Journal of Personality and Social Psychology

Truth Over Falsehood: Experimental Evidence on What Persuades and Spreads

Nicolas Fay, Keith J. Ransom, Bradley Walker, Piers D. L. Howe, Andrew Perfors, and Yoshihisa Kashima

Online First Publication, September 29, 2025. https://dx.doi.org/10.1037/pspa0000467

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Personality and Social Psychology:
Attitudes and Social Cognition

ISSN: 0022-3514

https://doi.org/10.1037/pspa0000467

Truth Over Falsehood: Experimental Evidence on What Persuades and Spreads

Nicolas Fay[1], Keith J. Ransom[2], Bradley Walker[1], Piers D. L. Howe[3], Andrew Perfors[3], and Yoshihisa Kashima[3]

[1] School of Psychological Science, The University of Western Australia
[2] School of Computer and Mathematical Sciences, The University of Adelaide
[3] School of Psychological Sciences, The University of Melbourne

The English poet John Milton portrayed truth as a powerful warrior capable of defeating falsehood in open combat. The spread of false information online suggests otherwise. Here, we test the persuasive power and transmission potential of true versus false messages in a controlled experimental setting, free from the effects of social media algorithms and bot amplification. Across four experiments (combined $N = 4{,}607$), we tested how perceived veracity affects message persuasion and shareability, using messages generated by both humans and large language models. Experiments 1 and 2 (persuasion game) involved participants creating and evaluating persuasive messages; Experiments 3 and 4 (attention game) focused on messages optimized to capture attention. Our findings consistently show that messages created with the intention of being truthful were more persuasive and more likely to be shared than those designed to be false. While perceived message truth was the main driver of persuasion, message transmission was primarily driven by positive emotion and social engagement, indicating that social connection is prioritized during information sharing. These results suggest that truth holds a competitive edge in the marketplace of ideas.

Statement of Limitations

The present research examines how message veracity influences persuasion and sharing within a controlled experimental setting. Given the mixed findings returned by computational social science studies using social media data, we prioritized internal validity. A primary limitation of our experiments is that the generalizability of our findings to different populations, contexts, platforms, or real-world environments remains uncertain. However, given the fundamental role of message veracity in effective decision making, we anticipate that our findings may extend across diverse populations. We also acknowledge that persuasion and message transmission depend on factors beyond message content, such as source credibility. Future research should explore these moderating factors to better understand the conditions under which truth wins.

Keywords: true, false, persuasion, attention, transmission

Supplemental materials: https://doi.org/10.1037/pspa0000467.supp

Mandy Huetter served as action editor.

Nicolas Fay https://orcid.org/0000-0001-9866-2800

Data, analytic code, study materials, and Supplemental Materials are available on the Open Science Framework, and preregistration documents are accessible on AsPredicted. The authors have no competing interests to declare. Nicolas Fay, Andrew Perfors, Piers D. L. Howe, and Yoshihisa Kashima received funding from the Office of National Intelligence and Australian Research Council Grant NI210100224.

Conceptualization: Nicolas Fay, Andrew Perfors, Piers D. L. Howe, and Yoshihisa Kashima; methodology: Nicolas Fay, Andrew Perfors, Piers D. L. Howe, Yoshihisa Kashima, Keith J. Ransom, and Bradley Walker; investigation: Keith J. Ransom and Bradley Walker; visualization: Nicolas Fay Funding; acquisition: Nicolas Fay, Andrew Perfors, Piers D. L. Howe, and Yoshihisa Kashima; project administration: Keith J. Ransom and Bradley Walker; writing–original draft: Nicolas Fay; writing–review and editing: Nicolas Fay, Andrew Perfors, Piers D. L. Howe, Yoshihisa Kashima, Keith J. Ransom, and Bradley Walker.

Nicolas Fay played a lead role in conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, visualization, and writing–original draft. Keith J. Ransom played a lead role in software and a supporting role in conceptualization, data curation, investigation, methodology, project administration, and writing–review and editing. Bradley Walker played a supporting role in conceptualization, investigation, and writing–review and editing. Piers D. L. Howe played a supporting role in conceptualization, funding acquisition, and writing–review and editing. Andrew Perfors played a supporting role in conceptualization, funding acquisition, and writing–review and editing. Yoshihisa Kashima played a supporting role in conceptualization, funding acquisition, and writing–review and editing.

Correspondence concerning this article should be addressed to Nicolas Fay, School of Psychological Science, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia. Email: nicolas.fay@gmail.com

> Let her [Truth] and Falsehood grapple; who ever knew Truth put to the worse, in a free and open encounter? (Milton, 1868)

In his defense of freedom of speech, the English poet John Milton portrayed truth as a powerful warrior who can defeat falsehood in open combat (Milton, 1868). The impact and spread of false information through online media suggests otherwise. False information has undermined public health (Pertwee et al., 2022), delayed climate action (Treen et al., 2020), eroded trust in institutions (Green et al., 2022), and manufactured societal problems that threaten the foundation of liberal democracies (Lewandowsky et al., 2023). False information has also been found to spread farther, faster, deeper, and more broadly (via resharing) on the social media platform Twitter (now X; Vosoughi et al., 2018). This may be due to its greater novelty and ability to elicit powerful negative emotions. Because falsehoods are not constrained by reality in the way that truths are, they may have an advantage in the marketplace of ideas (*Abrams vs. United States*, 1919; Dawkins, 1989). Concerns around the influence and spread of false information are compounded by evidence indicating that once false information is accepted, it is difficult to correct (Ecker et al., 2022; Walter & Tukachinsky, 2020) and by the potential of large language models (LLMs) to generate and disseminate false information at scale (Verma, 2023).

Truth matters; it is foundational to epistemology—the branch of philosophy concerned with the nature, origin, and limits of human knowledge (Steup & Neta, 2005)—and is critical to forming accurate beliefs and making effective decisions (Brownson et al., 1999; Savage, 1951). Fact-checking—the process of verifying the accuracy of information, statements, and claims—reflects epistemology in action. In this context, fact-checked posts on the Reddit platform that were rated as true are associated with stronger user engagement (volume of user comments and conversation length) than fact-checked posts rated as false (Bond & Garrett, 2023). So, the engagement patterns observed on Reddit differ from those observed on Twitter, where false information was found to attract stronger user engagement. This suggests that these differences may be due to platform-specific factors rather than an inherent human preference for true or false information. The present study tests a fundamental aspect of human nature—people's preference for true versus false information. This is done within a controlled experimental environment that is unaffected by the choice architecture and recommendation algorithms of social media platforms, as well as the bots that can amplify certain viewpoints (Orabi et al., 2020; Stella et al., 2018). In other words, the experiments reported prioritize internal validity over external or ecological validity.

Of course, how a message is interpreted is only part of the equation. Mass social influence requires two key elements: Messages must be persuasive, which depends on how recipients evaluate their content, and they must be received in the first place, which relies on successful transmission. Persuasion cannot occur without transmission. Likewise, merely receiving a message is not enough to exert influence—it must also prompt a favorable evaluation (Briñol & Petty, 2012; Greenwald, 1968; McGuire, 1985). The distinction between persuasion and transmission has parallels in the psychological literature, which has investigated them in two relatively independent streams of research. On the one hand, people's responses to messages—such as belief updating—are shaped by their persuasive power and personal relevance, as well as contextual factors (Druckman, 2022; Petty & Cacioppo, 1986). On the other hand, message transmission may occur with minimal cognitive reflection (Pennycook & Rand, 2020) and instead be driven by social and emotional motives—such as the desire to connect with others, express one's identity, or gain social approval (Berger & Milkman, 2012; Heath et al., 2001; Kashima et al., 2020). Although a message's impact depends on both its persuasiveness and spread, measures of engagement on social media primarily reflect spread. As such, the perceived truthfulness of a message may relate differently to its transmissibility than to its persuasive influence—a possibility tested in the studies reported here. We also systematically examine what other factors, beyond veracity, contribute to a message's impact. Recognizing that real-world messages vary across multiple dimensions, we measured eight psychological message attributes and investigated their relationship to two key outcomes: message influence (operationalized as persuasion and belief updating) and message spread (operationalized as self-reported intention to share the message online or offline). This multidimensional approach allowed us to identify the psychological features that most strongly predict a message's persuasive force and its likelihood of being transmitted, offering insights into the cognitive, affective, and social drivers of informational influence.

Experiments 1 and 2 report the findings of the persuasion game. In Experiment 1 human participants were instructed to write 15 persuasive messages, each supporting a different claim (e.g., prisoners should be forced to undertake manual labor) under one of three conditions: when instructed to produce true messages (i.e., messages they believe to be true), when instructed to produce false messages (i.e., messages they believe to be false), or messages unconstrained by veracity. A second group of participants then rated the messages across a range of dimensions, including persuasiveness and willingness to share. Experiment 2 replicated this design using persuasive messages generated by an LLM (GPT-3.5), allowing us to assess the robustness of the Experiment 1 findings when the messages originate from a model trained on a vast and diverse corpus of text (see also Acerbi & Stubbersfield, 2023; Bai et al., 2023; Dillion et al., 2023; Howe et al., 2023; Spitale et al., 2023). To examine the extent to which the goal of the message creator influences these outcomes, we ran the attention game in Experiments 3 and 4. In these studies participants were instructed to write messages designed to capture attention. Experiment 3 followed the same design as Experiment 1, using human-generated messages, while Experiment 4 used attention-grabbing messages generated by an LLM.

Across all four experiments, we found that messages generated to be true were also perceived as more truthful, and perceived truth was the primary driver of both persuasiveness and belief change. Although true messages were shared more often than false ones, the main drivers of message transmission were social in nature, with sharing intentions more strongly associated with positive emotion and perceived social engagement.

Method

Each experiment received approval from The University of Adelaide Ethics Committee. Participants viewed an information sheet before giving consent to take part in the experiment. All methods were performed in accordance with the guidelines from the National Health and Medical Research Council/Australian Research Council/University Australia's National Statement on Ethical Conduct in Human Research.

