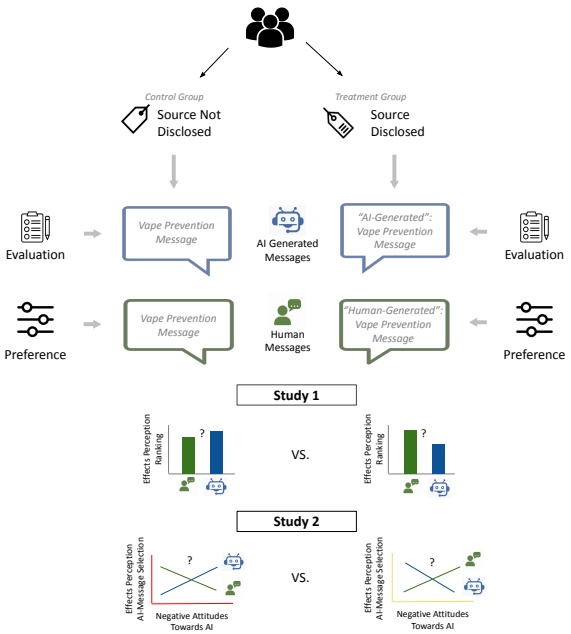

Figure 1: Conceptual Diagram of Study Design

2.1. *Vaping Prevention as Context to Examine the Source Effects of AI*

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 and 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 and 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 and Schmälzle, 2023).

[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.

2.2. The Current Study and Hypotheses

The current study examined how human participants respond to vaping prevention 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).

2.2.1. Effects Perception Ratings as Measure of Evaluation

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 used effects perception ratings as people's measure of the perceived effectiveness of the messages.

Since the influence of source disclosure is a relatively new area of research, to our knowledge, only one study specifically examined how source disclosure would impact people's ratings of health campaigns messages at the time of writing this manuscript. Karinshak et al. (2023) 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.

Our study had a few aspects that differed from Karinshak et al. (2023) study. For one, our comparison of human-generated messages were tweets, to take into account that much discussion about vaping occurs via social media platforms such as Twitter (Lyu et al., 2021; Wang et al., 2023). Second, we used effects perception measure specifically (rather than the general perceived message effectiveness) as a measure of message evaluation. Still, as existing literature suggests the existence of negative 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.

2.2.2. Ranking as Measure of Preference

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 and Ronaldson, 2012). Rankings have been used extensively in the social sciences to gather data about constructs such as values (Abalo et al., 2007; Alwin and 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). Though we do not know of any work that examined the influence of source disclosure on people's ranking of AI-generated vs. human-generated messages, we still predict that the negative bias against AI-generated messages will be exhibited in the ranking of the messages. Thus, we pose the following hypothesis:

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

2.3. Method

We pre-registered our hypotheses and procedures at as.predicted.

2.3.1. Participants

A total of 151 young adults (18-24 years of age) were recruited from two study pools and either received course credit (University study pool) or $2.80 (Prolific; Palan and Schitter (2018)) as compensation for participating in the study. We specifically selected the young adult age group because of the prevalence of vaping in this age demographic (Boakye et al., 2022). The local review board approved the study. We discarded the data from nine participants who did not complete the study or who completed the study in under five minutes, leaving 142 participants (m_{age} = 20.78, sd_{age} = 1.78]; 59% women) in the final dataset. Power calculations conducted a priori using the WebPower package in R (Zhang and Yuan, 2018) for a mixed ANOVA, with a medium effect size (f = 0.25) and significance level α = .05, showed that a total sample size of around 130 (about 65 per group) was enough to detect significant differences between groups at the power level of 0.8.

2.3.2. 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 and Schmälzle, 2023). For details, see 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 (JustAnotherArchivist, 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.

2.3.3. 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.

2.3.4. Measures

Study 1 included two main measures. First, we adopted and updated UNC's perceived message effects, otherwise named effects perceptions (EP), scale (Baig et al., 2019) to fit vaping. 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 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). 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. Finally, the participants answered demographic questions including age.

