Harvard Kennedy School Misinformation Review[1]
November 2025, Volume 6, Issue 6

DOI: https://doi.org/10.37016/mr-2020-189
Website: https://misinforeview.hks.harvard.edu

Research Note

People are more susceptible to misinformation with realistic AI-synthesized images that provide strong evidence to headlines

The development of artificial intelligence (AI) allows rapid creation of AI-synthesized images. In a pre-registered experiment, we examine how properties of AI-synthesized images influence belief in misinformation and memory for corrections. Realistic and probative (i.e., providing strong evidence) images predicted greater belief in false headlines. Additionally, we found preliminary evidence that paying attention to properties of images could selectively lower belief in false headlines. Our findings suggest that advances in photorealistic image generation will likely lead to greater susceptibility to misinformation, and that future interventions should consider shifting attention to images.

Authors: Sean Guo (1), Yiwen Zhong (2), Xiaoqing Hu (1) (3)
Affiliations: (1) Department of Psychology, The University of Hong Kong, China, (2) Department of Psychology and Human Development, Vanderbilt University, USA, (3) HKU-Shenzhen Institute of Research and Innovation, China
How to cite: Guo, S., Zhong, Y., & Hu, X. (2025). People are more susceptible to misinformation with realistic AI-synthesized images that provide strong evidence to headlines. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(6).
Received: July 16th, 2025. Accepted: October 27th, 2025. Published: November 10th, 2025.

Research questions

- Do the properties (realism, evidence strength, and surprisingness) of AI-synthesized images predict belief in false headlines?
- How do properties of AI-synthesized images influence correction effectiveness?
- How does examining realism, evidence strength, and surprisingness of images influence subsequent belief in headlines?
- How are digital literacy, conspiracist tendencies, and analytical thinking associated with headline discernment?

Research note summary

- This pre-registered study examined whether participants' subjective ratings of image realism, image surprisingness, and evidence strength predicted their belief in false headlines and the effectiveness of corrections to those headlines. We compared belief in headlines across two experiments to examine the effect of completing image ratings prior to belief ratings.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

- False headlines that were paired with realistic images and images that provided strong evidence were believed to a greater extent, and preliminary evidence showed that rating properties of images may decrease belief in false headlines. Corrections were generally memorable and effective at reducing belief.
- Our findings highlight that advances in generative AI image models could lead to increased misinformation susceptibility. Carefully examining the properties of images could be an effective intervention to reduce belief in misinformation, but further research is needed.

Implications

In past years, "photographs" of politicians being arrested, the pope in high fashion, and world landmarks ablaze circulated the internet. Many of these images were subsequently debunked as AI-generated and thus entirely fictional. Proliferation of AI-generated visual misinformation may exacerbate existing public health issues (Heley et al., 2022), reduce trust in media (Karnouskos, 2020; Ternovski et al., 2022), and influence voter attitudes (Diakopoulos & Johnson, 2021; Dobber et al., 2021). Despite this threat, how different properties of images are associated with initial belief in AI-generated visual misinformation and the effectiveness of corrections remains uncertain.

Findings from our experiment showed that the realism of images and the amount of evidence they provided to headlines were significant positive predictors of belief in false headlines. Therefore, we suggest that social media platforms should develop algorithms that can categorize these dimensions of images (i.e., realism and evidence strength) and detect whether these images may be AI-synthesized. When users upload images to social media, the algorithms could then identify highly realistic images that provide strong evidence to the accompanying text as greater potential misinformation risks, prioritizing content reviews of those posts. Fact-checking organizations could benefit from adopting similar algorithms. While fact-checking has proven effective in reducing the intent to share misinformation (Yaqub et al., 2020), the sheer volume of AI-generated images circulating online makes it challenging to address each one individually. Therefore, an automated system that identifies and prioritizes the most persuasive or potentially harmful misinformation could help fact checkers focus their efforts.

Following correction, belief in the misinformation was not significantly associated with perceived realism or evidence strength, indicating that corrective interventions are similarly effective across different image types. This finding further reinforces the idea that social media platforms or fact checkers should develop algorithms that automatically detect persuasive misinformation. If corrective measures yield comparable outcomes for both persuasive and non-persuasive misinformation, it may be more efficient to prioritize resources toward identifying and correcting persuasive misinformation.

Participants who scrutinized images and headlines to rate realism, evidence strength, and surprisingness prior to assessing headline believability exhibited selectively lower belief in misinformation headlines. This preliminary finding suggests that interventions that direct focus to images when evaluating the credibility of headlines could reduce belief in false headlines that are paired with convincing AI-synthesized images. Therefore, social media platforms and government organizations should consider updating existing media literacy guidelines with greater emphasis on paying attention to AI-synthesized images (Guo et al., 2025). However, given the small effect size, further research is warranted to clarify the role of image-focused attention in mitigating misinformation beliefs. For example, researchers could test accuracy nudges that prompt users to carefully examine images compared to those that only encourage checking the accuracy of a textual claim.

