Truth Wins: True Information is More Persuasive and Shareable than Falsehoods

Nicolas Fay[1]*, Keith J. Ransom[2], Bradley Walker[1], Piers D.L. Howe[3], Andrew Perfors[3], Yoshihisa Kashima[3]
[1]School of Psychological Science, University of Western Australia; Perth, Australia
[2]School of Computer and Mathematical Sciences, University of Adelaide; Adelaide, Australia
[3]School of Psychological Sciences, University of Melbourne

Running Head: Truth Wins

Keywords: Misinformation, Disinformation, True, False, Persuasion, Attention, Transmission, Large Language Model, LLM

*Corresponding author:
Nicolas Fay, School of Psychological Science, University of Western Australia
35 Stirling Highway, Crawley, WA 6009 Australia
Email: nicolas.fay@gmail.com; Tel: +61 (0)8 6488 2688; Fax: +61 (0)8 6488 1006

Word count (excluding title page, abstract, methods, figures captions & references): XXXX words

Abstract

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. Four experiments, involving human participants and Large Language Models (combined N=4607), are reported that compare the persuasive impact and transmission potential of true and false information. The results consistently showed that messages created with the intent of being true were more persuasive and more likely to be shared than those created to be false. While perceived message truth was the primary driver of persuasion, positive emotion and social engagement were the primary drivers of message transmission. Our findings indicate that in the marketplace of ideas, truth wins.

 2

Introduction

"*Let her [Truth] and Falsehood grapple; who ever knew Truth put to the worse, in a free and open encounter?*" John Milton, Areopagitica, 1644.

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 (*1*). The impact and spread of false information through online media suggests otherwise. False information has undermined public health (*2*), delayed climate action (*3*), eroded trust in institutions (*4*) and manufactured societal problems that threaten the foundation of liberal democracies (*5*). False information also spreads farther, faster, deeper, and more broadly (via re-sharing) than true information on the social media platform *Twitter* (now *X*; *6*). This is attributed to false information's greater novelty, and its ability to elicit powerful negative emotions. Unconstrained by reality, false information appears to thrive in the marketplace of ideas (*7*, *8*). Concerns around the impact and spread of false information are compounded by evidence indicating that once false information is accepted it is difficult to correct (*9*, *10*), and by the potential of large language models (hereafter LLMs) to generate and disseminate false information at scale (*11*).

Truth matters; it is foundational to epistemology—the branch of philosophy concerned with the nature, origin and limits of human knowledge (*12*)—and is critical to forming accurate beliefs and making effective decisions (*13*, *14*). 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 social media platform *Reddit* that were rated as true were associated with stronger user engagement (volume of user comments and conversation length) than fact-checked posts that were rated as false (*15*). 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 (*16*, *17*).

Mass social influence requires two key elements: the message must be received, and it must be persuasive. Reception relies on message transmission, while persuasion depends on the receivers' evaluation of the message content. Simply receiving a message is insufficient for persuasion; the message must also evoke a positive evaluation (*18–20*). Conversely, without transmission, persuasion is impossible. So, impact relies on the combination of message influence and message spread. The studies reported here experimentally test the persuasive influence and transmission potential of true and false information. 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 when unconstrained by message veracity. A second group of participants rated the messages across a range of dimensions, including persuasiveness and willingness to share. Experiment 2 used LLM-generated persuasive messages (GPT-3.5) to test the robustness of the Experiment 1 findings (*21–25*). Experiments 3 and 4 report the findings of the Attention Game, in which participants were instructed to write attention-grabbing messages. Experiment 3 used the same experimental design as Experiment 1, while Experiment 4 used LLM-generated attention-grabbing messages to test the robustness of the Experiment 3 findings. In each experiment we found that perceived message truth was the primary factor driving persuasion, while social connection (e.g., ratings of positive emotion and social engagement) was the key driver of message transmission.

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 to 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>.815). For each dimension, the Unconstrained-Condition messages were rated similarly to the True-Condition messages (ps>.099), and showed the same pattern of results as the True-Condition messages when compared to 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 1).

Next, we examined relationships between the different dimensions through a correlational analysis (see Figure 2, 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 backwards elimination stepwise regression (see Table 1). 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 to 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 to 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).

The correlation matrix for LLM-produced messages across the different dimensions mirrored that of the human-produced messages (Figure 2, Panel B). The correlation between the coefficients for the human- and LLM-produced messages was r=.99[1] (Figure 2, 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%).

1 The correlation between the coefficients for the human raters from Experiment 1 and Experiment 2 on the human-generated messages were equally high, r=0.99.