2.3.5. Data Analysis

All analyses were conducted in R. To examine H1, the responses for the four 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 conducted a mixed ANOVA that examined the influence of source disclosure (disclosed vs. undisclosed) and the message source (AI vs. human) on EP.

For the statistical difference in the mean ranks between the groups, 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 α = .05 to test for significance for both mixed ANOVA 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 and Hornik, 2011; Honnibal and Montani, 2020; Hutto and Gilbert, 2014; Reimers and 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 Appendix B for the results of the supplementary analysis.

2.3.6. 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 beyond the University sample and to increase the sample size. 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, for the rank data, we used the Wilcoxon test, which is a two-sample extension of the Kruskal-Wallis test.

2.4. Results

First, we present the results from the mixed ANOVA, which tested the influence of source disclosure on message ratings (see Table 1). We find that there was a significant interaction effect between source disclosure and the message source ($F(1,140)$ = 4.73, η^2 = .0018, p = .031). As illustrated in Figure 2, this interaction was due to the fact the difference between the AI-generated and human-generated messages was smaller when the source was disclosed compared to when it was not disclosed. This interaction qualified a main effect of message source ($F(1,140)$ = 10.25, η^2 = .0039, p = .0017), which indicated overall lower ratings for human-generated compared to AI-generated messages. Follow-up comparisons conducted separately for each message source (i.e. AI-generated and human-generated messages) revealed that the EP ratings for AI-generated messages were slightly lower and ratings for human-generated messages were slightly higher when the source was

disclosed, yet this difference was not statistically significant ($t(133)$ = .82; p > .05 for AI-generated messages; $t(125)$ = -.23; p > .05 for human-generated messages; see Table 2). Thus, H1 was partially supported.

Table 1: Influence of Source Disclosure on EP Scores

	F-score	***p-value***	η^2
SD: Disclosed (vs. Not Disclosed)	.072	.79	.00049
MS: Human (vs. AI)	10.25	**.0017**	.0039
SD: Disclosed & MS: Human	4.73	**.031**	.0018

Note. SD = Experimental Group; MS = Message Source.

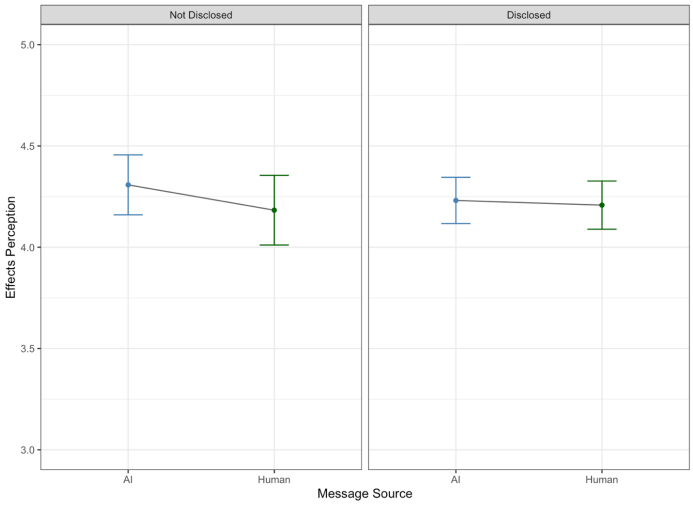

Figure 2: Mean Effects Perception Scores by Experimental Condition.

Table 2: The Effect of Source Disclosure by Message Source

	Mean (Standard Deviation)		t-score *(p-value)*
	Source Not Disclosed	Source Disclosed	
AI-Generated	4.31 (.63)	4.23 (.48)	.82 *(.41)*
Human-Generated	4.18 (.73)	4.21 (.50)	-.23 *(.82)*

To test H2, we compared the difference in the mean ranks of AI and human-generated messages (AI mean rank - Human mean rank; see Figure 3) using the Wilcoxon Sum Rank 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 = -.6, was slightly higher than the median difference in rank for participants who did not know the source, Mdn = -.87, though this difference was not statistically significant (W = 2652.5, p > .05).