Our results showed that individual levels of digital literacy are positively correlated with headline discernment, while a greater tendency to hold conspiracy beliefs is related to decreased discernment. These findings are promising and suggest that institutions can focus on expanding the reach of current

interventions aimed at improving digital literacy (e.g., integrating digital literacy courses into school curricula), which may be more practically feasible than developing and testing new interventions specifically tailored to AI-generated visual misinformation. Additionally, our findings suggest that existing interventions targeted at reducing conspiracy beliefs may improve detection of false headlines accompanied by AI-synthesized images, thereby extending the established benefits of such interventions. This broader efficacy could strengthen the case for their wider implementation by social media platforms, educational institutions, and governmental agencies. By demonstrating utility beyond conspiracy-related content, these interventions may be positioned as versatile tools for enhancing misinformation resilience across multiple formats and modalities. Despite prior associations between analytical thinking and headline accuracy discernment (Pennycook & Rand, 2021), we found that analytical thinking (as measured by the Cognitive Reflection Test) was not related to discernment. Therefore, if practitioners aim to implement existing interventions to enhance discernment between AI-synthesized and authentic image-headline pairs, those targeting digital literacy and reducing conspiracy beliefs should take precedence over interventions designed solely to increase analytical thinking.

Findings

Finding 1: People are more susceptible to false headlines with realistic and probative (i.e., providing strong evidence) images.

We first examined how image properties influence initial belief in false headlines using a model with initial false headline belief as the outcome variable, centered values of realism, and evidence strength and image surprisingness as predictors. To account for repeated measures, we included participant and headline as random intercept effects. As seen in Figure 1, realism (b = 0.36, SE = 0.01, p < .001) and evidence (b = 0.29, SE = 0.01, p < .001) are positive predictors of initial belief in headlines. Surprisingness does not predict initial belief (b = -0.01, SE = 0.02, p = .688). This suggests that images that were perceived as more realistic and more probative are associated with increased belief.

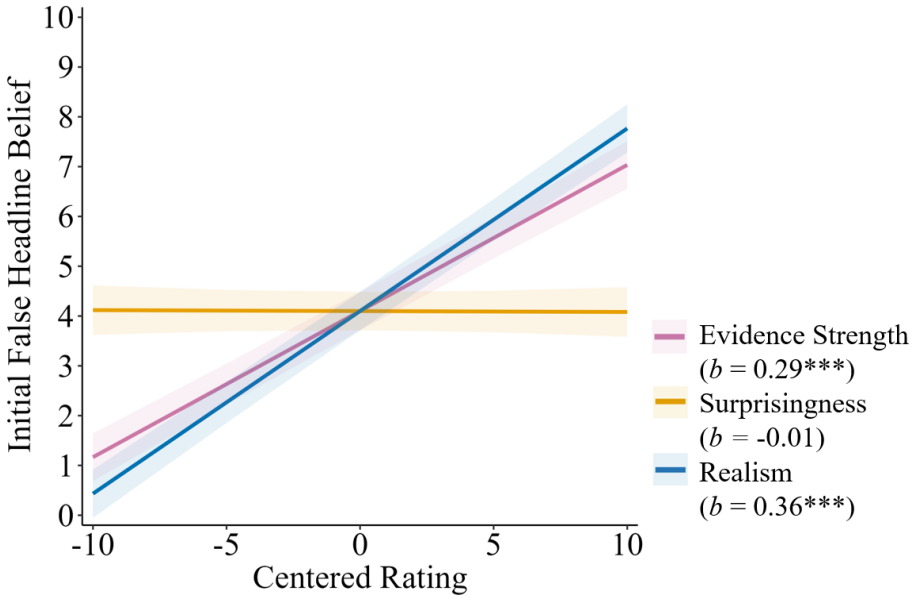

Figure 1. Regression lines between properties of images and initial false headline belief in Experiment 1 showing that greater realism and evidence strength predicted initial belief in false headlines*. Shaded regions represent 95% confidence intervals. *** p < .001.*

Finding 2: Corrections reduce belief in false headlines.

Before examining how properties of images influence belief reduction, we first verified that corrections reduced initial beliefs in false headlines using a model with belief as the outcome variable and phase (initial and immediate post-correction) as a predictor. To account for repeated measures, we included random intercept effects of participant and headline. Compared to initial beliefs (M = 4.14, *95% CI* = 4.05, 4.23), belief in false headlines is lower in the immediate post-correction phase (M = 2.13, *95% CI* = 2.05, 2.22), b = -2.01, SE = 0.05, p < .001, indicating that corrections successfully reduce beliefs in false headlines.

Finding 3: Properties of images are not related to post-correction belief reduction in headlines or memory for corrections.

We next investigated whether properties of images are linked to post-correction belief reduction. We used a model with post-correction beliefs as the outcome variable, centered values of realism, evidence strength, and image surprisingness as predictors, with initial belief and ratings of surprise to corrections as covariates. To account for repeated measures, we included random intercept effects of participant and headline. We found that neither realism (b = 0.02), evidence strength (b = 0.01), nor image surprisingness (b = 0.01) significantly predict post-correction belief in false headlines (ps > .080).

Using the same model but with memory for corrections as the outcome variable, we found that, again, neither realism (b = -0.01), evidence strength (b = -0.00), nor image surprisingness (b = -0.01) predict memory for corrections (ps > .504). In sum, the properties of images do not seem to be related to belief reduction or memory for corrections.

Finding 4: Rating image properties selectively decreases belief in false headlines.