General Methodology

People share a variety of information with others—everything from their personal views on current events, to social issues, to politics, to sports, and to pop culture. This information varies in its degree of truthfulness, ranging from outright falsehoods to misleading statements, half-truths, mostly accurate claims but with minor inaccuracies, to entirely truthful content. That is, truthfulness is not always binary and cannot always be fact-checked (e.g., in the case of personal opinions). The present study reflects this complexity by eliciting a large set of messages from both humans and LLMs (5,469 unique messages in total) that vary in both the intent behind their creation (truthful or not) and how truthful they are perceived to be by others. This allows us to test the extent to which a message's intended and perceived truth (along with a range of other dimensions) affects its persuasive force and transmission potential.

While the experiments reported rely on intended and perceived truth, one question is whether these more subjective measures of truth meaningfully align with objective or "ground" truth, insofar as it can be measured. Because hand checking 5,469 messages was prohibitive, we trained two LLMs (GPT-4 and Perplexity) to assess the objective truth of each message in Experiment 1 (see Supplemental Material 1 for the full analysis). Perplexity incorporates web searches to retrieve up-to-date information from trusted sources, and both models were first validated against a data set with a known ground truth. We then examined the degree of correspondence between the LLM-generated truth ratings and the human ratings. Human ratings were strongly correlated with those from GPT-4 ($r = .57$) and Perplexity ($r = .50$), suggesting that people's subjective evaluations aligned with our best measure of objective truth. This finding aligns with meta-analytic evidence indicating that people are generally effective at distinguishing between true and false news (Pfänder & Altay, 2025).

Experiments 1 and 2: The Persuasion Game

In Experiments 1 and 2, the task was to design persuasive messages. This was incentivized by offering a $100 reward to the participant who produced the most persuasive message in Experiment 1 (in addition to the payment for participation). In Experiment 1 human participants wrote 15 persuasive messages that supported 15 different claims under one of three conditions: when instructed to produce true messages (i.e., messages they believed to be true), when instructed to produce false messages (i.e., messages they believed to be false), or when unconstrained by message veracity (i.e., they were told they could use true and/or false information). We call this group the human producers. The human-produced messages were then evaluated by a second group of human participants who rated each message on a range of dimensions. We call this group the human raters.

In Experiment 2 the messages were produced by an LLM (GPT-3.5). The LLM was prompted to write 15 persuasive messages that supported the same 15 claims used in Experiment 1. Here we focused on the two conditions of primary interest: the true and false conditions. The LLM-produced messages were then evaluated by a third group of human participants who rated each message across a range of dimensions. To ensure consistency in ratings across the human and LLM-produced messages, 50% of the messages evaluated by the raters were sampled from the human producers in Experiment 1.

Participants

Experiment 1: Human Producers

A total of 285 participants were recruited as message producers through Amazon Mechanical Turk. Users were eligible to participate if they had previously passed a qualification study designed to test their English proficiency. A total of 116 participants self-identified as female, 165 as male, one as nonbinary, and one as trans female (the remainder chose not to provide gender information). Participants were aged 22–72 years ($M = 38.88$, $SD = 11.07$). Most participants were based in the United States (81%), and most self-reported being native English speakers (73%) or fluent English speakers (24%). Most message producers self-identified as White (63%, then 11% Asian, 8% Latinx, 6% Black) and were college educated (68%), politically progressive (58%; 23% conservative), and frequent social media users (87% were daily users). Participants were randomly assigned to the experimental conditions (true, false, unconstrained), with the allocation structured to ensure an equal number of participants in each condition ($N = 95$). Each participant was paid $5.50 for approximately 25–35 min of work (median duration = 29 min).

Experiment 1: Human Raters

A total of 1,710 participants were recruited as message raters through Amazon Mechanical Turk, having previously passed an English-proficiency qualification study. To reduce overall costs, the sample size was determined to ensure that nine out of 15 messages were sampled from each producer, and each message was rated by 10 independent raters. Messages were allocated to the raters using a randomized block design, which ensured that each rater saw one message for each of the 15 claim types. Each of the 15 messages allocated to each rater was sampled from a different producer and included five messages from each of the three experimental conditions (true, false, unconstrained). A total of 785 participants self-identified as female, 909 as male, eight as nonbinary, four as trans female, and four as trans male. Participants were aged 19–99 years ($M = 40.47$, $SD = 11.63$). Most participants were based in the United States (82%), and most self-reported being native English speakers (77%) or fluent English speakers (20%). Most message raters self-identified as White (68%, then 11% Asian, 7% Black, then 5% Latinx) and were college educated (69%), politically progressive (51%; 29% conservative), and frequent social media users (86% were daily users). Each participant was paid $3.50 for approximately 15–25 min of work (median duration = 18 min).

Experiment 2: LLM Producer (GPT-3.5)

LLM-produced messages were generated using GPT-3.5 Turbo (specifically, "gpt-3.5-turbo-0613") via the Python application programming interface. True- and false-condition messages were generated by instantiating one of the prompt templates shown in Table 1. Background information was included in the prompt to encourage GPT to produce misinformation rather than denying the request on ethical grounds (included in both conditions for consistency). The request parameters "temperature," "top-p sampling," "frequency penalty," and "presence penalty" were all set to 1. LLM-produced messages that exceeded the 280-character limit, were not standalone (i.e., continued a previous message), were empty, or were otherwise the result of a generation error were discarded. Each prompt

Table 1
Prompts Provided to GPT-3.5 to Generate Large Language Model-Produced Messages in Experiment 2

True condition	False condition
Background: We are studying the ways that people generate information and misinformation, and we wish to train a classifier to detect statements that are generated based on misinformation and exaggeration.	Background: We are studying the ways that people generate information and misinformation, and we wish to train a classifier to detect statements that are generated based on misinformation and exaggeration.
Your task: Write eight unrelated tweets each containing a different accurate, truthful, and unexaggerated fact that someone might write to support the claim: {claim}	Your task: Write eight unrelated tweets each containing a different inaccurate, misleading, or exaggerated fact that someone might write to support the claim: {claim}
The tweets should be as persuasive as possible and should contain correct information (information you believe to be true).	The tweets should be as persuasive as possible but must contain false facts (information you believe to be false).

invocation called for eight distinct messages to be produced (as opposed to a single message), as pilot testing indicated that this technique increased message diversity. Each prompt was executed 10 times to create a pool of 80 messages for each of the 15 claim types. For each condition and claim type, a sample of six messages was selected at random from the respective pool of 80 after discarding anomalies, yielding a total of 180 messages ($2 \times 15 \times 6$).

Experiment 2: Human Producers

To create an evenly balanced message pool consisting of 180 human-produced and 180 LLM-produced messages, 180 messages were selected at random from Experiment 1, with the constraint that each message came from a unique producer, and there were matching numbers of each claim type.

Experiment 2: Human Raters

A total of 300 adult participants were recruited as message raters through Amazon Mechanical Turk, having previously passed an English-proficiency qualification. Each person rated 12 messages, comprised of six human-produced messages (three from the true condition and three from the false condition) and six LLM-produced messages (three from the true condition and three from the false condition), sampled such that each rater evaluated 12 of 15 distinct claim types. This sample size ensured that each message was evaluated by 10 independent raters. A total of 114 participants self-identified as female, 182 as male, and one as nonbinary (the remainder chose not to provide gender information). Participants were aged 19–72 years (M = 38.80, SD = 11.01). Most participants were based in the United States (79%), and most self-reported being native English speakers (72%) or fluent English speakers (26%). Most message raters self-identified as White (62%, then 14% Asian, 7% Latinx, 6% Indian, and 5% Black) and were college educated (68%), politically progressive (47%; 32% conservative), and frequent social media users (89% were daily users). Each participant was paid $3.50 for approximately 15–25 min of work (median duration = 18 min).

Materials

Thirty-two claims were developed and pretested. Examples include the following: "Prisoners should be required to undertake manual labor," "Single-use plastic products should be banned" and "Dogs make better pets than cats." The claims were pretested by having 214 human participants rate their agreement with each claim, on a 101-point scale ranging from −50 (*strongly disagree*) to +50 (*strongly agree*), with 0 representing a neutral position. The distribution of agreement scores for each claim was assessed, with a preference to avoid claims that returned a strong consensus (e.g., most participants strongly disagreed with the claim "People should be required to donate 10% of their salary to charity") or showed strong political polarization (e.g., Progressives strongly agreed with the claim "COVID-19 vaccination should be required for school attendance" and Conservatives strongly disagreed with this claim). Fifteen claims were selected for use in the present experiment (see Supplemental Material 2 for details).

Measures

Each message was evaluated on 12 dimensions by each rater. Eight dimensions were treated as predictors: truth, relevant, familiar, interest, interest-if-true, social engagement, positive emotion, and negative emotion. Four dimensions were treated as outcomes: persuasion, belief update, online sharing, and offline sharing. Each dimension, with the exception of belief update, was rated on a 5-point Likert scale, for example, "To the best of your knowledge, how truthful is the post?"—*not at all* (1), *slightly* (2), *somewhat* (3), *very much* (4), and *extremely* (5). Agreement with each claim was rated on a 101-point scale ranging from −50 (*strongly disagree*) to +50 (*strongly agree*), with 0 representing a neutral position. The belief update score was computed for each rater by subtracting their agreement with the claim before reading the associated persuasive message from their agreement with the claim after reading the associated persuasive message. This difference score indicates the extent to which participants updated their beliefs on account of reading the persuasive message.

Task and Procedure

Producers

After giving informed consent, participants were shown an instructions page that explained the key elements of the task: for a series of claims, they would read the claim, indicate their agreement with it, and then write a message designed to persuade others of the claim. In the true condition, participants were told that their messages must be based on correct information (information they believe to be true); in the false condition, they were told that their messages must be based on misinformation (information they believe to be false); and in the unconstrained condition, they were told that their messages may be based on any information they like, regardless of whether they believe it to be true or false. Participants were told that the person who produced the most persuasive messages would be paid a $100 bonus. After the instructions page,

participants were asked three multiple-choice questions to demonstrate their understanding of the instructions (see Supplemental Material 3); they could proceed only if they answered all three questions correctly; otherwise, they were sent back to the instructions page to correct their misunderstanding and try again. Next, participants completed a short demographic questionnaire asking for their age, country of residence, gender, English proficiency, education, race/ethnicity, political orientation, and frequency of social media use. They then proceeded to the main task.