2.5. 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

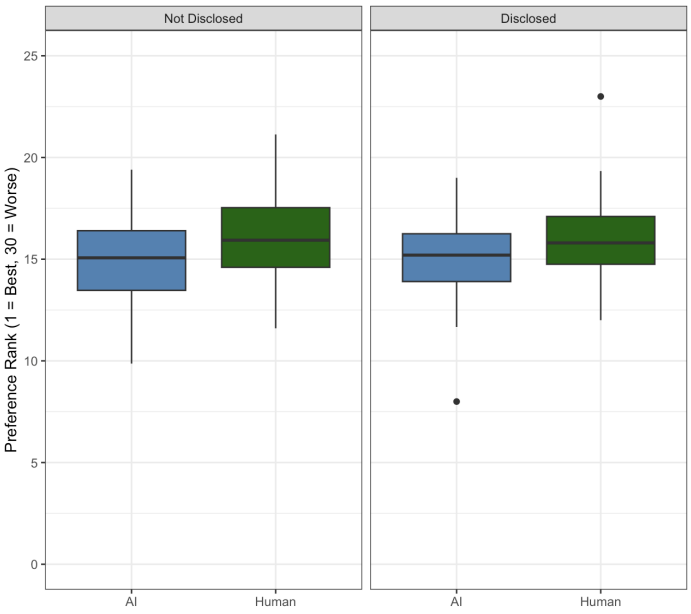

Figure 3: Mean Rank Difference Scores by Experimental Condition.

H1 was partially supported – source disclosure significantly decreased the ratings difference between AI and human-generated messages. However, follow-up mean comparisons by message source showed that ratings stayed statistical consistent between non source disclosure and source disclosure conditions. This finding is generally aligned with findings from Karinshak et al. (2023). However, 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. 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 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 Moderators

Schepman and Rodway (2023) created a scale about general attitudes toward AI (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. Adopting the negative attitudes towards AI subscale of GAAIS, we posited the following hypotheses:

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

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

As with study 1, we used two platforms to recruit participants, one administered by the university and the other by Prolific. A total of 216 adults recruited from the study pools either received course credit (University study pool) or $2.80 (Prolific; Palan and Schitter (2018)) as compensation for participating in the study. To generalize the findings of Study 1 beyond young adults, we extended the participant pool for the Prolific platform to all adults . The local review board approved the study. We discarded the data from 33 participants who did not complete the study, completed the study in under five minutes, and failed to pass the manipulation check questions, leaving 183 participants (m_{age} = 33.83, sd_{age} = 14.42; 56% women) in the final dataset.

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 the following modification: 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.

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 and Rodway, 2023). The overall mean was 3.04, with a standard deviation of .80. 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. 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 mixed effects linear regression model. The models allowed for the intercept to vary by participant, to take into consideration of the repeated measures design. 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 3), and then fitted a Poisson regression model.

3.3. Results

For the EP ratings, we found that there was a significant three-way interaction of source disclosure, message source (AI vs. Human), and the extent of having a negative attitude toward AI (b = -.14, SE = .047, p = .0029; see Table 3). In other words, the influence of source disclosure on the evaluation of the AI-generated messages vs. human-generated messages differed by the level of negative attitudes towards AI. A deeper inspection of the moderation effect shows that for both AI-generated and human-generated messages, source disclosure led to slightly higher EP ratings among participants with lower levels of negative attitudes towards AI, whereas it led to slightly lower EP ratings among those with higher levels of negative attitudes towards AI (see Table 4). Interestingly, when the source was disclosed, the more negative attitudes the participants had towards AI, the higher they rated the AI-generated messages, whereas the ratings of human-generated messages generally stayed flat (see Figure 4).