To examine the effect of rating image properties on subsequent belief ratings, we conducted supplementary Experiment S1 (n = 150), in which a new group of participants rated belief in the same headlines without first rating properties of images (Appendix A). We used a model with initial belief in headlines as the outcome variable and experiment and headline veracity as predictors. To account for repeated measures, we included random intercept effects of participant and headline. Headline veracity (b = 0.52, SE = 0.41, p = .210) and experiment (b = -0.13, SE = 0.13, p = .316) do not significantly predict belief in headlines, but there was a significant interaction (b = 0.29, SE = 0.09, p < .001). As seen in Figure 2, participants in Experiment S1 (M = 4.41, *95% CI* = 4.31, 4.52) have higher beliefs than those in Experiment 1 (M = 4.14, *95% CI* = 4.05, 4.23) in false headlines, b = 0.28, SE = 0.13, p = .041. Belief in true headlines does not differ between Experiment 1 (M = 4.81, *95% CI* = 4.71, 4.90) and S1 (M = 4.78, *95% CI* = 4.68, 4.89), b = -0.02, SE = 0.13, p = .882. This suggests that completing ratings of image realism, surprisingness and evidence strength selectively decreases beliefs in false headlines.

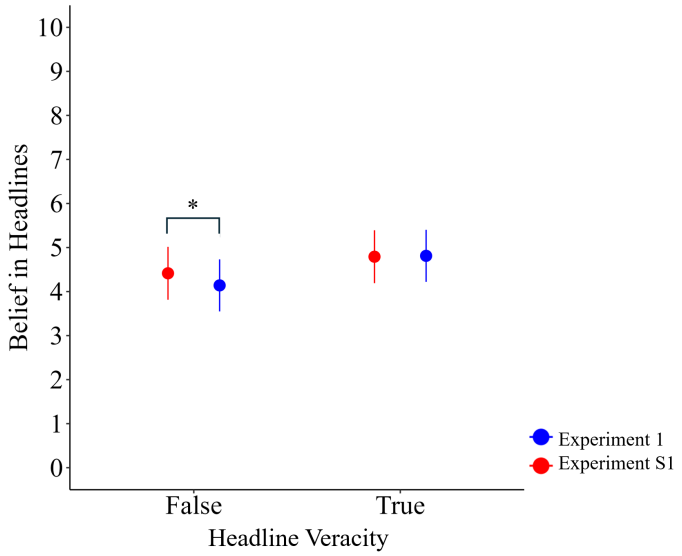

Figure 2. Comparison between belief in false and true headlines between Experiment 1 and Experiment S1, showing that belief in false headlines was lower in Experiment 1, while belief in true headlines remained similar across experiments*. Error bars represent 95% confidence intervals. *p < .05.*

Finding 5: Greater digital literacy and lower levels of conspiracist tendencies predict improved accuracy discernment.

Finally, we examined whether individual differences predict initial discernment between true and false headlines. Using a model with discernment (belief in true headlines minus belief in false headlines) as the outcome variable, and z-scored Cognitive Reflection Test (CRT), Generic Conspiracist Beliefs (GCB) and digital literacy scores as predictors (Figure 3), we found that greater GCB scores predict lower discernment (β = -0.26, *SE* = 0.10, *p* = .011), and greater digital literacy scores predict increased discernment (β = 0.21, *SE* = 0.10, *p* = .030), while CRT scores did not predict discernment, (β = 0.10, *SE* = 0.10, *p* = .304).

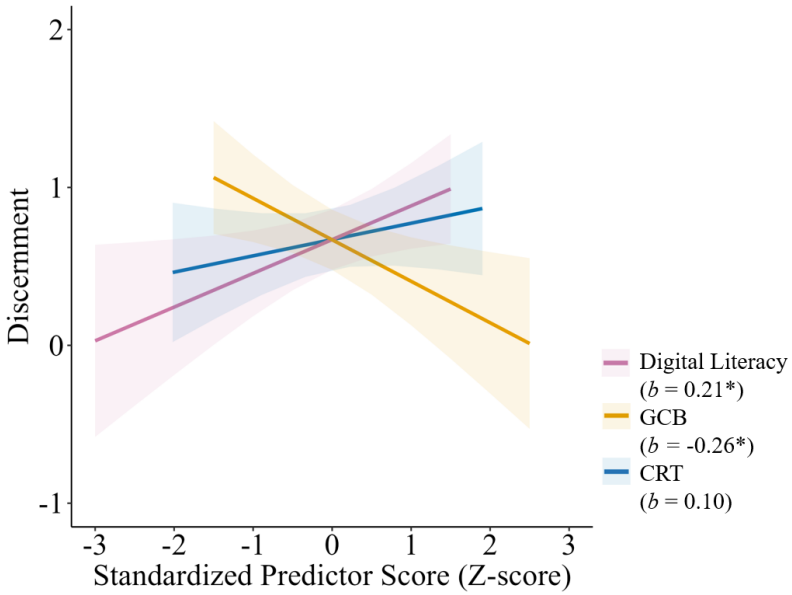

Figure 3. Regression lines between standardized CRT, digital literacy and GCB scores and initial accuracy discernment (belief in true headlines – belief in false headlines). *Accuracy discernment is positively predicted by digital literacy and negatively predicted by GCB scores. Shaded regions represent 95% confidence intervals. * p < .05.*

The full regression results are provided in Appendix B. We tested all prior findings with a sample that excluded participants who failed any attention checks and found results with the same direction and significance (see Appendix C). Additionally, possible differences in image-text co-reference (i.e., images may depict headline content to varying extents) may have influenced our results. Therefore, we defined co-reference as the proportion of words that are shown in the accompanying image and re-ran the analyses with this covariate, which yielded results with the same direction and significance (see Appendix D).