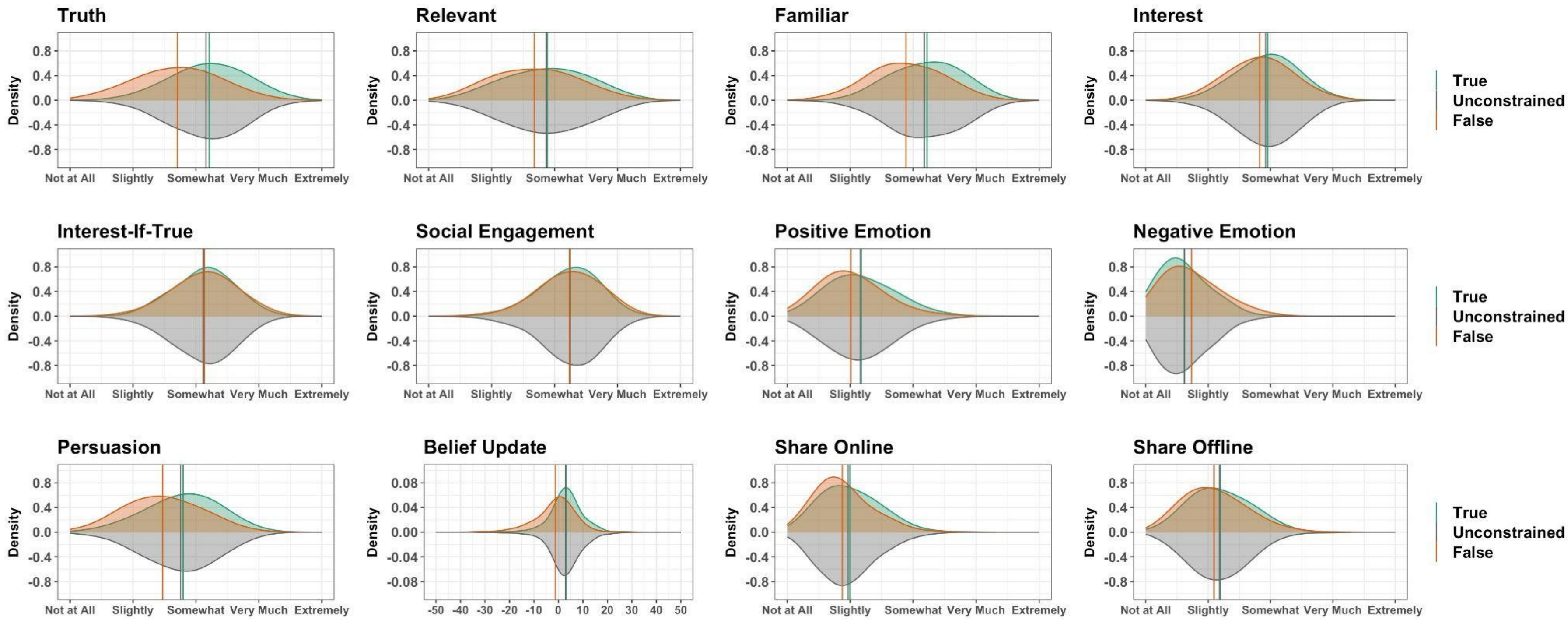

Fig. 1. Experiment 1 density plots for each dimension, with means indicated by the vertical lines: True Condition (green), Unconstrained (gray), and False (orange).

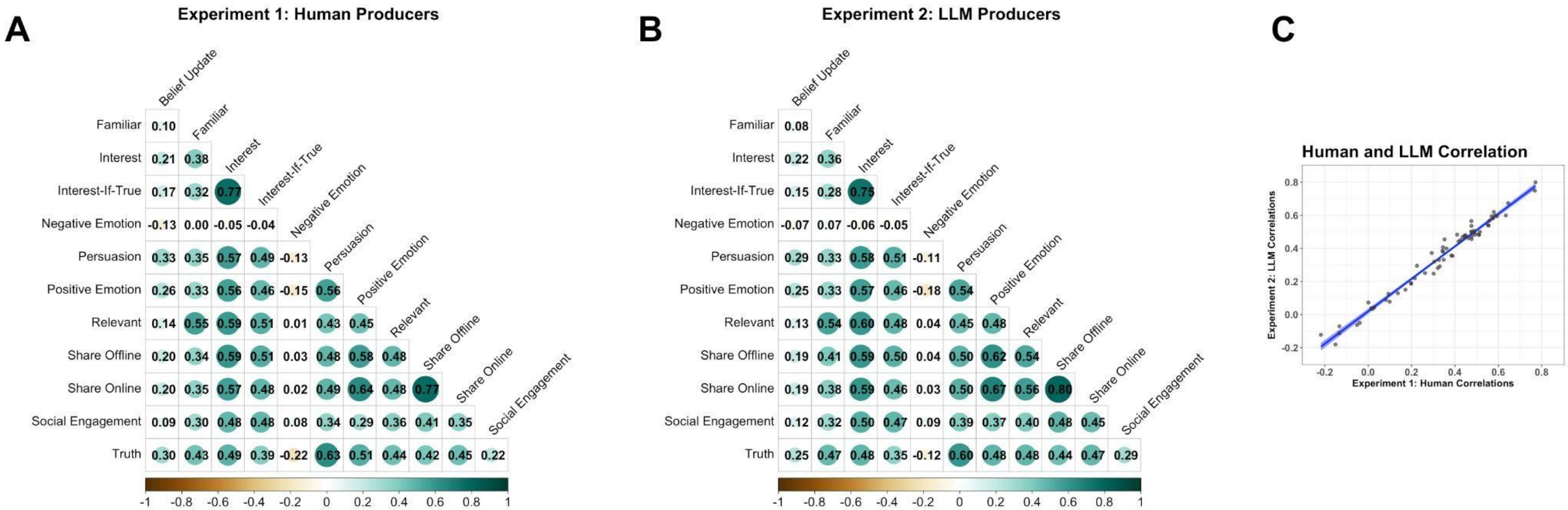

Fig. 2. Panel A: Correlation matrix for the human producers across dimensions (Experiment 1). Panel B: Correlation matrix for the LLM producers (GPT-3.5) across dimensions (Experiment 2). Panel C: Correlation between the correlation coefficients in Panel A (Human Producers) and Panel B (LLM Producers).