Table 3: Results from Mixed Effects Linear Model

Term	***Estimate***	***S.E.***	***t-score***	***p-value***
Intercept	3.32	.30	10.98	**<.001**
SD: Disclosed (vs. Not Disclosed)	.61	.45	1.34	.18
MS: Human (vs. AI)	-.46	.098	-4.69	**<.001**
Negative Attitude Towards AI	.25	.097	2.61	**.0098**
MS: Human & SD: Disclosed	.48	.15	3.26	**.0011**
SD: Disclosed & Negative Attitude	-.20	.14	-1.36	.18
MS: Human & Negative Attitude	.099	.031	3.15	**.0017**
MS: Human & SD: Disclosed & Neg Attitude	-.14	.047	-2.98	**.0029**

Note. SD = Experimental Group; MS = Message Source; S.E. = Standard Error

Table 4: Pairwise Comparison of EP by Negative Attitudes Towards AI

	Disclosed - NonDisclosed	***S.E.***	***z-score***	***p-value***
AI-Generated Messages				
*Negative Attitudes Towards AI = 2.24**	.17	.16	1.04	.30
Negative Attitudes Towards AI = 3.04	.013	.12	.11	.91
Negative Attitudes Towards AI = 3.84	-.14	.16	-.89	.37
Human-Generated Messages				
*Negative Attitudes Towards AI = 2.24**	.34	.16	2.06	**.04**
Negative Attitudes Towards AI = 3.04	.068	.12	.59	.56
Negative Attitudes Towards AI = 3.84	-.20	.16	-1.23	.22

*Note. S.E. = Standard Error; *3.04 = overall mean; 2.24 = mean - 1 standard deviation; 3.84 = mean + 1 standard deviation*