Methods

The experiment sought to answer the following research questions:

1) Is realism, evidence strength and surprisingness of AI-synthesized images related to belief in false headlines?
2) Are these properties of AI-synthesized images related to post-correction belief reduction and memory for corrections?
3) How does rating these properties of images influence subsequent belief in false headlines?
4) Are individual differences in digital literacy, conspiracist beliefs, and analytical thinking related to headline accuracy discernment?

Procedures

Stimulus ratings (phase 1): First, participants were instructed to rate how surprising and realistic images were, and how much evidence the image provided to a headline. Participants viewed 40 image-headline pairs individually in a random order. In each trial, they first saw an unlabeled real or AI-generated image and rated how surprising and realistic it was. Then, they saw the corresponding headline for that image and rated how much evidence the image provided to the headline.

Initial belief (phase 2): Participants were then told that they would see the same headlines again and rate how much they believed them. They then saw the same 40 headline-image pairs individually in a random order and rated their belief in the contents of the headline. We chose to include initial belief ratings in a separate task from phase 1 to minimize the effects of completing stimulus ratings on beliefs.

Correction/Affirmation task (phase 3): Next, participants were told that they would receive corrections or affirmations to the headlines they previously viewed. Participants read corrections and affirmations to all 40 headlines in a randomized order and rated how surprised they were upon receiving this information. Headlines were shown without images alongside a label stating “this headline is TRUE” for true headlines, and “this headline is FALSE” for false headlines.

Immediate post-correction belief (phase 4): Participants were then told to rate whether they thought headlines were true or false based on the information they received in phase 3. They then saw all 40 headlines again without images in a random order and rated their belief in the headline.

Delayed post-correction memory (phase 5): After one week, participants were instructed to complete a memory test by trying to remember whether headlines were true or false based on corrections and confirmations presented one week ago. Participants viewed all 40 headlines again individually in a randomized order and rated whether they remembered if the headline was corrected or affirmed in phase 3 of the experiment. It is important to note that ratings in each phase used discrete sliders (on a scale

from 0 to 10) instead of radio buttons, which may have introduced variability in participant responses due to potential ambiguity in scale interpretation.

Questionnaires (phase 6): Participants completed a battery of questionnaires, including a ten-item digital literacy scale (Hargittai, 2009), seven cognitive reflection test (CRT) questions (Oldrati et al., 2016; Thomson & Oppenheimer, 2016), and 15 questions from the Generic Conspiracist Beliefs scale (GCB) (Brotherton et al., 2013).

At the end of the experiment, participants were debriefed on the nature of the experiment. They then indicated which AI-generated and real images they had seen prior to the experiment. Trials containing these images were then excluded from analyses, as preregistered (18 AI-generated [0.4%] and 53 real [1.3%] trials excluded). For more details on the procedure, please refer to Appendix E.

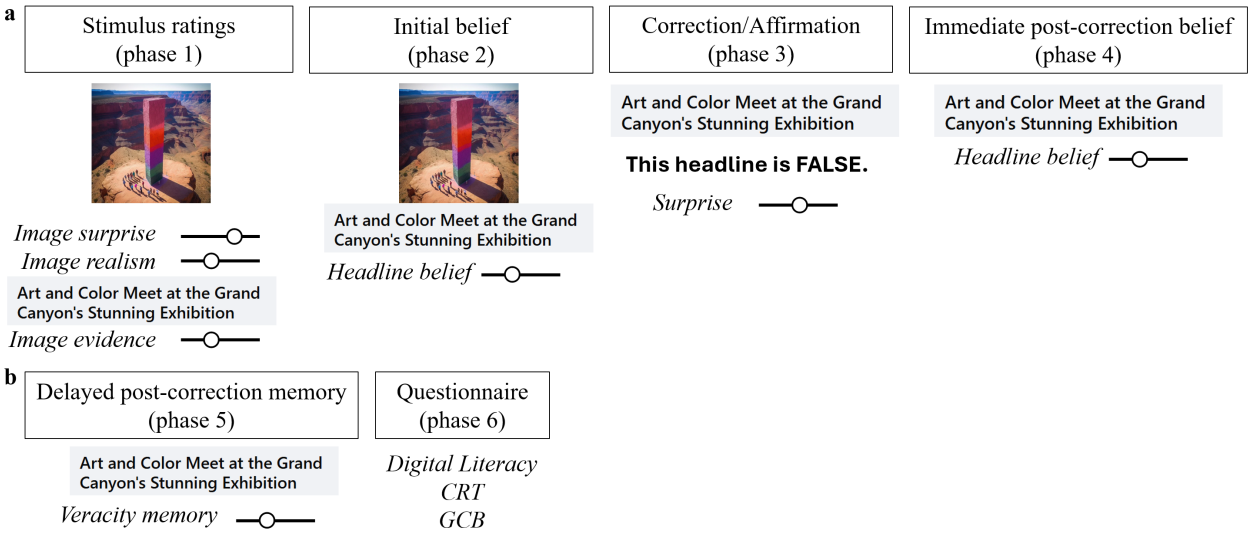

Figure 4. Procedure timeline and example trials for each phase in the false headline condition on a) Day 1 and b) Day 8. *Each slider (simplified for the figure) represents ratings participants made during each trial.*