Table 1. Experiment 1: Results of the hierarchical backwards elimination stepwise regression analysis for Persuasion, Belief Update, Share Online and Share Offline.

Persuasion							**Belief Update**					
Step	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**
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**					
Step	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**
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).

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 (*p*s<.001). 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 (*p*s<.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 (*p*s>.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 (*p*s<.003). The True-Condition messages were also rated as more persuasive and led to stronger belief updating (*p*s<.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 sharing and offline sharing (*p*s>.067). The Unconstrained-Condition messages showed the same pattern of results as the True-Condition messages when compared to 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 3).

Next, we examined relationships between different dimensions through a correlational analysis (see Figure 4, 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 backwards elimination stepwise regression (see Table 2). 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 (*p*s<.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 to the False-Condition messages (*p*s<.001). The False-Condition messages elicited stronger negative 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).

The correlation matrix for LLM-produced messages across the different dimensions mirrored that of the human-produced messages (Figure 4, Panel B). The correlation between the coefficients for the human- and LLM-produced messages was r=.94[2] (Figure 4, 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%).

[2] The correlation between the coefficients for the human raters from Experiment 3 and Experiment 4 on the human-generated messages were similarly high, r=.98.

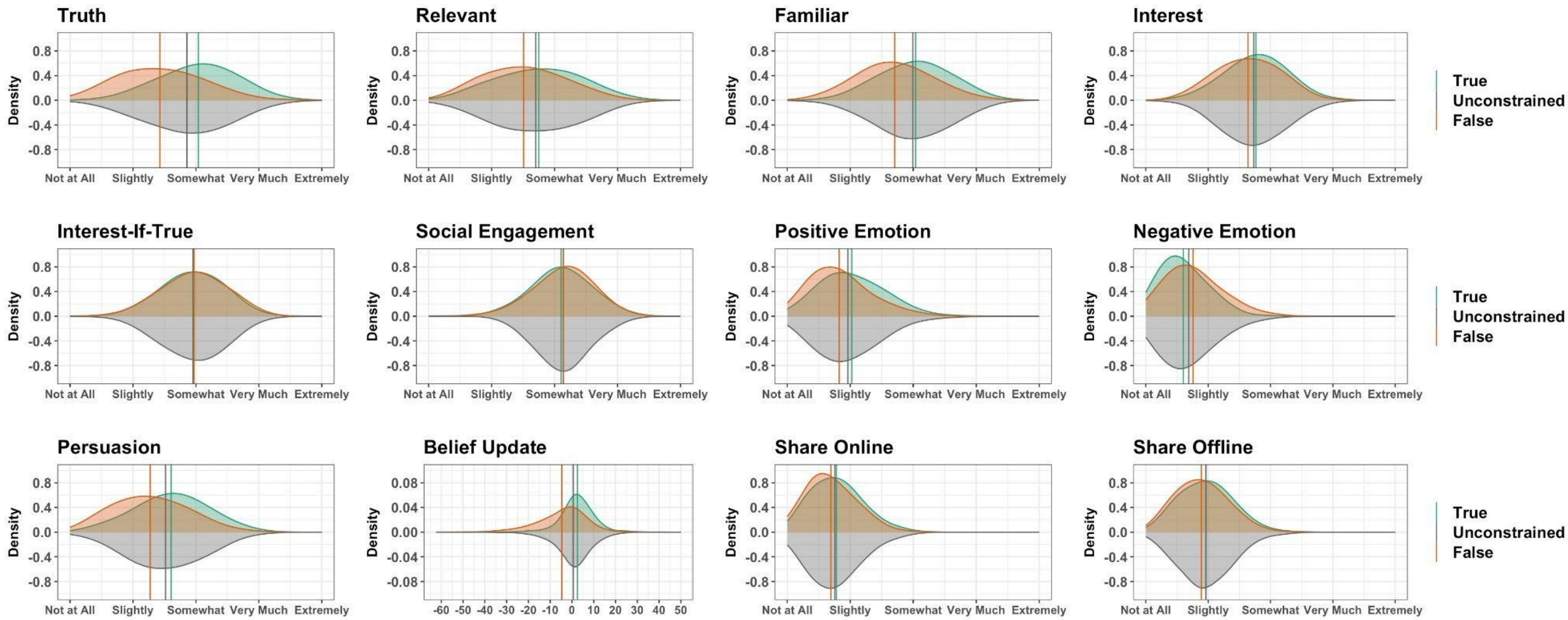

Fig. 3. Experiment 3 Density plots for each dimension, with means indicated by the vertical lines: True Condition (green), Unconstrained (gray), and False (orange).