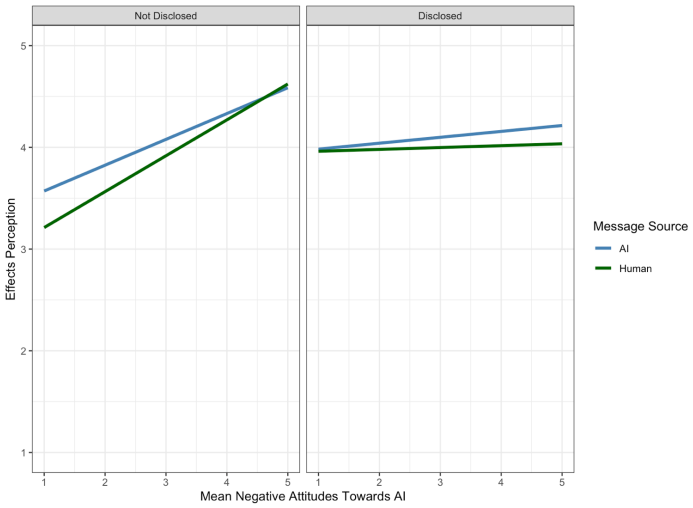

Figure 4: Predicted EP Ratings with Negative Attitudes as Moderator

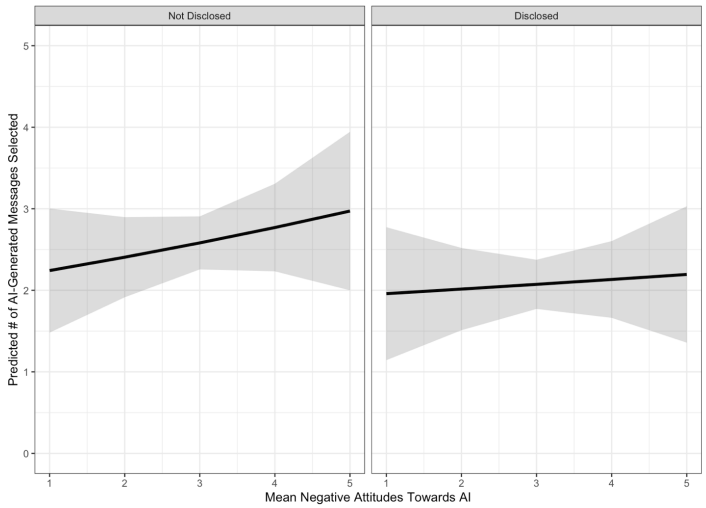

Figure 5: Predicted Number of AI-Generated Messages Selected

Table 5 shows the results for messages selection. There was no moderation effect of negative attitudes toward AI (b = -.042, SE = .12, $p > .05$), and H4 was not supported. A deeper inspection of the results showed that those who knew the source were likely to select less number of AI-generated messages compared to those who did not know the source for those with moderate level of negative attitudes towards AI (see Figure 5 and Table 6).

Table 5: Attitudes Towards AI and Selection of AI-Generated Messages

	b	*S.E.*	*z-score*	*p-value*
Intercept	.74	.25	2.97	**.003**
SD: Disclosed (vs. Not Disclosed)	-.09	.39	-.24	.81
Negative Attitudes Towards AI	.07	.079	.90	.37
SD: Disclosed & Negative Attitudes Towards AI	-.04	.12	-.34	.73

Note. S.E. = Standard Error

Table 6: Pairwise Comparison of AI-Message Selection by Negative Attitudes Towards AI

	Disclosed - NonDisclosed	*S.E.*	*z-score*	*p-value*
Negative Attitudes Towards AI = 2.24∗	-.42	.31	-1.33	.18
Negative Attitudes Towards AI = 3.04	-.51	.23	-2.27	**.023**
Negative Attitudes Towards AI = 3.84	-.62	.33	-1.88	.06

*Note. S.E. = Standard Error; *3.04 = overall mean; 2.24 = mean - 1 standard deviation; 3.84 = mean + 1 standard deviation*

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. Specifically, at lower levels of negative attitudes towards AI, the ratings for AI-generated and human-generated messages were slightly higher when the source was disclosed vs. not disclosed (albeit not statistically significant for AI-generated messages), whereas the opposite was observed at higher levels of negative attitudes towards AI (not statistically significant). This result suggests the existence of a slight bias against AI-generated messages.

However, for the participants who knew the source, the EP ratings for AI-generated messages compared to those for human-generated messages increased with the level of negative attitudes towards AI. While deeper inspection is needed to fully unpack this phenomenon, one explanation could be "source involvement". In other words, the level of negative attitudes towards AI could have determined how closely they examined the messages: those with greater levels of negative attitudes towards AI could have paid closer attention to the content of the messages compared to those with less negative attitudes towards AI. Another explanation is that the negative attitudes towards AI measure could be a bit too general. In the case of AI message generation, people are heavily involved in the process. In the case of AI message generation, people are heavily involved in the process (see Appendix A for details of how the messages used for this study were crafted). Thus, 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.

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 moderate levels of negative attitudes towards AI. These results further provide support for people's preference against AI-generated messages. 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 two studies provide qualified support for our hypotheses. We found that disclosing the source led to lower ratings of AI-generated messages (partially supporting H1 in Study 1) and that negative attitudes toward AI moderated this

effect (supporting H3 in Study 2). Though the analyses of ranking and message selection tasks did not support our hypotheses (H2 in Study 1 and H4 in Study 2), they 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 and Glikson, 2023; 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 and 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 say that there are small but significant effects of source disclosure, consistent with a small preferential treatment for human-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, which 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. 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).

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. The work was additionally supported by the Strosacker Grant from Michigan State University's Health and Risk Communication Center.

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, 115–121.

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, 21–44.

Allem, J., Escobedo, P., Chu, K., Soto, D., Cruz, T., Unger, J., 2017. Campaigns and counter campaigns: reactions on twitter to e-cigarette education. Tobacco Control 26, 226–229.

Alwin, D., Krosnick, J., 1985. The measurement of values in surveys: A comparison of ratings and rankings. Public Opinion Quarterly 49, 535–552.