Participants

After exclusions, the final sample included 212 participants from Prolific. We excluded participants if there were multiple submissions from the same user (n = 2), if there was a high likelihood of an automated submission, flagged by Qualtrics security measures (n = 7), if they provided more than 80% identical responses in the initial belief rating or immediate post-correction belief rating stage (n = 2), if they failed at least two out of three attention checks (i.e., the headline read "For this question, please rate 7 on the scale below. This is an attention check.") on the first day of the experiment (n = 12), or if they did not return one week later for the second part of the experiment (n = 20). After exclusions, the final sample included 119 men and 92 women, and one who did not report their gender from Prolific (M_{age} = 42.4, SD = 13.6, range = 20 to 104 years old). Participants currently resided in the United States, spoke English as their first language, completed at least 10 prior submissions, and had an approval rating of at least 90% on Prolific.

Stimuli

We wrote our own false headlines to match the content of AI-generated images, which we took from a Reddit community dedicated to sharing AI generated images from the website Midjourney (reddit.com/r/midjourney). We chose images that emulated photographs in composition, content, and style, as they have the greatest potential to mislead. We found true headlines and corresponding real images online from reputable news sources, and we re-used some from on a previous study on the role of images in news (Smelter & Calvillo, 2020). We created both types of headlines to emulate the style of news headlines on Facebook to improve external validity (see Figure 5).

We pretested images to ensure a wide distribution of realism and evidence strength, and pretested headlines to ensure a wide distribution of believability. The final experiment contained 40 headline-image pairs (20 AI-generated images with false headlines and 20 real images with true headlines). Both false and true headlines contained news about animals, accidents, natural disasters, and strange phenomena. We avoided political and health-related topics in our headlines to minimize effects of prior attitudes on belief.

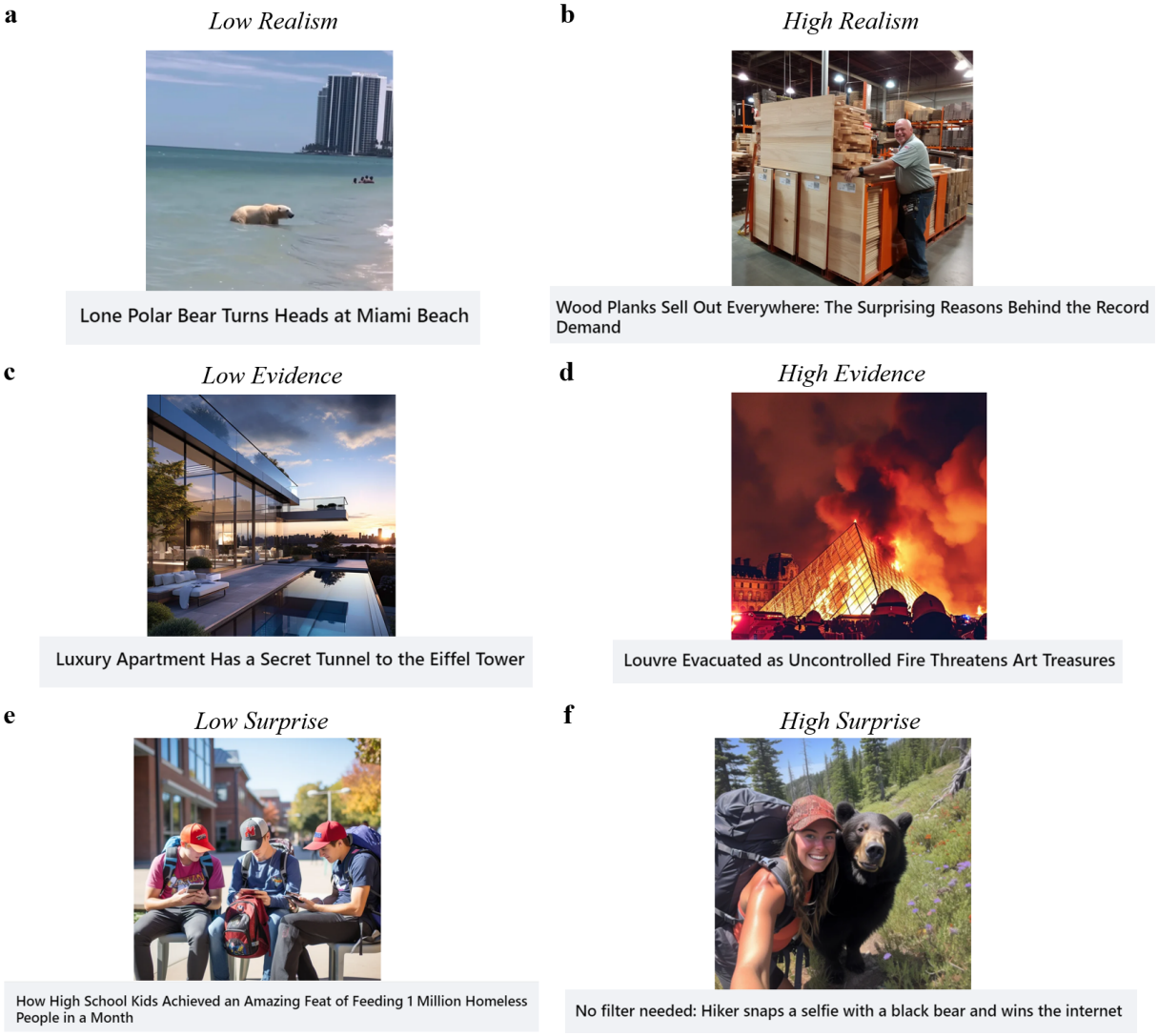

Figure 5. Example false headlines and AI-generated images. *a) Low realism image, b) high realism image, c) low evidence image, d) high evidence image, e) low surprisingness image, f) high surprisingness image. Images were found on a Reddit community dedicated to sharing AI-generated images created by the software Midjourney (reddit.com/r/midjourney), and headlines were written to match image content.*