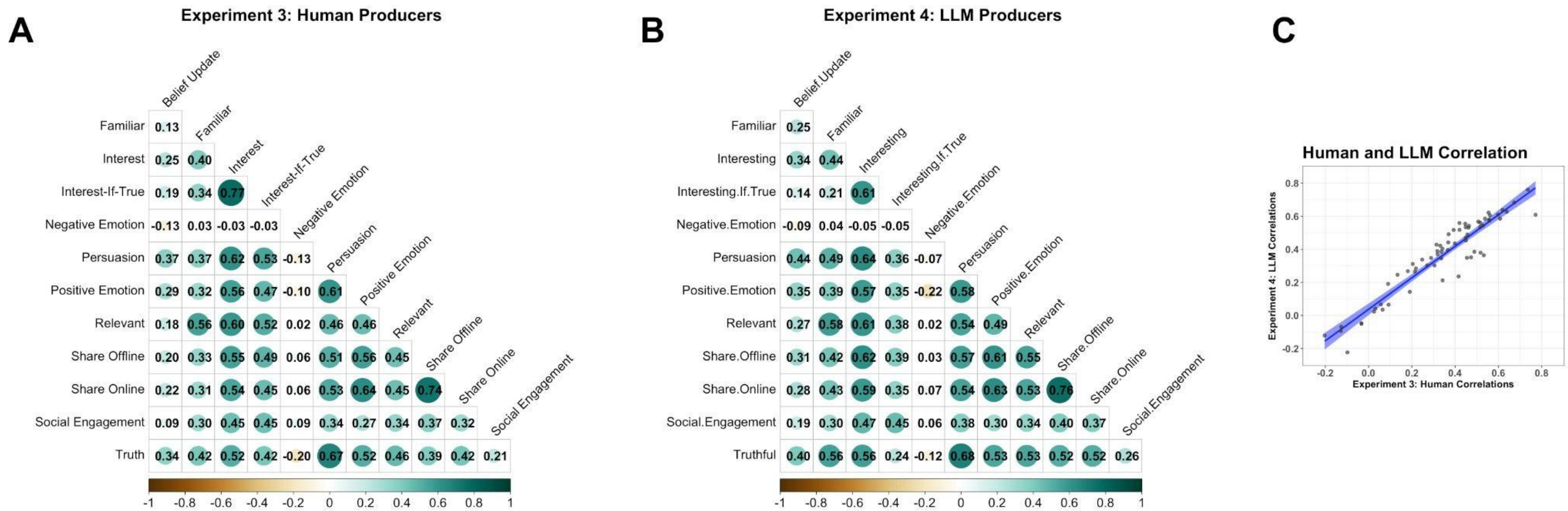

Fig. 4. Panel A: Correlation matrix for the human producers across dimensions (Experiment 3). Panel B: Correlation matrix for the LLM (GPT-3.5) producers across dimensions (Experiment 4). Panel C: Correlation between the correlation coefficients in Panel A (Human Producers) and Panel B (LLM Producers).

Table 2. Experiment 3: Results of the hierarchical backwards stepwise regression analysis for Persuasion, Belief Update, Share Online and Share Offline.

Persuasion							**Belief Update**					
Step	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**
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**					
Step	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**	**Dimension**	**Coefficient**	**95% CI**	**t**	**p**	**Marginal R^2**
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 ($p<.05$).

Discussion

The experiments reported indicate that, in the marketplace of ideas, truth wins. In each experiment, the True-Condition messages were more persuasive, led to stronger belief updating, and were more likely to be re-shared (online and offline) than the False-Condition messages (Experiment 1–4). In short, the True-Condition messages had more impact than the False-Condition messages. While the True-Condition messages increased participants' belief in the claims (Experiment 1–4), the False-Condition messages either did not change their belief in the claims (Experiment 2) or, more commonly, decreased their 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 (Experiment 3). 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, 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'. Yet, in the present study, relaxing the truth did not increase engagement; social engagement and intent to re-share 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 (*26*, *27*).

Our results also differentiate between the factors that drive message influence and spread. The main 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). So, in the experiments reported, truth was the gatekeeper of informational influence. This aligns with research showing that people update person-impressions only when the newly encountered information is believable (*28*). More broadly, it is consistent with the view of humans as 'information foragers', who analytically search the environment for valuable information (*29*). Truth was not the main driver of message spread. Instead, message spread—the intention to share the messages online or offline—was primarily driven by the positive emotions and anticipated social engagement the messages elicited (Experiments 1–4). This finding is testament to the importance of emotions in human decision-making (*30–32*), and aligns with research showing that messages that elicit high-arousal positive emotions tend to be more viral (*33*). The importance of positive emotions and social engagement indicates that people may prioritize social connection during information transmission (see also *34*, *35*), consistent with their behavior being guided by the core social motive to belong (*36*).

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 (*37–39*), and has informed strategies designed to combat it (*40–42*). However, the metaphor has been criticized for oversimplifying a complex issue, in large part because it conflates information spread with influence (*43*, *44*). 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 (*45*), 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 the finding that false information on the social media platform *Facebook* had no effect on COVID-19 vaccination intent (*46*). By contrast, the same study found that true-but-misleading content (e.g., *A healthy doctor died two 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 (see also *47–49*). Other contextual factors that moderate informational influence include: the communication channel (*50*, *51*), information frequency (*52–54*), perceived consensus (*55–57*), and the characteristics of the audience, especially their political identity (*58*). 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 Supplement 1 for details).