The main task consisted of two pages. On the first page, participants were shown a claim and rated their agreement with it. On the second page, an input box was shown below the claim, containing placeholder text asking the participant to write a persuasive message supporting the claim (see Figure 1). The box was formatted like a social media post, with a person's silhouette as a profile picture and the name "Anonymous Poster." Below the input box was a reminder of the condition instructions (e.g., "Must be based on TRUE information" for the true condition) and an indication of the message's length out of the maximum 280 characters (this was updated as the participant typed their message). There was also a button to bring up an emoji menu, so that participants could add emojis to their message if desired, and a "Submit" button to proceed to the next trial after writing a message (participants were required to write at least three characters to continue). A panel on the right side of the page reminded participants of the instructions (i.e., write a message supporting the claim, with the goal of being as persuasive as possible, and where the message is based on correct information/misinformation/any information they like).

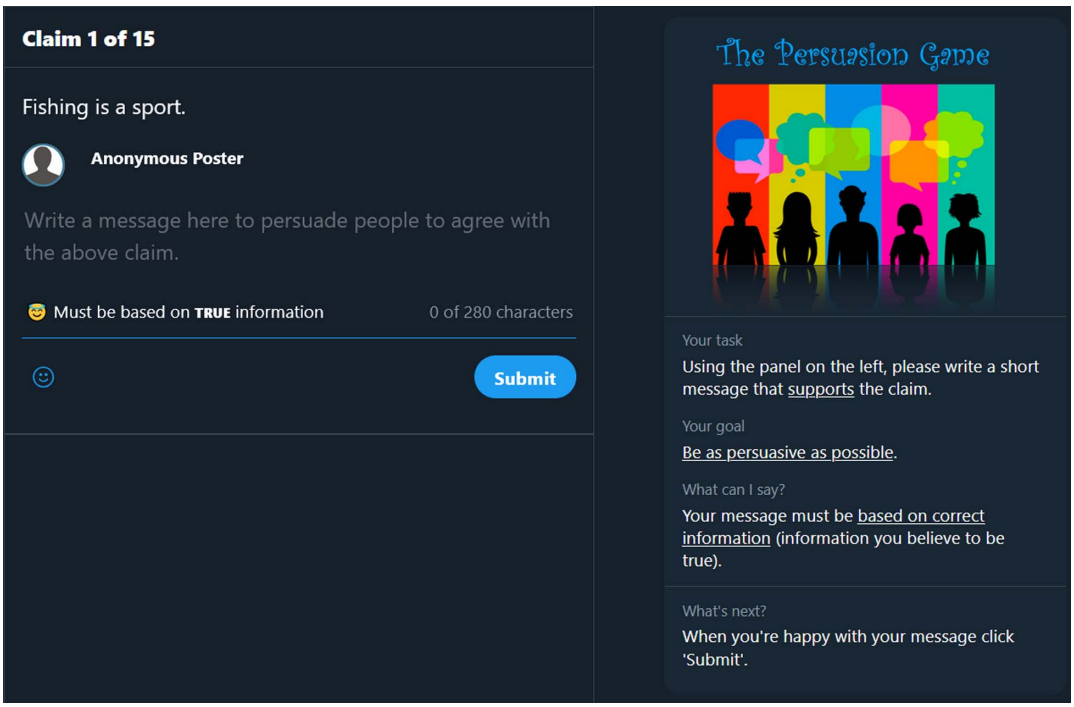
After writing a message for each of the 15 claims, participants were taken to a debriefing page and given a completion code to submit on Mechanical Turk.

Raters

After giving informed consent, participants were shown a series of messages. In each case participants were first shown the claim and were asked to rate their agreement with it (the first belief rating used to measure belief update). They were then shown the message, below the claim in the format of a social media post, with a person's silhouette as a profile picture and the name Anonymous Poster, and the ostensible date and location of the post underneath ("April 2022," "location withheld"). After reading the message, participants again rated their agreement with the claim (the second belief rating). On the next page, they rated the message on the 11 other dimensions (truth, relevant, interest, interest-if-true, familiar, persuasion, social engagement, positive emotion, negative emotion,

Figure 1

A Screenshot From the Experiment 1 Producer Task in the True Condition



Note. In the false condition, the prompt below the message input box read "Must be based on FALSE information" (with a devil emoji), and the panel on the right (in the "What can I say?" section) read "Your message must be *based on misinformation* (information you believe to be false)." In the unconstrained condition, the prompt read "May be based on TRUE or FALSE information" (with a grinning emoji), and the panel read "Your message may be based on *any information you like* regardless of whether you believe it to be true or false." In Experiment 3 the title in the top right was changed to "The Attention Game," and the "Your goal" text was changed to *Gain as much attention as possible*. Screenshots for each condition in each experiment are included in Supplemental Material 4. See the online article for the color version of this figure.

online sharing, offline sharing). The order of the two pages after reading the message was counterbalanced across participants. Participants rated one claim/message at a time, and after rating every message, they were taken to a debrief page and given a completion code.

Statistical Analysis

The data were analyzed using linear mixed-effects modeling (including the backward stepwise regression analyses). The fixed effect (message veracity: true, false, unconstrained) was treatment coded. The random effects structure included by-producer, by-rater, and by-claim random intercepts. This allowed us to account for variations among the producers, the raters, and the claims. All analyses were performed, and all figures were created in R (R Core Team, 2013). Statistical models were estimated using the lmer() function of the lmerTest (Bates et al., 2013; Kuznetsova et al., 2017) package. The statistical analyses were preregistered at https://aspredicted.org/see_one.php, and the data, R Notebooks, and Supplemental Materials are provided on the Open Science Framework (Fay et al., 2024).

Experiments 3 and 4: The Attention Game

In Experiments 3 (human producers) and 4 (LLM producers), the task was to design attention-grabbing messages. This was incentivized by offering a $100 reward to the participant who produced the most attention-grabbing messages in Experiment 3 (in addition to the payment for participation). Aside from this change to the goal, the materials, measures, experimental procedure, and statistical analyses were identical to the persuasion game.

Experiment 3: Human Producers

A total of 285 participants were recruited as message producers through Amazon Mechanical Turk. Users were eligible to participate if they had previously passed a qualification study designed to test their English proficiency. A total of 139 participants self-identified as female, 142 as male, one as nonbinary, and one as trans female (the remainder chose not to provide gender information). Participants were aged 19–73 years ($M = 40.43$, $SD = 10.92$). Most participants were based in the United States (88%), and most self-reported being native English speakers (80%) or fluent English speakers (18%). Most message producers self-identified as White (68%, then 10% Asian, 6% Latinx, 6% Black) and were college educated (65%), politically progressive (49%; 31% conservative), and frequent social media users (91% were daily users). Participants were randomly assigned to the experimental conditions (true, false, unconstrained), with the allocation structured to ensure an equal number of participants in each condition ($N = 95$). Each participant was paid $5.50 for approximately 25–35 min of work (median duration = 26 min).

Experiment 3: Human Raters

A total of 1,710 participants were recruited as message raters through Amazon Mechanical Turk, having previously passed an English-proficiency qualification study. To reduce overall costs, the sample size was determined to ensure that nine out of 15 messages were sampled from each producer, and each message was rated by 10 independent raters. Messages were allocated to the raters using a randomized block design, which ensured that each rater saw one message for each of the 15 claim types. Each of the 15 messages allocated to each rater was sampled from a different producer and included five messages from each of the three experimental conditions (true, false, unconstrained). A total of 740 participants self-identified as female, 905 as male, 28 as nonbinary, seven as trans female, and eight as trans male (the remainder chose not to provide gender information). Participants were aged 18–94 years ($M = 40.96$, $SD = 13.45$). With two exceptions, all participants were based in the United States, and most self-reported being native English speakers (86%) or fluent English speakers (14%). Most message raters self-identified as White (69%, then 12% Black, 8% Asian, 6% Latinx) and were college educated (60%), politically progressive (56%; 24% conservative), and frequent social media users (87% were daily users). Each participant was paid $3.50 for approximately 15–25 min of work (median duration = 18 min).

Experiment 4: LLM Producer (GPT-3.5)

As per Experiment 2, the LLM-produced messages were generated via the Python Application Programming Interface for GPT-3.5 Turbo. The prompts for each condition are given in Table 2.

Experiment 4: Human Producers

Following the same design as Experiment 2, a total of 180 messages were selected at random from Experiment 3, with the constraint that each message came from a unique producer and that there were matching numbers of each claim type.

Table 2
Prompts Provided to GPT-3.5 to Generate Large Language Model-Produced Messages in Experiment 4

True condition	False condition
Background: We are studying the ways that people generate information and misinformation, and we wish to train a classifier to detect statements that are generated based on misinformation and exaggeration. Your task: Write eight unrelated tweets each containing a different accurate, truthful, and unexaggerated fact that someone might write to support the claim: {claim} The tweets should be as attention grabbing and shareable as possible and should contain correct information (information you believe to be true).	Background: We are studying the ways that people generate information and misinformation, and we wish to train a classifier to detect statements that are generated based on misinformation and exaggeration. Your task: Write eight unrelated tweets each containing a different inaccurate, misleading, or exaggerated fact that someone might write to support the claim: {claim} The tweets should be as attention grabbing and shareable as possible but must contain false facts (information you believe to be false).

Experiment 4: Human Raters

A total of 302 adult participants were recruited as message raters through Amazon Mechanical Turk, having previously passed an English-proficiency qualification study. As per Experiment 2, each participant rated 12 messages, comprised of six human-produced messages (three from the true condition and three from the false condition) and six LLM-produced messages (three from the true condition and three from the false condition), sampled such that each person saw 12 out of 15 distinct claim types. This sample size ensured that each message was evaluated by 10 independent raters. A total of 118 participants self-identified as female, 177 as male, one as nonbinary, and one as trans male (the remainder chose not to provide gender information). Participants were aged 18–73 years ($M = 41.70$, $SD = 11.37$). Most participants were based in the United States (85%), and most self-reported being native English speakers (77%) or fluent English speakers (18%). Most message raters self-identified as White (64%, then 14% Asian, 7% Indian, 5% Latinx, 5% Black) and were college educated (70%), politically progressive (49%; 30% conservative), and frequent social media users (87% were daily users). Each participant was paid $3.50 for approximately 15–25 min of work (median duration = 19 min).