Bibliography

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00279

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media and Society*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Guo, S., Swire-Thompson, B., & Hu, X. (2025). Specific media literacy tips improve AI-generated visual misinformation discernment. *Cognitive Research: Principles and Implications*, *10*(1), 1–11. https://doi.org/10.1186/s41235-025-00648-z

Hargittai, E. (2009). An update on survey measures of web-oriented digital literacy. *Social Science Computer Review*, 27(1), 130–137. https://doi.org/10.1177/0894439308318213

Heley, K., Gaysynsky, A., & King, A. J. (2022). Missing the bigger picture: The need for more research on visual health misinformation. *Science Communication*, *44*(4), 514–527. https://doi.org/10.1177/10755470221113833

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138–147. https://doi.org/10.1109/tts.2020.3001312

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia*, *91*, 499–508. https://doi.org/10.1016/j.neuropsychologia.2016.09.010

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Smelter, T. J., & Calvillo, D. P. (2020). Pictures and repeated exposure increase perceived accuracy of news headlines. *Applied Cognitive Psychology*, *34*(5), 1061–1071. https://doi.org/10.1002/acp.3684

Ternovski, J., Kalla, J., & Aronow, P. (2022). Negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety*, *1*(2). https://doi.org/10.54501/jots.v1i2.28

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/s1930297500007622

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020, April 21). Effects of credibility indicators on social media news sharing intent. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376213

Acknowledgements
The authors would like to thank Dr. Briony Swire-Thompson for providing valuable comments on the manuscript and Dr. Ziqing Yao for assisting with data analysis.

Funding
X.H. discloses support for the research from the Ministry of Science and Technology of China STI2030-Major Projects (No. 2022ZD0214100), National Natural Science Foundation of China (No. 32171056), and the General Research Fund of Hong Kong Research Grants Council (No. 17614922).

Competing interests
The authors declare no competing interests.

Ethics
This research was approved by the Human Research Ethics Committee of the University of Hong Kong (EA210341). Participants provided informed consent prior to completing the study. Gender information was determined from demographic information collected and defined by Prolific.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/OJKYPQ. Pre-registration is available at: https://osf.io/fw863

Appendix A: Experiment S1

Experiment S1 examined belief in headlines without completing stimulus ratings prior to belief ratings.

Participants

We recruited 160 participants with Prolific Academic. We excluded participants due to multiple submissions from the same user (n = 2) and high likelihood of an automated submission informed via Qualtrics security measures (n = 4). We also excluded participants if they admitted to answering randomly in the honesty check (n = 2), or if they admitted to looking up headlines during the study without specifying which ones (n = 1). One participant's responses were not recorded on Qualtrics due to technical issues (n = 1). In our final sample of 150 participants, there were 82 males and 68 females. Ages ranged between 21 to 81 (M = 45.75, SD = 14.66).

Procedure

The instructions we provided to participants were identical to those in the initial belief phase (phase 2) in the main experiment. After providing informed consent, we showed 20 false and 20 true headlines individually in a random order to participants. For each, they indicated their belief on a 0 (*definitely false*) to 10 (*definitely true*) scale.

Appendix B: Linear mixed models results

Table B1. *Fixed and random effects of a model measuring initial belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.14	0.19	22.36	< .001
Image Surprise	-0.01	0.02	-0.40	.688
Realism	0.36	0.01	24.80	< .001
Evidence	0.29	0.01	20.36	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.80		1.34	
Headline (Intercept)	0.49		0.70	
Model equation: initial belief in false headlines ~ 1 + realism + evidence + image surprise + (1 | subject) + (1 | headline)				

Table B2. *Fixed and random effects of a model measuring difference in headline belief between the main experiment and Experiment S1.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.54	0.21	21.23	< .001
Veracity	0.52	0.41	1.28	.210
Experiment	-0.13	0.13	-1.00	.316
Veracity*Experiment	0.29	0.09	3.39	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.25		1.12	
Headline (Intercept)	1.67		1.29	
Model equation: initial belief ~ 1 + veracity + experiment + veracity* experiment + (1 | subject) + (1 | headline)				

Note: Veracity and Experiment were contrast coded such that '0.5' represented true headlines and the main experiment, while -0.5 represented false headlines and Experiment S1.