Misinformation is a significant societal issue, as demonstrated by the hyper-partisan false claim that the 2020 US presidential election was rigged, which in turn fueled the riots at the US Capitol (*59*). When stripped of contextual factors, the present study demonstrates that truthful messages persuade, untruthful messages dissuade, and these outcomes are driven by perceived message truth. Although messages from the True Condition were more likely to be shared than those from the False Condition, this 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. These findings indicate that people are predisposed to the truth—both as information producers and consumers—consistent with the finding that the majority of online misinformation is spread by a small group of supersharers (*60*). Taken together, the experiments reported indicate that in the marketplace of ideas truth wins. We note that the experiments reported sampled participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (*61*), and that LLMs are predominantly trained on data from English-speaking people from WEIRD societies (*62*). It is therefore important to test if our findings replicate in non-WEIRD societies.

Methods

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, including their personal views on current events, social issues, politics and pop culture. The information they share will vary along a continuum of truthfulness, ranging from complete falsehoods to misleading statements, half-truths, mostly accurate but with minor inaccuracies, to complete truths. That is, truthfulness is not always binary, and cannot always be fact-checked (e.g., in the case of personal opinions). The present study recognises this, and elicits a large number of messages from participants (humans and LLMs; 5469 unique messages in total) that vary in how truthful they are perceived to be by others. This allows us to test the extent to which a message's perceived truth (along with a range of other dimensions) affects its persuasive impact and transmission potential.

Experiments 1 and 2: The Persuasion Game

In Experiments 1 and 2 the task was to design persuasive messages. This was incentivised by offering a US$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 producing 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 Condition. 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. 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. 116 participants self-identified as female, 165 as male, 1 as non-binary and 1 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 US (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), 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 US$5.50 for approximately 25-35 minutes work (median duration 29 minutes).

Experiment 1: Human Raters. 1710 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 9 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 randomised 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 were sampled from a different producer and included five messages from each of the three experimental conditions (True, False, Unconstrained). 785 participants self-identified as female, 909 as male, 8 as non-binary, 4 as trans female and 4 as trans male. Participants were aged 19–99 years (*M* = 40.47, *SD* = 11.63). Most participants were based in the US (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), college educated (69%), politically progressive (51%; 29% conservative) and frequent social media users (86% were daily users). Each participant was paid US$3.50 for approximately 15-25 minutes work (median duration 18 minutes).

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 API. True- and False-Condition messages were generated by instantiating one of the prompt templates shown in Table 3. 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 invocation called for 8 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 6 messages was selected at random from the respective pool of 80 after discarding anomalies, yielding a total of 180 messages (2 x 15 x 6).

Table 3. Prompts provided to GPT-3.5 to generate LLM-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. Your task: write 8 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 persuasive 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 8 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 but must contain false facts (information you believe to be false).

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. 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 6 human-produced messages (3 from the True Condition and 3 from the False Condition) and 6 LLM-produced messages (3 from the True Condition and 3 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. 114 participants self-identified as female, 182 as male and 1 as non-binary (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 US (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), college educated (68%), politically progressive (47%; 32% conservative) and frequent social media users (89% were daily users). Each participant was paid US$3.50 for approximately 15-25 minutes work (median duration 18 minutes).

Materials

Thirty-two claims were developed and pre-tested. These included, "*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 pre-tested 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 were 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 Supplement 1 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, e.g., ‘*To the best of your knowledge, how truthful is the post?*’, *Not at All* (1), *Slightly* (2), *Somewhat* (3), *Very Much* (4), *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, then write a message designed to persuade others of the claim. In the True Condition participants were told 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 the person who produced the most persuasive messages would be paid a US$100 bonus. After the instructions page, participants were asked three multiple-choice questions to demonstrate their understanding of the instructions (see Supplement 2); 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 (see Figure 5) an input box was shown below the claim, containing placeholder text asking the participant to write a persuasive message supporting the claim. 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 3 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.