Example Messages

Table 3 provides examples of the true- and false-condition messages that were rated as high and low on the persuasion dimension, sampled from the persuasion game in Experiment 1 (human producers) and Experiment 2 (LLM producers), plus examples of the true- and false-condition messages that were rated as high and low on the share online dimension, sampled from the attention game in Experiment 3 (human producers) and Experiment 4 (LLM producers).

Results

Experiment 1: The Persuasion Game (Human Producers)

We first tested how the messages produced under the experimental conditions (true, false, unconstrained) differed across the dimensions of interest. True-condition messages were rated as more truthful than false-condition messages ($p < .001$), confirming the success of the experimental manipulation. True-condition messages were also rated as more relevant, familiar, and interesting and elicited stronger positive emotions than the false-condition messages ($ps < .001$). Importantly, the true-condition messages were also rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline compared with the false-condition messages ($ps < .001$). By contrast, the false-condition messages elicited stronger negative emotions ($p < .001$). The true- and false-condition messages were rated similarly with respect to the interest-if-true and social engagement dimensions ($ps > .814$). For each dimension, the unconstrained-condition messages were rated similarly to the true-condition messages ($ps > .098$) and showed the same pattern of results as the true-condition messages when compared with the false-condition messages. Whereas the true- and unconstrained-condition messages increased belief in the claim (+3.20 and +2.92 points, respectively; $ps < .001$), the false-condition messages decreased belief in the claim (−1.33 points; $p = .026$; see Figure 2 and Table 4).

Next, we examined the relationships between the different dimensions through a correlational analysis (see Figure 3, Panel A). Correlations ranged from negligible ($r = .00$ for negative emotion and familiarity) to strong ($r = .77$ for online sharing and offline sharing), with most dimensions showing moderate positive correlations. We then identified which dimensions best predicted the key outcomes, persuasion and belief update, plus online and offline sharing, using hierarchical backward elimination stepwise regression (see Table 5). For persuasion, the retained dimensions explained 52% of the variance, mostly driven by message truth (36%), positive emotion (+9%), and message interest (+6%). For belief update, 77% of the variance was accounted for, mainly by prior belief in the claim (69%) and message truth (+6%). For online sharing and offline sharing, the retained dimensions explained 40% and 42% of the variance, respectively. In both cases, most of the variance was explained by positive emotion (29%, 26%), social engagement (+4%, +7%), and message interest (+5%, +6%).

Experiment 2: The Persuasion Game (LLM Producers)

The Experiment 2 results replicated the key findings from Experiment 1. LLM-produced true-condition messages were rated by humans as more truthful, familiar, and interesting and elicited stronger positive emotions compared with the false-condition messages ($ps < .032$). Again, the true-condition messages were rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline compared with the false-condition messages ($ps < .031$). LLM-produced true-condition messages increased belief in the claim (+4.59 points; $p < .001$). Conversely, there was no statistical evidence that LLM-produced false-condition messages changed belief in the claim (+1.34 points; $p = .180$). Compared with the human-produced messages sampled from Experiment 1, the LLM-produced messages were rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline ($ps < .034$). This pattern was consistent across the true- and false-condition messages.

The correlation matrix for LLM-produced messages across the different dimensions mirrored that of the human-produced messages (Figure 3, Panel B). The correlation between the coefficients for the human- and LLM-produced messages was $r = .99$[1] (Figure 3, Panel C). Reflecting this strong correlation, the stepwise regression analyses replicated the Experiment 1 results. Persuasion (52% of variance accounted for by the retained dimensions) was mostly driven by message truth (34%), positive emotion (+10%), and message interest (+7%). Belief update (81% of variance) was mostly driven by prior belief in the claim (75%) and message truth (+5%). Online sharing and offline sharing (45% and 49% of the variance, respectively) were mostly driven by positive emotion (29%, 30%) and social engagement (+7%, +8%).

Experiment 3: The Attention Game (Human Producers)

In the attention game, the true-condition messages were rated as more truthful, relevant, familiar, and interesting and elicited stronger positive emotions than the false-condition messages ($ps < .001$).

[1] The correlation between the coefficients for the human raters from Experiments 1 and 2 on the human-generated messages was equally high ($r = 0.99$).

Table 3
Example Messages From Experiments 1–4

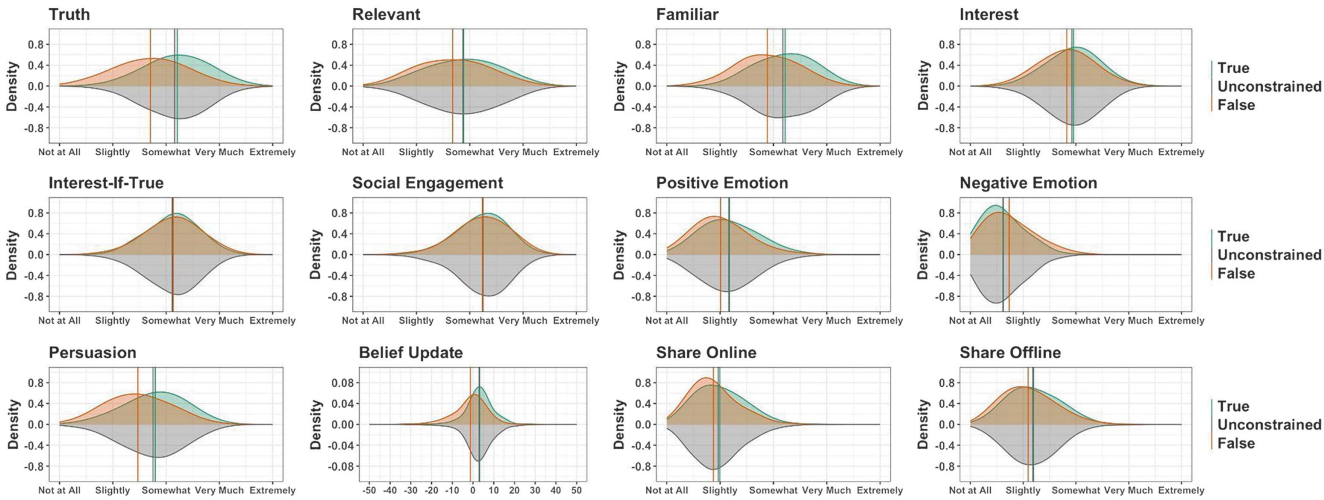
Condition	High persuasion	Low persuasion
	Persuasion game	
Experiment 1: Human producers		
True	Tobacco is a leading cause of lung disease and cancer in the USA. It contains harmful carcinogens and should not be easily accessible. [*Mean persuasion rating = 4.30*]	Smoking tobacco does no good to anyone, and it is unnecessarily killing people and harming secondhand smokers. [*Mean persuasion rating = 2.30*]
False	Medical evidence reports that Tobacco smoking is linked to over 96% of lung cancer cases in adults. Where also passive smoke is responsible for 98% of cases of lung cancer in people under 18 years of age. Tobacco smoking must be banned immediately. [*Mean persuasion rating = 3.80*]	Tobacco smoking causes people to act in ways that are harmfull to society as a whole. They act irrationally and take actions that have no logic because smoking reduces brain power. [*Mean persuasion rating = 1.60*]
Experiment 2: LLM producer		
True	Single-use plastic bags can take up to 500 years to decompose, causing harm both on land and in our oceans. Let's ban them and make reusable options more accessible! [*Mean persuasion rating = 4.00*]	Many countries like France have already banned single-use plastics successfully with positive results! It's time we follow suit—let's ensure a sustainable future for us all! #NoMoreSingleUsePlastics [*Mean persuasion rating = 2.70*]
False	Over half of all plastic waste in the ocean comes from single-use products like bags and utensils. We must stop this by banning them. [*Mean persuasion rating = 3.90*]	Did you know that single-use plastic products are responsible for 90% of all ocean pollution? It's time to ban them once and for all! #BanSingleUsePlastic [*Mean persuasion rating = 2.80*]
Condition	High share online	Low share online
	Attention game	
Experiment 3: Human producers		
True	Every dog owner knows dogs make better pets. Can cats fetch the newspaper? No! Best they can do is pee on it. Will a cat protect from an intruder? An intruder who is not allergic to cats? I think not. Dogs rock! [*Mean share online rating = 2.70*]	There is a reason the saying "Dogs are a man's best friend" exists (sorry all you cat people!) lol [*Mean share online rating = 1.20*]
False	Studies have shown that dog owners live up to 5 years longer than cat owners, on average. [*Mean share online rating = 3.00*]	Having a cat in your house increases the chances that you'll contract toxoplasmosis by 10,000%! [*Mean share online rating = 1.10*]
Experiment 4: LLM producer		
True	Dreaming of a greener future? Well, did you know that satellites help monitor deforestation and climate change patterns from above? Increasing investments in space exploration means a better understanding and protection of our planet! 🌍🚀 #GoGreenWithSpace [*Mean share online rating = 2.80*]	Space is the place!!! And we're just a pale blue dot. [*Mean share online rating = 1.10*]
False	Incredible but true: The surface of Mercury is covered with sparkling diamonds as far as the eye can see 💎💥 Expanding investments into space exploration will enable humankind to finally claim ownership over this luxurious extraterrestrial diamond mine! #MercurysDiamondRush [*Mean share online rating = 2.00*]	We have to increase spending because soon enough we will all be able to live on the planet of our choosing. [*Mean share online rating = 1.20*]

Note. The example messages provided were in response to the claims: *Tobacco smoking should be banned, Single-use plastic products should be banned, Dogs make better pets than cats, and Governments should increase their investment in space exploration*. LLM = large language model. See the online article for the color version of this table.