Asscher, O., Glikson, E., 2023. Human evaluations of machine translation in an ethically charged situation. New Media & Society 25, 1087–1107.

Baig, S., Noar, S., Gottfredson, N., Boynton, M., Ribisl, K., Brewer, N., 2019. Unc perceived message effectiveness: validation of a brief scale. Annals of Behavioral Medicine 53, 732–742.

Bellaiche, L., Shahi, R., Turpin, M., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., Seli, P., 2023. Humans versus ai: whether and why we prefer human-created compared to ai-created artwork. Cognitive Research 8, 1–22.

Berlo, D., 1960. The Process of Communication. Holt, Rinehart, and Winston.

Bigscience, 2022. Bigscience rail license v1.0. https://huggingface.co/spaces/bigscience/license.

Boakye, E., Osuji, N., Erhabor, J., Obisesan, O., Osei, A., Mirbolouk, M., Blaha, M., 2022. Assessment of patterns in e-cigarette use among adults in the us, 2017-2020. JAMA Network Open 5, e2223266–e2223266.

Boynton, M., Sanzo, N., Brothers, W., Kresovich, A., Sutfin, E., Sheeran, P., Noar, S., 2023. Perceived effectiveness of objective elements of vaping prevention messages among adolescents. Tobacco Control 32, e228–e235.

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 preprint arXiv:2303.12712 .

Castelo, N., Ward, A., 2021. Conservatism predicts aversion to consequential artificial intelligence. Plos One 16, e0261467.

Chambers, J., Freeny, A., Heiberger, R., 1992. Analysis of variance; designed experiments, in: Statistical Models in S. Routledge, pp. 145–193.

Chen, S., Chaiken, S., 1999. The heuristic-systematic model in its broader context, in: Dual-process theories in social psychology. The Guilford Press, pp. 73–96.

Clark, E., Jones, C., Williams, J., Kurti, A., Norotsky, M., Danforth, C., Dodds, P., 2016. Vaporous marketing: Uncovering pervasive electronic cigarette advertisements on twitter. PloS One 11, e0157304.

Clark, J., Wegener, D., Habashi, M., Evans, A., 2012. Source expertise and persuasion: The effects of perceived opposition or support on message scrutiny. Personality and Social Psychology Bulletin 38, 90–100.

Collins, L., Glasser, A., Abudayyeh, H., Pearson, J., Villanti, A., 2019. E-cigarette marketing and communication: How e-cigarette companies market e-cigarettes and the public engages with e-cigarette information. Nicotine and Tobacco Research 21, 14–24.

DeWilde, B., 2020. Textacy: Nlp, before and after spacy. https://github.com/chartbeat-labs/textacy. Accessed September 10, 2022.

von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust ai. Philosophy & Technology 34, 1607–1622.

Grummon, A., Hall, M., Mitchell, C., Pulido, M., Sheldon, J., Noar, S., Ribisl, K., Brewer, N., 2022. Reactions to messages about smoking, vaping and covid-19: Two national experiments. Tobacco Control 31, 402–410.

Grün, B., Hornik, K., 2011. topicmodels: An r package for fitting topic models. Journal of Statistical Software 40, 1–30.

Hirschberg, J., Manning, C., 2015. Advances in natural language processing. Science 349, 261–266. doi:10.1126/science.aaa8685.

Honnibal, M., Montani, I., 2020. spacy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303.

Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 216–225.

Ismagilova, E., Slade, E., Rana, N., Dwivedi, Y., 2020. The effect of characteristics of source credibility on consumer behaviour: A meta-analysis. Journal of Retailing and Consumer Services 53, 101736.

Jakesch, M., French, M., Ma, X., Hancock, J., 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 .

JustAnotherArchivist, 2021. snscrape: A social networking service scraper in python. Github. URL: https://github.com/JustAnotherArchivist/snscrape.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R., 2023. Challenges and applications of large language models. arXiv .