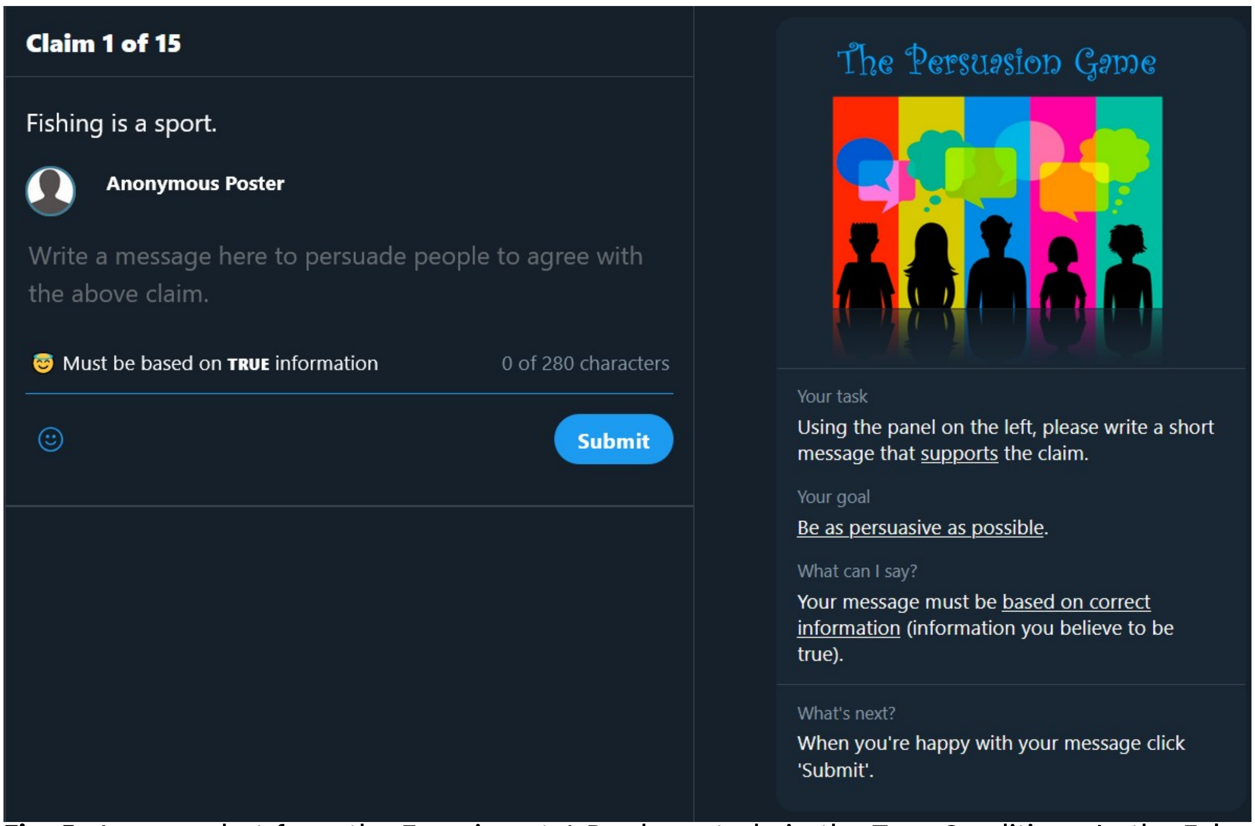

Fig. 5. A screenshot from the Experiment 1 Producer task, in the True Condition. In the False Condition the prompt below the message input box read “Must be based on FALSE information” (with a devil emoji) and the panel on the right (in the “What can I say?” section) read “Your message must be <u>based on misinformation</u> (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 <u>any information you like</u> 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 “<u>Gain as much attention as possible</u>”. Screenshots for each condition in each experiment are included in Supplement 3.

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, 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 each message they were taken to a debrief page and given a completion code. The completion time was approximately 18 minutes.

Statistical Analysis

The data were analyzed using linear mixed effects modeling (including the backwards stepwise regression analyses). The random effects structure included by-producer, by-rater and by-claim random intercepts. This allowed us to account for variation among the producers, the raters and the claims. All analyses were performed and all figures were created in R (*64*). Statistical models were estimated using the lmer() function of the lmerTest (*65*, *66*) package. The statistical analyses were pre-registered: https://aspredicted.org/see_one.php and the data, R Notebooks and Supplementary Materials are provided on the Open Science Framework: https://osf.io/t6sq4/

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 incentivised by offering a US$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. 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. 139 participants self-identified as female, 142 as male, 1 as non-binary and 1 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 US (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), 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 US$5.50 for approximately 25-35 minutes work (median duration 26 minutes).

Experiment 3: Human Raters. 1710 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 9 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 randomised 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 were sampled from a different producer and included five messages from each of the three experimental conditions (True, False, Unconstrained). 740 participants self-identified as female, 905 as male, 28 as non-binary, 7 as trans female and 8 as trans male (the remainder chose not to provide gender information). Participants were aged 18–94 years (M = 40.96, SD = 13.45). With 2 exceptions, all participants were based in the US, 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), college educated (60%), politically progressive (56%; 24% conservative) and frequent social media users (87% were daily users). Each participant was paid US$3.50 for approximately 15-25 minutes work (median duration 18 minutes).

Experiment 4: LLM Producer (GPT-3.5). As per Experiment 2, the LLM-produced messages were generated via the Python API for GPT-3.5 Turbo. The prompts for each condition are given in Table 4.

Table 4. Prompts provided to GPT-3.5 to generate LLM-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 8 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 8 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 Producers. Following the same design as Experiment 2, 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.

Experiment 4: Human Raters. 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 6 human-produced messages (3 from the True Condition and 3 from the False Condition) and 6 LLM-produced messages (3 from the True Condition and 3 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. 118 participants self-identified as female, 177 as male, 1 as non-binary and 1 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 US (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), college educated (70%), politically progressive (49%; 30% conservative) and frequent social media users (87% were daily users). Each participant was paid US$3.50 for approximately 15-25 minutes work (median duration 19 minutes).

Example Messages. Table 5 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 the True- and False-Condition messages that were rated as High and Low on the Share Online dimension, sampled from the The Attention Game in Experiment 3 (Human Producers) and Experiment 4 (LLM Producers).

Table 5. Examples Messages from Experiments 1-4.

Persuasion Game			
		High Persuasion	**Low Persuasion**
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]**
Attention Game			
		High Share Online	**Low Share Online**
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]**
	TRUE	Dreaming of a greener future? Well, did you know that satellites help monitor	Space is the place!!! And we're just a pale blue dot.

Experiment 4: LLM Producer		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]**	**[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*, respectively.