True-condition messages were also rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline than the false-condition messages ($ps < .001$). By contrast, the false-condition messages elicited stronger negative emotions ($p < .001$). The true- and false-condition messages were rated similarly with respect to the interest-if-true and social engagement dimensions ($ps > .204$). These findings replicate the Experiment 1 persuasion game results. Unlike Experiment 1, the true-condition messages were rated as more truthful and elicited stronger positive emotions than the unconstrained-condition messages ($ps < .003$). The true-condition messages were also rated as more persuasive and led to stronger belief updating ($ps < .038$). The unconstrained-condition messages elicited stronger negative emotions than the true-condition messages ($p < .001$). The true- and unconstrained-condition messages were rated similarly with respect to the other dimensions: relevant, familiar, interest, interest-if-true, social engagement, online

Figure 2

Experiment 1 Density Plots for Each Dimension (With Means Indicated by the Vertical Lines): True Condition (Green), Unconstrained (Gray), and False (Orange)



Note. See the online article for the color version of this figure.

sharing, and offline sharing ($ps > .067$). The unconstrained-condition messages showed the same pattern of results as the true-condition messages when compared with the false-condition messages. While the true-condition messages increased belief in the claim (+2.52 points; $p < .001$), the false-condition messages decreased belief in the claim (−4.66 points; $p < .001$). There was no statistical evidence that the unconstrained-condition messages affected belief in the claim (+0.58 points; $p = .390$; see Figure 4 and Table 6).

Next, we examined the relationships between different dimensions through a correlational analysis (see Figure 5, Panel A). Again, the correlations ranged from negligible ($r = .02$ for negative emotion and relevance) to strong ($r = .74$ for online sharing and offline sharing), with most dimensions showing moderate positive correlations. We then identified which dimensions best predicted the key outcomes, persuasion and belief update, plus online and offline sharing, using hierarchical backward elimination stepwise regression (see Table 7). For persuasion, the retained dimensions explained 59% of the variance, mostly driven by message truth (41%), positive emotion (+12%), and message interest (+5%). For belief update, 71% of the variance was accounted for, mainly by prior belief in the claim (58%) and message truth (+11%). For both online sharing and offline sharing, the retained dimensions explained 39% of the variance. Most of the variance was explained by positive emotion (30%, 25%), social engagement (+3%, +5%), and message persuasion (online sharing; +3%) or message interest (offline sharing; +6%).

Experiment 4: The Attention Game (LLM Producers)

The Experiment 4 results replicated the findings from Experiment 3. LLM-produced true-condition messages were rated by humans as more truthful, relevant, familiar, and interesting and elicited stronger positive emotions than the false-condition messages ($ps < .001$). Again, the true-condition messages were rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline compared with the false-condition messages ($ps < .001$). The false-condition messages elicited stronger negative

Table 4

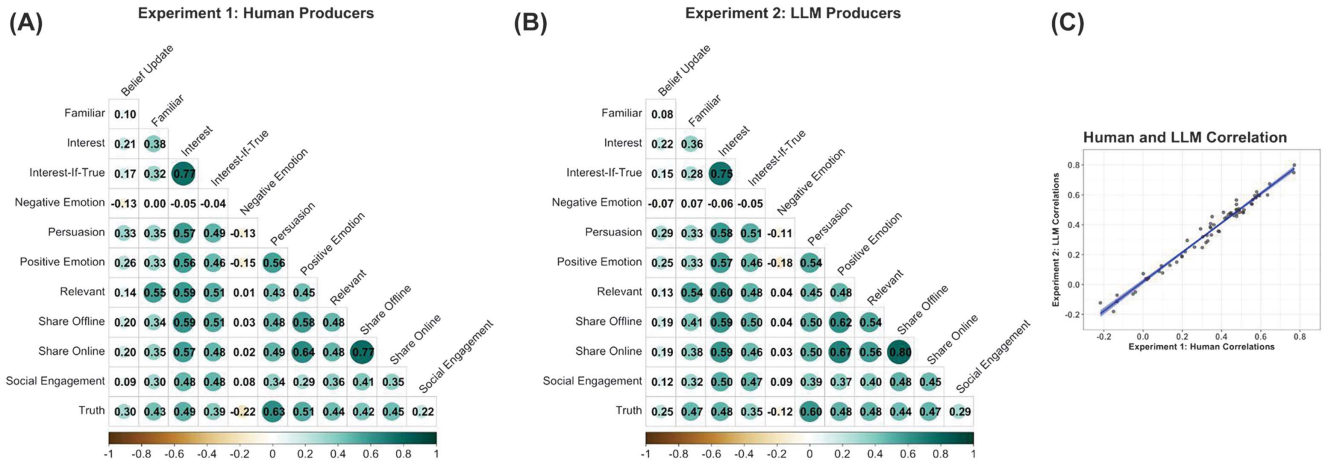
Experiment 1: Results of the Linear Mixed-Effects Model Across the Message Veracity Conditions for Each Dimension

	True versus false				True versus unconstrained				Unconstrained versus false			
Dimension	Coefficient	95% CI	*t*	*p*	*B*	95% CI	*t*	*p*	*B*	95% CI	*t*	*p*
Truth	−0.50	[−0.58, −0.43]	−12.91	**<.001**	−0.05	[−0.13, 0.03]	−1.26	.207	−0.46	[−0.53, −0.38]	−11.65	**<.001**
Relevant	−0.21	[−0.25, −0.16]	−8.30	**<.001**	−0.01	[−0.06, 0.03]	−0.59	.552	−0.19	[−0.24, −0.14]	−7.71	**<.001**
Familiar	−0.33	[−0.38, −0.28]	−12.95	**<.001**	−0.04	[−0.09, 0.01]	−1.65	.099	−0.29	[−0.34, −0.24]	−11.29	**<.001**
Interest	−0.13	[−0.19, −0.07]	−4.13	**<.001**	−0.03	[−0.09, 0.03]	−0.98	.326	−0.10	[−0.16, −0.04]	−3.15	**.002**
Interest-if-true	0.01	[−0.05, 0.06]	0.23	.815	−0.01	[−0.07, 0.04]	−0.43	.665	0.02	[−0.04, 0.07]	0.67	.504
Social engagement	0.00	[−0.06, 0.06]	0.03	.980	0.01	[−0.05, 0.07]	0.38	.703	−0.01	[−0.07, 0.05]	−0.36	.722
Positive emotion	−0.16	[−0.22, −0.11]	−6.10	**<.001**	−0.01	[−0.06, 0.04]	−0.40	.691	−0.15	[−0.21, −0.10]	−5.70	**<.001**
Negative emotion	0.12	[0.08, 0.16]	6.38	**<.001**	0.01	[−0.03, 0.04]	0.40	.693	0.11	[0.08, 0.15]	5.98	**<.001**
Persuasion	−0.33	[−0.42, −0.23]	−6.83	**<.001**	−0.04	[−0.14, 0.05]	−0.89	.374	−0.28	[−0.38, −0.19]	−5.94	**<.001**
Belief update	−4.53	[−5.51, −3.55]	−9.07	**<.001**	−0.28	[−1.26, 0.70]	−0.56	.573	−4.25	[−5.23, −3.27]	−8.50	**<.001**
Share online	−0.12	[−0.16, −0.08]	−5.88	**<.001**	−0.03	[−0.07, 0.01]	−1.38	.168	−0.09	[−0.13, −0.05]	−4.50	**<.001**
Share offline	−0.10	[−0.14, −0.06]	−4.58	**<.001**	−0.01	[−0.05, 0.03]	−0.53	.598	−0.09	[−0.13, −0.05]	−4.05	**<.001**

Note. Values presented in bold indicate statistically significant (at $p < .05$) and values not in bold indicate not statistically significant (at $p > .05$). CI = confidence interval.

Figure 3

Experiments 1 and 2: Correlation Structures of Human and LLM Producers Across Dimensions



Note. Panel A: Correlation matrix for the human producers across dimensions (Experiment 1). Panel B: Correlation matrix for the LLM producers (GPT-3.5) across dimensions (Experiment 2). Panel C: Correlation between the correlation coefficients in Panel A (human producers) and Panel B (LLM producers). LLM = large language model. See the online article for the color version of this figure.

emotions than the true-condition messages ($p < .001$). While the LLM-produced true-condition messages increased belief in the claim (+5.08 points; $p < .001$), the LLM-produced false-condition messages decreased belief in the claim (−7.11 points; $p < .001$). Compared with the human-produced messages sampled from Experiment 3, the LLM-produced true-condition messages were rated as more persuasive, led to stronger belief updating, and were more likely to be transmitted online and offline ($ps < .036$). By contrast, the LLM-produced false-condition messages did not differ to the human-produced messages, except that they led to weaker belief updating ($p = .049$).

The correlation matrix for LLM-produced messages across the different dimensions mirrored that of the human-produced messages (Figure 5, Panel B). The correlation between the coefficients for the human- and LLM-produced messages was $r = .94$[2] (Figure 5, Panel C). Reflecting this strong correlation, the stepwise regression analyses replicated the main Experiment 3 results. Message persuasion (60% of variance accounted for by the retained dimensions) was mostly driven by message truth (44%), message interest (+11%), and positive emotion (+2%). Belief update (77% of variance) was mostly driven by prior belief in the claim (63%) and message truth (+11%). Online sharing and offline sharing (45% and 51% of the variance, respectively) were mostly driven by positive emotion (30%, 33%) and (lower) negative emotion (online sharing; +6%) or message interest (offline sharing; +10%).

Discussion

The studies reported provide experimental evidence suggesting that, in the marketplace of ideas, truth wins. In all four experiments, the true-condition messages were consistently more persuasive, led to stronger belief updating, and were more likely to be reshared (online and offline) than the false-condition messages. In short, the true-condition messages had more impact than the false-condition messages. Moreover, messages generated by LLMs were more impactful than those written by humans, particularly when conveying true information. While the true-condition messages reliably increased participants' belief in the claims (Experiments 1–4), the false-condition messages either had no effect (Experiment 2) or, more commonly, reduced participants' belief in the claims (Experiments 1, 3, and 4). Furthermore, when the participants' goal was to create persuasive messages and they were unconstrained by message veracity (Experiment 1), they produced messages that were rated as similarly truthful to those in the true condition. This default tendency toward truthfulness was relaxed when the goal was to create attention-grabbing messages (Experiments 3 and 4). Here, when message veracity was unconstrained, participants produced messages that were rated as slightly less truthful than those in the true condition but still substantially more truthful than those in the false condition. This suggests that while people tend to prioritize the truth as a default, they are willing to sacrifice it to some extent for the sake of creating more engaging messages, as per the phrase "never let the truth get in the way of a good story." However, relaxing the truth did not increase engagement; social engagement and intent to reshare the message were unaffected. This is consistent with other research suggesting that exaggerated press releases about scientific findings do not lead to increased media coverage (Sumner et al., 2014, 2016).