Table B3. *Fixed and random effects of a model measuring difference in false headline belief before and after corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.14	0.21	20.04	< .001
Post-correction phase	-2.01	0.05	-36.86	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.72		1.31	
Headline (Intercept)	0.66		0.81	
Model equation: belief in false headlines ~ 1 + phase + (1 | subject) + (1 | headline)				

Table B4. *Fixed and random effects of a model measuring post-correction belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	1.44	0.16	8.97	<.001
Realism	0.02	0.01	1.13	.260
Evidence	0.01	0.01	0.61	.542
Image surprise	0.01	0.01	0.52	.606
Correction surprise	0.03	0.02	1.91	.056
Initial belief (covariate)	0.17	0.02	9.99	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.73		1.65	
Headline (Intercept)	0.14		0.38	
Model equation: immediate post-correction belief ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + (1 | subject) + (1 | headline)				

Table B5. *Fixed and random effects of a model measuring delayed post-correction memory for corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	8.14	0.20	41.15	<.001
Realism	-0.01	0.02	-0.66	.510
Evidence	-0.00	0.02	-0.19	.853
Image surprise	-0.01	0.02	-0.67	.504
Correction surprise	0.00	0.02	0.15	.880
Initial belief (covariate)	-0.16	0.02	-8.06	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.57		1.60	
Headline (Intercept)	0.38		0.62	
Model equation: delayed post-correction memory ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + (1 | subject) + (1 | headline)				

Appendix C: Linear mixed models results with co-reference

The following tables add an additional co-variate (co-reference) which accounts for the proportion of words from the headline that are also depicted in the image. This co-reference score was calculated by providing the following prompt to ChatGPT 4o, via the OpenAI API: "Create a list of objects in the attached image. Then, lemmatize the following sentence and create a list of lemmatized words: [HEADLINE TEXT HERE]. Finally, calculate the percentage of words from the lemmatized word list that also appear in the list of image objects, and output that as the score". The percentage was then z-scored and added as a covariate. The prompt was run on 14th August 2025 with a temperature of 0 and the default system prompt (i.e., "You are a helpful assistant").

Table C1. *Fixed and random effects of a model measuring initial belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	0.60	0.22	2.71	.009
Image Surprise	-0.01	0.02	-0.78	.436
Realism	0.38	0.01	26.71	< .001
Evidence	0.29	0.01	21.14	< .001
Co-reference	-0.18	0.23	-0.79	.442
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	0.75		0.86	
Headline (Intercept)	0.49		0.70	
Model equation: initial belief in false headlines ~ 1 + realism + evidence + image surprise + co-reference + (1 | subject) + (1 | headline)				

Table C2. *Fixed and random effects of a model measuring difference in headline belief between the main experiment and Experiment S1.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.54	0.21	21.33	< .001
Veracity	0.58	0.41	1.41	.168
Experiment	-0.13	0.13	-1.00	.316
Co-reference	-0.25	0.21	-1.19	.243
Veracity*Experiment	0.29	0.09	3.39	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.25		1.12	
Headline (Intercept)	1.65		1.29	
Model equation: initial belief ~ 1 + veracity + experiment + co-reference + veracity* experiment + (1 | subject) + (1 | headline)				

Note: Veracity and Experiment were contrast coded such that '0.5' represented true headlines and the main experiment, while -0.5 represented false headlines and Experiment S1.

Table C3. *Fixed and random effects of a model measuring difference in false headline belief before and after corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.13	0.21	19.45	< .001
Post-correction phase	-2.01	0.05	-36.86	< .001
Co-reference	-0.13	0.27	-0.49	.634
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.72		1.31	
Headline (Intercept)	0.69		0.83	
Model equation: belief in false headlines ~ 1 + phase + co-reference (1 | subject) + (1 | headline)				

Table C4. *Fixed and random effects of a model measuring post-correction belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	1.44	0.16	8.85	< .001
Realism	0.02	0.01	1.13	.261
Evidence	0.01	0.01	0.63	.529
Image surprise	0.01	0.01	0.56	.578
Correction surprise	0.03	0.02	1.91	.056
Initial belief (covariate)	0.17	0.02	9.97	< .001
Co-reference	-0.06	0.13	-0.46	.652
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.73		1.65	
Headline (Intercept)	0.15		0.38	
Model equation: immediate post-correction belief ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + co-reference + (1 | subject) + (1 | headline)				

Table C5. *Fixed and random effects of a model measuring delayed post-correction memory for corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	8.14	0.20	40.41	< .001
Realism	-0.01	0.02	-0.66	.508
Evidence	-0.00	0.02	-0.19	.848
Image surprise	-0.01	0.02	-0.68	.494
Correction surprise	0.00	0.02	0.16	.877
Initial belief (covariate)	-0.16	0.02	-8.04	< .001
Co-reference	0.07	0.21	0.31	.759
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.57		1.60	
Headline (Intercept)	0.40		0.63	
Model equation: delayed post-correction memory ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + co-reference (1 | subject) + (1 | headline)				

Appendix D: Linear mixed models results excluding all attention check failures

Our primary analyses excluded participants who failed at least two out of three attention checks. However, including participants who failed one attention check may have biased our results. As a robustness check, we conduct the same analyses excluding participants who failed one attention check as well (*N* = 11).