References

1. J. Milton, *Areopagitica*, *1644* (1868).
2. E. Pertwee, C. Simas, H. J. Larson, An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nat Med* **28**, 456–459 (2022).
3. K. M. d'I. Treen, H. T. P. Williams, S. J. O'Neill, Online misinformation about climate change. *WIREs Climate Change* **11**, e665 (2020).
4. J. Green, W. Hobbs, S. McCabe, D. Lazer, Online engagement with 2020 election misinformation and turnout in the 2021 Georgia runoff election. *Proceedings of the National Academy of Sciences* **119**, e2115900119 (2022).
5. S. Lewandowsky, U. K. H. Ecker, J. Cook, S. van der Linden, J. Roozenbeek, N. Oreskes, Misinformation and the epistemic integrity of democracy. *Current Opinion in Psychology* **54**, 101711 (2023).
6. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
7. *Abrams vs United States* (1919)vol. 250.
8. R. Dawkins, *The Selfish Gene* (Oxford University Press, 1989).
9. N. Walter, R. Tukachinsky, 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**, 155–177 (2020).
10. U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, M. A. Amazeen, The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol* **1**, 13–29 (2022).
11. P. Verma, The rise of AI fake news is creating a 'misinformation superspreader,' *Washington Post* (2023). https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/.
12. M. Steup, R. Neta, Epistemology. (2005).
13. R. C. Brownson, J. G. Gurney, G. H. Land, Evidence-Based Decision Making in Public Health. *Journal of Public Health Management and Practice* **5**, 86 (1999).
14. L. J. Savage, The Theory of Statistical Decision. *Journal of the American Statistical Association* **46**, 55–67 (1951).
15. R. M. Bond, R. K. Garretta, Engagement with fact-checked posts on Reddit. *PNAS Nexus*, pgad018 (2023).
16. M. Stella, E. Ferrara, M. D. Domenico, Bots increase exposure to negative and inflammatory content in online social systems. *PNAS* **115**, 12435–12440 (2018).
17. M. Orabi, D. Mouheb, Z. Al Aghbari, I. Kamel, Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management* **57**, 102250 (2020).
18. A. G. Greenwald, "Cognitive Learning, Cognitive Response to Persuasion, and Attitude Change" in *Psychological Foundations of Attitudes*, A. G. Greenwald, T. C. Brock, T. M. Ostrom, Eds. (Academic Press, 1968; https://cir.nii.ac.jp/crid/1360013170389570944), pp. 147–170.
19. W. J. McGuire, "Attitudes and Attitude Change" in *Handbook of Social Psychology*, L. Gardner, E. Aronson, Eds. (Random House, New York, 1985; https://cir.nii.ac.jp/crid/1571135650731642368)vol. 2, pp. 233–346.
20. P. Briñol, R. E. Petty, "A history of attitudes and persuasion research" in *Handbook of the History of Social Psychology* (Psychology Press, New York, NY, US, 2012), pp. 283–320.
21. A. Acerbi, J. M. Stubbersfield, Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences* **120**, e2313790120 (2023).

22. P. D. L. Howe, N. Fay, M. Saletta, E. Hovy, ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology* **14** (2023).
23. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends in Cognitive Sciences*, doi: 10.1016/j.tics.2023.04.008 (2023).
24. (Max) Hui Bai, J. G. Voelkel, johannes C. Eichstaedt, R. Willer, Artificial Intelligence Can Persuade Humans on Political Issues. OSF [Preprint] (2023). https://doi.org/10.31219/osf.io/stakv.
25. G. Spitale, N. Biller-Andorno, F. Germani, AI model GPT-3 (dis)informs us better than humans. *Science Advances* **9**, eadh1850 (2023).
26. P. Sumner, S. Vivian-Griffiths, J. Boivin, A. Williams, L. Bott, R. Adams, C. A. Venetis, L. Whelan, B. Hughes, C. D. Chambers, Exaggerations and Caveats in Press Releases and Health-Related Science News. *PLOS ONE* **11**, e0168217 (2016).
27. P. Sumner, S. Vivian-Griffiths, J. Boivin, A. Williams, C. A. Venetis, A. Davies, J. Ogden, L. Whelan, B. Hughes, B. Dalton, F. Boy, C. D. Chambers, The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ* **349**, g7015 (2014).
28. J. Cone, K. Flaharty, M. J. Ferguson, Believability of evidence matters for correcting social impressions. *PNAS*, 201903222 (2019).
29. P. Pirolli, S. Card, Information foraging. *Psychological Review* **106**, 643–675 (1999).
30. J. S. Lerner, Y. Li, P. Valdesolo, K. S. Kassam, Emotion and decision making. *Annual Review of Psychology* **66**, 799–823 (2015).
31. N. Schwarz, G. L. Clore, Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology* **45**, 513 (1983).
32. Y. Kashima, A. Coman, J. V. T. Pauketat, V. Yzerbyt, Emotion in Cultural Dynamics. *Emotion Review*, 175407391987521 (2019).
33. J. Berger, K. L. Milkman, What Makes Online Content Viral? *Journal of Marketing Research* **49**, 192–205 (2012).
34. A. Lyons, Y. Kashima, How Are Stereotypes Maintained Through Communication? The Influence of Stereotype Sharedness. *Journal of Personality and Social Psychology* **85**, 989–1005 (2003).
35. A. E. Clark, Y. Kashima, 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**, 1028 (2007).
36. S. T. Fiske, *Social Beings: Core Motives in Social Psychology* (John Wiley & Sons, 2018; 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).
37. S. van der Linden, *Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity* (W. W. Norton & Company, New York, 2023).
38. J. Zarocostas, How to fight an infodemic. *The Lancet* **395**, 676 (2020).
39. D. J. Rothkopf, Opinion | When the Buzz Bites Back, *Washington Post* (2003). https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/.
40. A. Kozyreva, P. Lorenz-Spreen, S. M. Herzog, U. K. H. Ecker, S. Lewandowsky, R. Hertwig, A. Ali, J. Bak-Coleman, S. Barzilai, M. Basol, A. J. Berinsky, C. Betsch, J. Cook, L. K. Fazio, M. Geers, A. M. Guess, H. Huang, H. Larreguy, R. Maertens, F. Panizza, G. Pennycook, D. G. Rand, S. Rathje, J. Reifler, P. Schmid, M. Smith, B. Swire-Thompson, P. Szewach, S. van der Linden, S. Wineburg, Toolbox of individual-level interventions against online misinformation. *Nat Hum Behav*, 1–9 (2024).