Our results also distinguish between the factors that drive message influence and those that drive message spread. The primary driver of message influence—persuasion and belief update (after accounting for prior belief in the claim)—was the perceived truth of the message (Experiments 1–4). This finding—that truth acts as the gatekeeper of informational influence—aligns with research showing that people update their impressions of others only when the new information is credible (Cone et al., 2019) and that belief in conspiracy theories declines when people encounter compelling

[2] The correlation between the coefficients for the human raters from Experiments 3 and 4 on the human-generated messages was similarly high ($r = .98$).

Table 5

Experiment 1: Results of the Hierarchical Backward Elimination Stepwise Regression Analysis for Persuasion, Belief Update, Share Online, and Share Offline

Step	Dimension	Coefficient	95% CI	*t*	*p*	Marginal R^2	Dimension	Coefficient	95% CI	*t*	*p*	Marginal R^2
Persuasion							Belief update					
1	Truth	0.35	[0.34, 0.37]	68.15	**<.001**	0.36	Prior belief	0.67	[0.66, 0.68]	179.32	**<.001**	0.69
2	Positive emotion	0.17	[0.16, 0.18]	29.94	**<.001**	0.45	Truth	4.81	[4.60, 5.03]	43.47	**<.001**	0.75
3	Interest	0.18	[0.16, 0.19]	23.78	**<.001**	0.51	Persuasion	2.70	[2.48, 2.92]	23.73	**<.001**	0.76
4	Belief update	0.00	[0.00, 0.01]	17.33	**<.001**	0.51	Negative emotion	−2.08	[−2.28, −1.88]	−20.55	**<.001**	0.77
5	Social engagement	0.09	[0.08, 0.11]	16.48	**<.001**	0.52	Positive emotion	1.40	[1.20, 1.60]	13.51	**<.001**	0.77
6	Interest-if-true	0.08	[0.07, 0.09]	11.20	**<.001**	0.52	Interest	0.77	[0.54, 0.99]	6.70	**<.001**	0.77
7	Negative emotion	−0.01	[−0.03, −0.00]	−2.61	**.009**	0.52	Familiar	−0.55	[−0.74, −0.35]	−5.58	**<.001**	0.77
8							Social engagement	−0.53	[−0.73, −0.32]	−5.10	**<.001**	0.77
Online sharing							Offline sharing					
1	Positive emotion	0.28	[0.27, 0.29]	50.70	**<.001**	0.29	Positive emotion	0.26	[0.25, 0.27]	42.48	**<.001**	0.26
2	Social engagement	0.10	[0.09, 0.11]	18.41	**<.001**	0.33	Social engagement	0.13	[0.11, 0.14]	21.60	**<.001**	0.33
3	Interest	0.11	[0.10, 0.12]	16.32	**<.001**	0.38	Interest	0.16	[0.15, 0.18]	21.43	**<.001**	0.39
4	Relevant	0.06	[0.05, 0.07]	13.01	**<.001**	0.39	Relevant	0.08	[0.07, 0.09]	13.78	**<.001**	0.41
5	Persuasion	0.07	[0.06, 0.08]	12.20	**<.001**	0.39	Negative emotion	0.07	[0.06, 0.08]	12.23	**<.001**	0.42
6	Truth	0.06	[0.05, 0.07]	11.32	**<.001**	0.39	Persuasiveness	0.07	[0.06, 0.08]	11.16	**<.001**	0.42
7	Interest-if-true	0.05	[0.04, 0.06]	7.97	**<.001**	0.39	Interest-if-true	0.07	[0.05, 0.08]	9.48	**<.001**	0.42
8	Negative emotion	0.04	[0.03, 0.05]	7.53	**<.001**	0.40	Truth	0.04	[0.03, 0.05]	6.69	**<.001**	0.42
9	Familiar	0.03	[0.02, 0.04]	5.18	**<.001**	0.40	Familiar	0.03	[0.02, 0.05]	6.00	**<.001**	0.42

Note. For each outcome, we used a linear mixed model with all the predictors included. We sequentially removed the predictor with the lowest *t* value and used maximum likelihood estimation for model comparison. Predictors were removed if their exclusion did not reduce model fit ($p > .05$), continuing this process until removal reduced model fit ($p < .05$). Values presented in bold indicate statistically significant (at $p < .05$) and values not in bold indicate not statistically significant (at $p > .05$). CI = confidence interval.

Figure 4

Experiment 3 Density Plots for Each Dimension (With Means Indicated by the Vertical Lines): True Condition (Green), Unconstrained (Gray), and False (Orange)

Note. See the online article for the color version of this figure.

fact-based counterarguments (Costello et al., 2024). More broadly, it supports the view of humans as "information foragers," who analytically search their environment for valuable information (Pirolli & Card, 1999). By contrast, although true messages were shared more than false ones, truth was not the primary driver of message spread. Instead, the intention to share messages—both online and offline—was largely driven by the positive emotions they evoked and the anticipated social engagement they generated (Experiments 1–4). This finding is a testament to the importance of emotions in human decision making (Lerner et al., 2015; Schwarz & Clore, 1983) and aligns with research showing that messages eliciting high-arousal positive emotions tend to be more viral (Berger & Milkman, 2012). The importance of positive emotions and social engagement indicates that people may prioritize social connection when sharing information (Clark & Kashima, 2007; Lyons & Kashima, 2003), consistent with the idea that their behavior is guided by the core social motive to belong (Fiske, 2018).

The metaphor of misinformation as a virus—as reflected by the term "infodemic"—has been used to describe the rapid spread and harmful impact of false information (Rothkopf, 2003; van der Linden, 2023; Zarocostas, 2020) and has informed strategies designed to combat it (Blair et al., 2024; Kozyreva et al., 2024; Tay et al., 2022). However, the metaphor has been criticized for oversimplifying a complex issue, in large part because it conflates information spread with influence (Altay et al., 2023; Simon & Camargo, 2023). Unlike a virus, where infection is involuntary, people can choose to accept or reject the information they encounter. Rather than viewing people as passive information consumers, it may be more accurate to see them as skeptical and discerning information evaluators (Mercier, 2020), as our findings demonstrate—participants were persuaded by messages in the true condition and dissuaded by those in the false condition. This position is supported by a meta-analysis showing that people are good at discerning between true and false news (Pfänder & Altay, 2025) and the finding that false information on the social media

Table 6

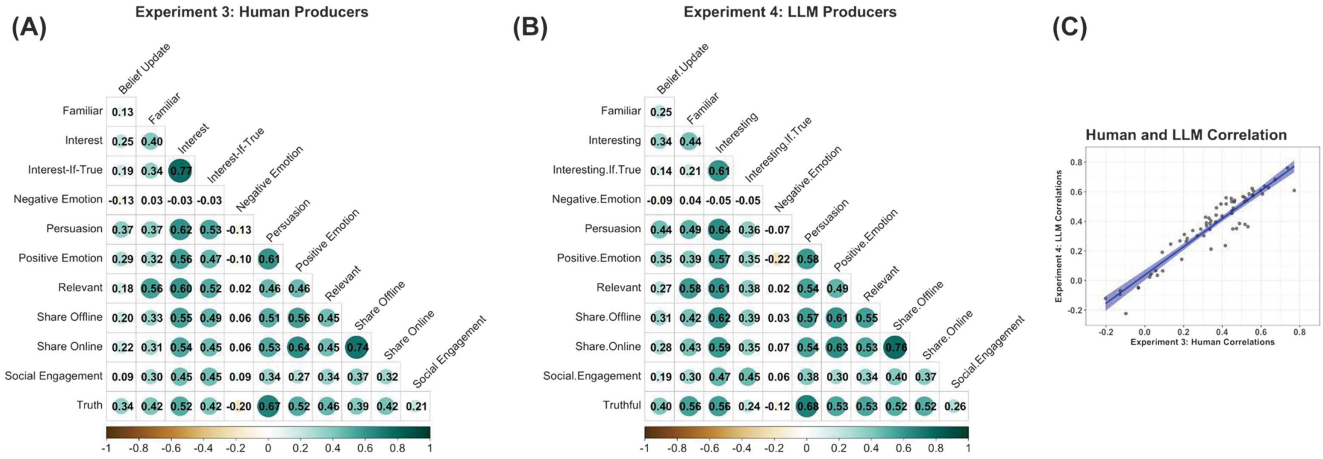
Experiment 3: Results of the Linear Mixed-Effects Model Across the Message Veracity Conditions for Each Dimension