Table D1. *Fixed and random effects of a model measuring initial belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.10	0.20	20.90	< .001
Image Surprise	-0.00	0.02	-0.12	.907
Realism	0.37	0.01	24.76	< .001
Evidence	0.29	0.01	20.33	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.82		1.35	
Headline (Intercept)	0.57		0.75	
Model equation: initial belief in false headlines ~ 1 + realism + evidence + image surprise + (1 | subject) + (1 | headline)				

Table D2. *Fixed and random effects of a model measuring difference in headline belief between the main experiment and Experiment S1.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.54	0.21	21.17	< .001
Veracity	0.52	0.41	1.27	.211
Experiment	-0.13	0.13	-1.06	.288
Veracity*Experiment	0.30	0.09	3.41	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.25		1.12	
Headline (Intercept)	1.68		1.30	
Model equation: initial belief ~ 1 + veracity + experiment + veracity* experiment + (1 | subject) + (1 | headline)				

Note: Veracity and Experiment were contrast coded such that '0.5' represented true headlines and the main experiment, while -0.5 represented false headlines and Experiment S1.

Table D3. *Fixed and random effects of a model measuring difference in false headline belief before and after corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	4.10	0.21	19.57	< .001
Post-correction phase	-2.07	0.06	-37.30	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	1.64		1.28	
Headline (Intercept)	0.68		0.83	
Model equation: belief in false headlines ~ 1 + phase + (1 | subject) + (1 | headline)				

Table D4. *Fixed and random effects of a model measuring post-correction belief in false headlines.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	1.33	0.16	8.40	< .001
Realism	0.02	0.02	1.22	.222
Evidence	0.01	0.01	0.99	.321
Image surprise	0.00	0.01	0.29	.776
Correction surprise	0.03	0.02	1.90	.058
Initial belief (covariate)	0.17	0.02	9.81	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.60		1.61	
Headline (Intercept)	0.12		0.35	
Model equation: immediate post-correction belief ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + (1 | subject) + (1 | headline)				

Table D5. *Fixed and random effects of a model measuring delayed post-correction memory for corrections.*

	Fixed Effects			
	Est/Beta	*SE*	*t*	*p*
Intercept	8.25	0.20	41.88	< .001
Realism	-0.01	0.02	-0.67	.504
Evidence	-0.00	0.02	-0.09	.926
Image surprise	-0.01	0.02	-0.58	.560
Correction surprise	0.01	0.02	0.32	.752
Initial belief (covariate)	-0.16	0.02	-7.84	< .001
	Random Effects			
	Variance		*SD*	
Subject (Intercept)	2.40		1.55	
Headline (Intercept)	0.37		0.61	
Model equation: delayed post-correction memory ~ 1 + realism + evidence + image surprise + correction surprise + initial belief + (1 | subject) + (1 | headline)				

Appendix E: Additional methodological details

Below, we provide more details on experiment pretesting, procedure, and participants.

Pretesting

Pretesting for headlines and images occurred in two waves. The first wave doubled as a behavioral pilot in which 39 participants first provided headline believability ratings for 40 image-headline pairs (20 false, 20 true), received corrections and affirmations to headlines, and rated their belief in headlines after a one-week delay. Then, we showed them the same 40 image-headline pairs again and asked them to rate the realism of images and the strength of evidence they provided to paired headlines. After we examined the rating distribution, we found that we did not have enough false headlines that were both unrealistic and provided weak evidence, and headlines that were both realistic and provided strong evidence. We therefore attempted to create more headlines-image pairs to round out the distribution and conducted a second wave of pretesting. Because we had already completed a behavioral pilot in the first wave, we decided to simplify the procedure in the second wave. Therefore, 49 participants directly provided believability, evidence strength and realism ratings to 14 image-headline pairs (11 false, 3 true).

Procedure

Stimulus ratings (phase 1): First, we told participants that they would be shown a series of images and headlines, that they would rate how surprising and realistic each image was, and how much evidence the image provided to a headline. They completed ratings on discrete scales ranging from 0 (*surprise: very unsurprising, realism: definitely fake, evidence: very weak*) to 10 (*surprise: very surprising, realism: definitely real, evidence: very strong*).

Initial belief (phase 2): We then told participants that they would see the same headlines again and rate how much they believed them. For each pair, they indicated their belief on a 0 (*definitely false)* to 10 (*definitely true)* discrete scale.

Correction/Affirmation task (phase 3): Next, we told participants that they would receive corrections or affirmations to the headlines they previously viewed, and that they would rate how surprised they were upon receiving the correction or affirmation. For each correction and affirmation, participants rated how surprised they were to learn this information on a discrete scale from 0 (*very unsurprised*) to 10 (*very surprised*).

Immediate post-correction belief (phase 4): We then told participants to rate whether they thought headlines were true or false based on the information they received in phase 3. They saw all 40 headlines again without images individually in a random order and completed belief ratings on a discrete scale from 0 (*definitely false*) to 10 (*definitely true*).

Delayed post-correction memory (phase 5): We invited participants to complete the second part of the experiment approximately seven days after completion of phase 4 (*M* interval between sessions = 165 hours, 58 minutes [6.9 days]). After receiving the invitation, they had 24 hours to complete the experiment. We told participants: “The following section will be a memory test. For each headline, please try to remember whether it was true or false, as indicated in the corrections and confirmations presented one week ago. Please adjust your rating according to your confidence. For example, if you are very

confident that the headline was true, rate "10." If you are slightly less confident, rate "9," and so on. If you have no recollection at all of whether it was true or false, rate "5." Please try your best to remember whether each headline was true or false". Participants then viewed all 40 headlines again individually in a randomized order and completed memory ratings using a discrete scale from 0 (*definitely false*) to 10 (*definitely true*). These memory ratings were subsequently reverse-coded such that higher values indicated greater memory for corrections.