41. R. A. Blair, J. Gottlieb, B. Nyhan, L. Paler, P. Argote, C. J. Stainfield, Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology* **55**, 101732 (2024).
42. L. Q. Tay, M. J. Hurlstone, T. Kurz, U. K. H. Ecker, A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology* **113**, 591–607 (2022).
43. S. Altay, M. Berriche, A. Acerbi, Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media + Society* **9**, 20563051221150412 (2023).
44. F. M. Simon, C. Q. Camargo, Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic.' *New Media & Society* **25**, 2219–2240 (2023).
45. H. Mercier, *Not Born Yesterday: The Science of Who We Trust and What We Believe* (Princeton University Press, 2020; https://www.degruyter.com/document/doi/10.1515/9780691198842/html).
46. J. Allen, D. J. Watts, D. G. Rand, Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science* **384**, eadk3451 (2024).
47. G. T. Kumkale, D. Albarracín, P. J. Seignourel, 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**, 1325–1356 (2010).
48. T. Prike, L. H. Butler, U. K. H. Ecker, Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online. *Sci Rep* **14**, 6900 (2024).
49. P. Briñol, R. E. Petty, Source factors in persuasion: A self-validation approach. *European Review of Social Psychology* **20**, 49–96 (2009).
50. P. Breves, Persuasive communication and spatial presence: a systematic literature review and conceptual model. *Annals of the International Communication Association* **47**, 222–241 (2023).
51. S. Chaiken, A. H. Eagly, Communication modality as a determinant of message persuasiveness and message comprehensibility. *Journal of Personality and Social Psychology* **34**, 605–614 (1976).
52. A. Hassan, S. J. Barber, The effects of repetition frequency on the illusory truth effect. *Cogn. Research* **6**, 38 (2021).
53. L. Hasher, D. Goldstein, T. Toppino, Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior* **16**, 107–112 (1977).
54. G. Pennycook, T. D. Cannon, D. G. Rand, Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General* **147**, 1865–1880 (2018).
55. S. Lewandowsky, J. Cook, N. Fay, G. E. Gignac, Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Mem Cogn* **47**, 1445–1456 (2019).
56. L. H. Butler, N. Fay, U. K. H. Ecker, Social Endorsement Influences the Continued Belief in Corrected Misinformation. *Journal of Applied Research in Memory and Cognition*, doi: 10.1037/mac0000080 (2022).
57. S. E. Asch, Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes* **58**, 295–303 (1951).
58. J. J. V. Bavel, A. Pereira, The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences* **22**, 213–224 (2018).
59. J. Heine, The Attack on the US Capitol: An American Kristallnacht. *Protest* **1**, 126–141 (2021).
60. S. Baribi-Bartov, B. Swire-Thompson, N. Grinberg, Supersharers of fake news on Twitter. *Science* **384**, 979–982 (2024).
61. J. Henrich, S. J. Heine, A. Norenzayan, The weirdest people in the world? *Behav Brain Sci* **33**, 61–83; discussion 83-135 (2010).
62. M. Atari, M. J. Xue, P. S. Park, D. Blasi, J. Henrich, Which humans? (2023).
63. S. Altay, E. de Araujo, H. Mercier, "If This account is True, It is Most Enormously Wonderful":

Interestingness-If-True and the Sharing of True and False News. *Digital Journalism* **0**, 1–22 (2021).
64. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing (2013); http://www.R-project.org/.
65. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82**, 1–26 (2017).
66. D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear mixed-effects models using Eigen and S4. *R package version* **1** (2013).
67. I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. 'Sandy' Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, M. Wellman, Machine behaviour. *Nature* **568**, 477–486 (2019).
68. N. E. Friedkin, F. Bullo, How truth wins in opinion dynamics along issue sequences. *Proceedings of the National Academy of Sciences* **114**, 11380–11385 (2017).

Acknowledgements

Funding: Office of National Intelligence and Australian Research Council grant NI210100224 (N.F., A.P., P.D.L.H., and Y.K.).

Author contributions: Conceptualization: N.F., A.P., P.D.L.H., and Y.K. Methodology: N.F., A.P., P.D.L.H., Y.K., K.R., and B.W. Investigation: K.R., and B.W. Visualization: N.F. Funding acquisition: N.F., A.P., P.D.L.H., and Y.K. Project administration: K.R. and B.W. Writing – original draft: N.F. Writing – review & editing: N.F., A.P., P.D.L.H., Y.K., K.R., and B.W.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: Data, Analytic Code, Study Materials and Supplementary Materials available on the OSF, and pre registration documents accessible in AsPredicted.