	True versus false				True versus unconstrained				Unconstrained versus false			
Dimension	Coefficient	95% CI	*t*	*p*	*B*	95% CI	*t*	*p*	*B*	95% CI	*t*	*p*
Truth	−0.61	[−0.69, −0.53]	−14.86	**<.001**	−0.18	[−0.26, −0.10]	−4.43	**<.001**	−0.43	[−0.51, −0.35]]	−10.43	**<.001**
Relevant	−0.24	[−0.29, −0.19]	−9.01	**<.001**	−0.05	[−0.10, 0.00]	−1.83	.067	−0.19	[−0.24, −0.14]	−7.18	**<.001**
Familiar	−0.33	[−0.38, −0.28]	−12.09	**<.001**	−0.04	[−0.10, 0.01]	−1.64	.101	−0.29	[−0.34, −0.23]	−10.45	**<.001**
Interest	−0.13	[−0.18, −0.07]	−4.28	**<.001**	−0.04	[−0.09, 0.02]	−1.19	.234	−0.09	[−0.15, −0.03]	−3.09	**.002**
Interest-if-true	0.02	[−0.04, 0.08]	0.66	.512	0.01	[−0.05, 0.07]	0.33	.741	0.01	[−0.05, 0.07]	0.32	.745
Social engagement	0.04	[−0.02, 0.09]	1.27	.205	0.03	[−0.02, 0.09]	1.23	.220	0.00	[−0.05, 0.06]	0.04	.968
Positive emotion	−0.20	[−0.25, −0.15]	−7.88	**<.001**	−0.06	[−0.11, −0.01]	−2.44	**.015**	−0.14	[−0.19, −0.09]	−5.45	**<.001**
Negative emotion	0.16	[0.11, 0.20]	7.08	**<.001**	0.09	[0.04, 0.13]	3.94	**<.001**	0.07	[0.03, 0.11]	3.14	**.002**
Persuasion	−0.33	[−0.42, −0.25]	−7.90	**<.001**	−0.09	[−0.17, −0.01]	−2.09	**.037**	−0.25	[−0.33, −0.16]	−5.81	**<.001**
Belief update	−7.19	[−8.42, −5.96]	−11.45	**<.001**	−1.94	[−3.17, −0.71]	−3.09	**.002**	−5.25	[−6.48, −4.02]	−8.36	**<.001**
Share online	−0.09	[−0.12, −0.05]	−5.03	**<.001**	−0.02	[−0.06, 0.01]	−1.36	.175	−0.06	[−0.10, −0.03]	−3.67	**<.001**
Share offline	−0.07	[−0.11, −0.03]	−3.50	**<.001**	0.00	[−0.04, 0.04]	0.10	.924	−0.07	[−0.11, −0.03]	−3.59	**<.001**

Note. Values presented in bold indicate statistically significant (at $p < .05$) and values not in bold indicate not statistically significant (at $p > .05$). CI = confidence interval.

Figure 5

Experiments 3 and 4: Correlation Structures of Human and LLM Producers Across Dimensions



Note. Panel A: Correlation matrix for the human producers across dimensions (Experiment 3). Panel B: Correlation matrix for the LLM (GPT-3.5) producers across dimensions (Experiment 4). Panel C: Correlation between the correlation coefficients in Panel A (human producers) and Panel B (LLM producers). LLM = large language model. See the online article for the color version of this figure.

platform Facebook had no effect on COVID-19 vaccination intent (Allen et al., 2024). By contrast, the same Facebook study found that true-but-misleading content (e.g., "A healthy doctor died 2 weeks after getting a COVID vaccine") from mainstream news organizations reduced vaccination intent by 2.28 percentage points. This emphasizes the persuasive potential of gray-area content—as distinct from outright falsehoods—and highlights the moderating role of source credibility on message impact (Briñol & Petty, 2009; see also Kumkale et al., 2010; Prike et al., 2024). Other contextual factors that moderate informational influence include the communication channel (Breves, 2023; Chaiken & Eagly, 1976), information frequency (Hasher et al., 1977; Hassan & Barber, 2021; Pennycook et al., 2018), perceived consensus (Asch, 1951; Butler et al., 2023; Lewandowsky et al., 2019), and the characteristics of the audience, especially their political identity (Van Bavel & Pereira, 2018). In fact, in our study, participants' political identity moderated their belief in the claims, even those one would expect to be nonpartisan, such as the claim that "dogs make better pets than cats" (see Supplemental Material 2 for details).

Misinformation is a significant societal issue, as demonstrated by the hyper-partisan false claim that the 2020 U.S. presidential election was rigged, which in turn fueled the riots at the U.S. Capitol (Heine, 2021). Our work suggests that, when stripped of contextual factors like platform affordances, truthful messages persuade and untruthful messages dissuade—and these outcomes are not the result of confounding factors but are driven primarily by perceived message truth. For message transmission, the impact of truth is more nuanced. Although messages from the true condition were more likely to be shared than those from the false condition, this effect was driven by factors associated with message truth—positive emotion and social engagement—rather than truth itself. Furthermore, when participants could design persuasive messages without being constrained to use only true information, the messages they produced were rated as equally truthful as the true condition messages. However, this preference for truth diminished slightly when the goal was to create attention-grabbing messages. Taken together, our findings suggest that people are predisposed to the truth—both as information producers and consumers. This is consistent with the finding that the majority of online misinformation is spread by a small group of supersharers (Baribi-Bartov et al., 2024).

In the context of the experiments reported, we conclude that—within the boundaries of our design—truth holds a competitive edge in the marketplace of ideas. To be clear, the experiments prioritized internal validity over external or ecological validity. It remains unclear to what extent these findings will generalize to other populations or real-world settings (see Table 8 limitations). For example, our participant samples were drawn from Western, Educated, Industrialized, Rich, and Democratic societies (Henrich et al., 2010) and from LLMs predominantly trained on English-language data from Western, Educated, Industrialized, Rich, and Democratic societies (Atari et al., 2023). It is therefore important to test whether our findings replicate in different populations. Given the fundamental importance of truth to effective decision making, we anticipate that our findings will extend across diverse populations. Our results also provide a baseline against which to compare persuasion and sharing in situations that are more likely to be shaped by platform affordances, social context, and other real-world factors. Although true information was more persuasive, this advantage was shaped by multiple psychological factors, among them interest and positivity, that are not inherent to factual accuracy. This aligns with research showing that repeated exposure to falsehoods can render them compelling, as familiarity and fluency increase their perceived truth (Henderson et al., 2021; Pennycook et al., 2018). Such mechanisms can enhance the credibility of falsehoods, independent of their objective veracity. In addition, in naturalistic environments, falsehoods may be propagated by people who sincerely believe them, and these falsehoods may be reinforced by other interconnected falsehoods. Such mechanisms and dynamics may help explain the persistence of conspiracy theories and urban myths as stable features of the information environment (Heath et al., 2001; Pelletier & Drozda-Senkowska, 2020; Van Bavel et al., 2021). We encourage researchers to investigate the factors that moderate the conditions under which truth wins and when compelling but false information gains the upper hand (see Table 8 future research).

Table 7

Experiment 3: Results of the Hierarchical Backward Stepwise Regression Analysis for Persuasion, Belief Update, Share Online, and Share Offline

Step	Dimension	Coefficient	95% CI	*t*	*p*	Marginal R^2	Dimension	Coefficient	95% CI	*t*	*p*	Marginal R^2
Persuasion							Belief update					
1	Truth	0.36	[0.35, 0.37]	71.99	**<.001**	0.41	Prior belief	0.60	[0.59, 0.61]	144.56	**<.001**	0.58
2	Positive emotion	0.22	[0.21, 0.23]	39.02	**<.001**	0.53	Truth	5.68	[5.43, 5.92]	45.51	**<.001**	0.69
3	Interest	0.17	[0.16, 0.18]	24.11	**<.001**	0.58	Persuasion	3.35	[3.08, 3.62]	24.33	**<.001**	0.70
4	Belief update	0.00	[0.00, 0.00]	17.40	**<.001**	0.58	Negative emotion	−2.28	[−2.50, −2.06]	−20.11	**<.001**	0.71
5	Social engagement	0.08	[0.07, 0.09]	15.86	**<.001**	0.59	Positive emotion	1.48	[1.23, 1.72]	11.79	**<.001**	0.71
6	Interest-if-true	0.08	[0.07, 0.10]	12.89	**<.001**	0.59	Social engagement	−0.80	[−1.03, −0.58]	−7.01	**<.001**	0.71
7	Negative emotion	−0.03	[−0.04, −0.02]	−6.55	**<.001**	0.59	Interest	0.90	[0.64, 1.16]	6.89	**<.001**	0.71
8	Familiar	0.02	[−0.04, −0.02]	4.55	**<.001**	0.59	Familiar	−0.41	[−0.63, −0.20]	−3.78	**<.001**	0.71
Online sharing							Offline sharing					
1	Positive emotion	0.29	[0.27, 0.30]	52.85	**<.001**	0.30	Positive emotion	0.25	[0.24, 0.26]	41.96	**<.001**	0.25
2	Social engagement	0.08	[0.07, 0.09]	16.80	**<.001**	0.33	Social engagement	0.10	[0.09, 0.11]	18.83	**<.001**	0.30
3	Persuasion	0.09	[0.08, 0.10]	16.27	**<.001**	0.36	Interest	0.13	[0.12, 0.15]	18.34	**<.001**	0.36
4	Interest	0.09	[0.07, 0.10]	13.10	**<.001**	0.37	Persuasion	0.09	[0.08, 0.11]	16.91	**<.001**	0.36
5	Relevant	0.05	[0.04, 0.06]	11.08	**<.001**	0.38	Relevant	0.06	[0.05, 0.07]	11.90	**<.001**	0.37
6	Negative emotion	0.05	[0.04, 0.06]	9.53	**<.001**	0.39	Negative emotion	0.06	[0.05, 0.07]	11.86	**<.001**	0.39
7	Truth	0.04	[0.03, 0.04]	7.28	**<.001**	0.39	Interest−if−true	0.06	[0.05, 0.07]	8.96	**<.001**	0.39
8	Interest-if-true	0.04	[0.02, 0.05]	6.02	**<.001**	0.39	Familiar	0.03	[0.02, 0.04]	5.28	**<.001**	0.39
9	Belief update	−0.00	[−0.00, −0.00]	−3.44	**.001**	0.39						
10	Familiar	0.02	[0.01, 0.03]	3.40	**.001**	0.39						

Note. For each outcome, we used a linear mixed model with all the predictors included. We sequentially removed the predictor with the lowest *t* value and used maximum likelihood estimation for model comparison. Predictors were removed if their exclusion did not reduce model fit ($p > .05$), continuing this process until removal reduced model fit ($p < .05$). Values presented in bold indicate statistically significant (at $p < .05$) and values not in bold indicate not statistically significant (at $p > .05$). CI = confidence interval.