Karinshak, E., Liu, S., Park, J., Hancock, J., 2023. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. Proceedings of the ACM on Human-Computer Interaction (CSCW) 7, 1–29.

Lagerkvist, C., 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, 77–88.

Lasswell, H., 1948. The structure and function of communication in society, in: Bryson, L. (Ed.), The Communication of Ideas. New York: Institute for

Religious and Social Studies, pp. 37–51.
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.
Liu, S., Yang, J., 2020. Incorporating message framing into narrative persuasion to curb e-cigarette use among college students. Risk Analysis 40, 1677–1690.
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., 2019. Resistance to medical artificial intelligence. Journal of Consumer Research 46, 629–650.
Luger, G., 2005. Artificial Intelligence: Structures and strategies for complex problem solving. London: Pearson Education.
Lyu, J., Luli, G., Ling, P., 2021. Vaping discussion in the covid-19 pandemic: An observational study using twitter data. PloS One 16, e0260290.
Ma, T., Atkin, D., 2017. User generated content and credibility evaluation of online health information: A meta analytic study. Telematics and Informatics 34, 472–486.
Marcus, G., 2018. Deep learning: A critical appraisal. arXiv .
Marcus, G., 2020. The next decade in ai: four steps towards robust artificial intelligence. arXiv URL: 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, 20552076211063012.
Mitchell, M., 2019. Artificial Intelligence: A guide for thinking humans. Penguin UK, London.
Murphy, J.J., 1981. Rhetoric in the Middle Ages: A history of rhetorical theory from Saint Augustine to the Renaissance. University of California Press, Berkeley, CA.
Nahon, K., Hemsley, J., 2013. Going viral. Polity.
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.
O'Keefe, D.J., 2015. Persuasion: Theory and research. Sage Publications.
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, 6429–6437.
Palan, S., Schitter, C., 2018. Prolific. ac—a subject pool for online experiments. Journal of Behavioral and Experimental Finance 17, 22–27.
Petty, R.E., Cacioppo, J.T., 1986. The elaboration likelihood model of persuasion, in: Springer New York, pp. 1–24.
Pornpitakpan, C., 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. Journal of Applied Social Psychology 34, 243–281.
Prasetyo, Y.T., Dewi, R.S., Balatbat, N.M., Antonio, M.L.B., Chuenyindee, T., Perwira Redi, A.A.N., Young, M.N., Diaz, J.F.T., Kurata, Y.B., 2021. The evaluation of preference and perceived quality of health communication icons associated with covid-19 prevention measures. Healthcare 9, 1115.
Ragot, M., Martin, N., Cojean, S., 2020. Ai-generated vs. human artworks. a perception bias towards artificial intelligence?, in: Extended abstracts of the 2020 CHI conference on human factors in computing systems, pp. 1–10.
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, 52–62.
Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv URL: http://arxiv.org/abs/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, 1222–1230.
Russell, S., Norvig, P., 2021. Artificial intelligence: A modern approach 4th Edition. Prentice Hall, Hoboken.
Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Manica, M., 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint 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, 2724–2741.
Schmälzle, R., Wilcox, S., 2022. Harnessing artificial intelligence for health message generation: The folic acid message engine. Journal of Medical Internet Research 24, e28858.
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, 676.
Shannon, C., 1948. A mathematical theory of communication. Bell Systems Technical Journal 27, 379–423.
Slovic, P., 1987. The psychometric paradigm. Science 236, 280–285.
Tunstall, L., von Werra, L., Wolf, T., 2022. Natural language processing with Transformers. O'Reilly Media, Inc.
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.
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, e2111336–e2111336.
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, 1666–1676.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., V. Le, Q., 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.
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 URL: https://doi.org/10.48550/arXiv.2309.01219.
Zhang, Z., Yuan, K.H., 2018. Practical statistical power analysis using Webpower and R. ISDSA Press.
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, 20552076231155679.