Table 8
Table of Limitations

Limiting factor	Limitation for interpretation	Recommendation for future research
Experimental method	The study design prioritized internal validity over external and ecological validity, making it uncertain whether the findings generalize to other modes of data collection.	Conduct similar studies in more naturalistic environments, such as social media platforms, to improve external and ecological validity through field experiments.
Participant sampling	Participants were U.S.-based adults recruited via Amazon Mechanical Turk, allowing for a larger and more diverse sample than typical undergraduate samples. However, our findings may be limited to adult participants from Western, Educated, Industrialized, Rich, and Democratic societies.	Extend research to in-person laboratory settings to enhance experimental control. Expand sampling to include younger and older adults, as well as participants from non-Western, Educated, Industrialized, Rich, and Democratic societies, to assess the generalizability of our findings.
Stimuli	Of the 32 claims tested, 15 were selected for the experiments based on criteria such as moderate prior agreement and minimal political polarization.	Use a broader and more diverse set of claims in future studies to test the robustness of our findings across different types of messages.
Predictors	Eight dimensions (e.g., interesting) were used to predict four key outcomes (e.g., persuasion).	Investigate additional predictors, such as message complexity, to gain a more complete understanding of message persuasion and sharing.
Moderators	The study focused on the message characteristics that drive persuasion and sharing, though other contextual factors are known to influence these outcomes (e.g., perceived consensus).	Examine the extent to which our "truth wins" findings are moderated by factors beyond the message content, such as the political ideology of the source and audience.

References

Abrams vs. United States. 250 U.S. 616. (1919).

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(44), Article e2313790120. https://doi.org/10.1073/pnas.2313790120

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), Article eadk3451. https://doi.org/10.1126/science.adk3451

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, *9*(1), Article 412. https://doi.org/10.1177/20563051221150412

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Organizational influence processes* (Vol. 58, pp. 295–303). Carnegie Press.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* https://doi.org/10.31234/osf.io/5b26t

Bai, M. H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2023). *Artificial intelligence can persuade humans on political issues*. https://doi.org/10.31219/osf.io/stakv

Baribi-Bartov, S., Swire-Thompson, B., & Grinberg, N. (2024). Supersharers of fake news on Twitter. *Science*, *384*(6699), 979–982. https://doi.org/10.1126/science.adl4435

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *Lme4: Linear mixed-effects models using Eigen and S4*. R Package Version 1.0-6.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, *55*, Article 101732. https://doi.org/10.1016/j.copsyc.2023.101732

Bond, R. M., & Garrett, R. K. (2023). Engagement with fact-checked posts on Reddit. *PNAS Nexus*, *2*(3), Article pgad018. https://doi.org/10.1093/pnasnexus/pgad018

Breves, P. (2023). Persuasive communication and spatial presence: A systematic literature review and conceptual model. *Annals of the International Communication Association*, *47*(2), 222–241. https://doi.org/10.1080/23808985.2023.2169952

Briñol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96. https://doi.org/10.1080/10463280802643640

Briñol, P., & Petty, R. E. (2012). A history of attitudes and persuasion research. In A. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 283–320). Psychology Press.

Brownson, R. C., Gurney, J. G., & Land, G. H. (1999). Evidence-based decision making in public health. *Journal of Public Health Management and Practice*, *5*(5), 86–97. https://doi.org/10.1097/00124784-199909000-00012

Butler, L. H., Fay, N., & Ecker, U. K. H. (2023). Social endorsement influences the continued belief in corrected misinformation. *Journal of Applied Research in Memory and Cognition*, *12*(3), 364–375. https://doi.org/10.1037/mac0000080

Chaiken, S., & Eagly, A. H. (1976). Communication modality as a determinant of message persuasiveness and message comprehensibility. *Journal of Personality and Social Psychology*, *34*(4), 605–614. https://doi.org/10.1037/0022-3514.34.4.605

Clark, A. E., & Kashima, Y. (2007). Stereotypes help people connect with others in the community: A situated functional analysis of the stereotype consistency bias in communication. *Journal of Personality and Social Psychology*, *93*(6), 1028–1039. https://doi.org/10.1037/0022-3514.93.6.1028

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, *116*(20) 9802–9807. https://doi.org/10.1073/pnas.1903222116

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), Article eadq1814. https://doi.org/10.1126/science.adq1814

Dawkins, R. (1989). *The selfish gene*. Oxford University Press.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Druckman, J. N. (2022). A framework for the study of persuasion. *Annual Review of Political Science*, *25*, 65–88. https://doi.org/10.1146/annurev-polisci-051120-110428

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to

correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Fay, N., Ransom, K., Walker, B., Howe, P., Perfors, A., & Kashima, Y. (2024). *Truth over falsehood: Experimental evidence on what persuades and spreads*. https://doi.org/10.17605/OSF.IO/T6SQ4

Fiske, S. T. (2018). *Social beings: Core motives in social psychology*. Wiley. https://books.google.com/books?hl=en&lr=&id=zE6MDwAAQBAJ&oi=fnd&pg=PR15&dq=susan+fiske+Social+Beings:+A+Core+Motives+Approach+to+Social+Psychology&ots=R_4SvG2m5n&sig=tQdfzh97zqecAtarZMQEmFX6KWo

Green, J., Hobbs, W., McCabe, S., & Lazer, D. (2022). Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(34), Article e2115900119. https://doi.org/10.1073/pnas.2115900119

Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and attitude change. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes* (pp. 147–170). Academic Press. https://doi.org/10.1016/B978-1-4832-3071-9.50012-X

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, *6*(1), Article 38. https://doi.org/10.1186/s41235-021-00301-5

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology*, *81*(6), 1028–1041. https://doi.org/10.1037/0022-3514.81.6.1028

Heine, J. (2021). The attack on the U.S. capitol: An American Kristallnacht. *Protest*, *1*(1), 126–141. https://doi.org/10.1163/2667372X-01010004

Henderson, E. L., Simons, D. J., & Barr, D. J. (2021). The trajectory of truth: A longitudinal study of the illusory truth effect. *Journal of Cognition*, *4*(1), Article 29. https://doi.org/10.5334/joc.161

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology*, *14*, Article 1281255. https://doi.org/10.3389/fpsyg.2023.1281255

Kashima, Y., Coman, A., Pauketat, J. V. T., & Yzerbyt, V. (2020). Emotion in cultural dynamics. *Emotion Review*, *12*(2), 48–64. https://doi.org/10.1177/1754073919875215

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

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

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

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019). Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, *47*(8), 1445–1456. https://doi.org/10.3758/s13421-019-00948-y

Lewandowsky, S., Ecker, U. K. H., Cook, J., van der Linden, S., Roozenbeek, J., & Oreskes, N. (2023). Misinformation and the epistemic integrity of democracy. *Current Opinion in Psychology*, *54*, Article 101711. https://doi.org/10.1016/j.copsyc.2023.101711

Lyons, A., & Kashima, Y. (2003). How are stereotypes maintained through communication? The influence of stereotype sharedness. *Journal of Personality and Social Psychology*, *85*(6), 989–1005. https://doi.org/10.1037/0022-3514.85.6.989

McGuire, W. J. (1985). Attitudes and attitude change. In L. Gardner & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, pp. 233–346). Penguin Random House. https://cir.nii.ac.jp/crid/1571135650731642368

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press. https://doi.org/10.1515/9780691198842

Milton, J. (1868). *Areopagitica, 1644*. Alex Murray & Son.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management*, *57*(4), Article 102250. https://doi.org/10.1016/j.ipm.2020.102250

Pelletier, P., & Drozda-Senkowska, E. (2020). Towards a socially situated rumouring: Historical and critical perspectives of rumour transmission. *Social and Personality Psychology Compass*, *14*(6), Article e12532. https://doi.org/10.1111/spc3.12532

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pertwee, E., Simas, C., & Larson, H. J. (2022). An epidemic of uncertainty: Rumors, conspiracy theories and vaccine hesitancy. *Nature Medicine*, *28*(3), 456–459. https://doi.org/10.1038/s41591-022-01728-z

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag. https://doi.org/10.1007/978-1-4612-4964-1

Pfänder, J., & Altay, S. (2025). Spotting false news and doubting true news: A systematic review and meta-analysis of news judgements. *Nature Human Behaviour*, *9*(4), 688–699. https://doi.org/10.1038/s41562-024-02086-1

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

Prike, T., Butler, L. H., & Ecker, U. K. H. (2024). Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online. *Scientific Reports*, *14*(1), Article 6900. https://doi.org/10.1038/s41598-024-57560-7

R Core Team. (2013). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rothkopf, D. J. (2003). Opinion|when the buzz bites back. *Washington Post*. https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67. https://doi.org/10.1080/01621459.1951.10500768

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523. https://doi.org/10.1037/0022-3514.45.3.513

Simon, F. M., & Camargo, C. Q. (2023). Autopsy of a metaphor: The origins, use and blind spots of the "infodemic". *New Media & Society*, *25*(8), 2219–2240. https://doi.org/10.1177/14614448211031908

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), Article eadh1850. https://doi.org/10.1126/sciadv.adh1850

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings*

of the National Academy of Sciences of the United States of America, *115*(49), 12435–12440. https://doi.org/10.1073/pnas.1803470115

Steup, M., & Neta, R. (2005). *Epistemology*. https://plato.stanford.edu/entries/epistemology/?utm_medium=podcast&utm_source=bcast&utm_campaign=gold-exchange-with-keith-weiner

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R., Venetis, C. A., Whelan, L., Hughes, B., & Chambers, C. D. (2016). Exaggerations and caveats in press releases and health-related science news. *PLOS ONE*, *11*(12), Article e0168217. https://doi.org/10.1371/journal.pone.0168217

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., Boy, F., & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *The BMJ*, *349*, Article g7015. https://doi.org/10.1136/bmj.g7015

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2022). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, *113*(3), 591–607. https://doi.org/10.1111/bjop.12551

Treen, K. M., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, *11*(5), Article e665. https://doi.org/10.1002/wcc.665

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

van der Linden, S. (2023). *Foolproof: Why misinformation infects our minds and how to build immunity*. W.W. Norton.

Verma, P. (2023, December 18). The rise of AI fake news is creating a "misinformation superspreader". *Washington Post*. https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, *47*(2), 155–177. https://doi.org/10.1177/0093650219854600

Zarocostas, J. (2020). How to fight an infodemic. *Lancet*, *395*(10225), Article 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Received December 10, 2024
Revision received July 22, 2025
Accepted August 8, 2025 ■