



# Data Visualization or Visual Exemplars? Testing the Differential Effects of AI-Generated Visual Correction Enhancements

Yibing Sun<sup>1</sup>, Liwei Shen<sup>1,2</sup>, Ji Soo Choi<sup>1</sup>,  
Porismita Borah<sup>3</sup>, Michael W. Wagner<sup>1</sup>,  
Dhavan V. Shah<sup>1</sup>, and Sijia Yang<sup>1</sup>

## Abstract

This study investigates the efficacy of AI-generated visuals and potential credibility cues in correcting health-related misinformation. Using a pre-registered, factorial experiment ( $N=1,751$ ), we tested two AI-generated visual enhancements (visual exemplars, infographics) and two credibility boosters (source tagging, partisan neutrality in posting history). Findings revealed small but significant advantages of visual exemplars, but not infographics, over text-only corrections in reducing misbeliefs, primarily through mitigating psychological reactance. However, credibility cues did not significantly enhance correction effectiveness. Overall, the results suggest that AI-generated visual exemplars offer incremental persuasive benefits beyond textual correction alone.

## Keywords

misinformation, multimodal communication, LLM-generated visuals, psychological reactance

The rapid advancement of generative artificial intelligence (AI) introduces opportunities and challenges in combating health-related misinformation (Capraro et al., 2024; C. Chen & Shu, 2023; Shao et al., 2018). On the one hand, AI facilitates the creation of highly sophisticated multimodal “deepfakes” that blur the line between authentic

<sup>1</sup>School of Journalism and Mass Communication, University of Wisconsin-Madison, USA

<sup>2</sup>Department of Communication Arts, University of Wisconsin-Madison, USA

<sup>3</sup>Edward R. Murrow College of Communication, Washington State University, Pullman, USA

## Corresponding Author:

Sijia Yang, School of Journalism and Mass Communication, University of Wisconsin-Madison, 5160 Vilas Hall, 821 University Avenue, Madison, WI 53706, USA.

Email: [sijia.yang@alumni.upenn.edu](mailto:sijia.yang@alumni.upenn.edu)

and fabricated content (Barari et al., 2025; Q. Peng et al., 2024). On the other hand, multimodal large language models (LLMs) also have the potential to identify misleading content at scale (Liu et al., 2024) and produce compelling visuals (Barman et al., 2024), thereby potentially enhancing corrective efforts in health communication. However, recent findings suggest that fact-checking outputs from LLMs may sometimes impair users' ability to discern headline accuracy (DeVerna et al., 2024), underscoring the risks of over-relying on AI-generated corrections. While textual misinformation and correction have received extensive academic attention, research on the use of multimodal LLMs for visual correction is lagging behind these models' rapidly advancing visual capabilities. Admittedly, current evidence on the efficacy of incorporating visuals to enhance misinformation correction is mixed (Hameleers et al., 2020, 2023; Sundar et al., 2021; Young et al., 2018). One potential explanation is that many existing studies tend to lump different types of visuals together, dichotomously testing the presence or absence of visuals as an undifferentiated category. Not all visuals are created equal. Therefore, it is a crucial next step to unpack potential heterogeneous effects between different types of visuals and identify those that can effectively support misinformation correction. This study aims to leverage the visual-generating capacities of multimodal LLMs and makes an initial attempt to fill this gap.

Another possible explanation for the mixed evidence on the efficacy of visuals is that their effects might be contingent on the presence of other message features. One promising candidate is credibility cues, given the overall established roles of source credibility in persuasion (C. Shen et al., 2019; Vraga & Bode, 2017) and the importance of credibility, as well as its related concepts including trustworthiness and authenticity, in human-AI interaction (E. J. Lee, 2020). Given the move by both Meta and TikTok toward AI transparency by requiring identification and labeling of AI-generated content, the issue of signaling and establishing the credibility of AI-generated correction messages becomes even more relevant. Although individuals might still employ machine heuristics and perceive algorithms and AI as more objective than humans (Sundar & Kim, 2019), this perception is increasingly being challenged. In particular, efforts to use AI for misinformation correction face skepticism due to growing public awareness of AI hallucinations, where the model generates false, misleading, or nonsensical information (Y. Zhang et al., 2023). On issues related to medicine and health, it is important to empirically test whether adding expert endorsement might help boost the credibility of AI-produced visuals in correction messages and enhance these messages' effectiveness.

Furthermore, a growing number of health-related misinformation topics are increasingly politicized in the U.S. context, such as those related to COVID-19 masking and vaccination (Neumann et al., 2024; Zhou et al., 2024). Previous research has found that AI sources significantly reduced partisan-based motivated reasoning in evaluating a correction message's credibility (Moon et al., 2022). Building on these insights, this study further investigates whether and how partisan neutrality in posting history might influence correction effectiveness. We do not claim to exhaust all possible types of credibility cues that could have been deployed. Rather than evaluating correction messages in isolation, we focus on testing the effects of two types of credibility cues:

source tagging and partisan neutrality in posting history. Source tagging refers to explicitly labeling the source of the information, such as endorsement from experts, which may increase perceptions of credibility. In the U.S. context, partisan neutrality in posting history involves showing that the AI account has corrected misinformation from both Democratic and Republican parties, potentially reducing perceived political bias. Through studying these two types of credibility cues, we aim to highlight the broader importance of considering auxiliary message features when evaluating the effectiveness of AI-produced visual corrections.

This report shares findings from a pre-registered study that systematically varied and tested the efficacy of two types of AI-produced visuals (i.e., visual exemplars, data visualization presented as an infographic), along with two types of credibility cues (i.e., source tagging, partisan neutrality in posting history), in a between-subject factorial experiment with a large and diverse national sample. After detailing the theoretical rationale that motivated the experimental design and selection of message features, we present methodological specifics and key results regarding the main effects of AI-produced visuals and the moderation effects of incorporating credibility cues. Moreover, we further explored the indirect effects through two complementary mediating mechanisms (i.e., credibility perceptions, psychological reactance) to shed light on relevant psychological processes underlying effective misinformation correction, or the lack thereof, using AI-produced visuals. Although the mediation analyses were not part of our pre-registration, these indirect pathways help highlight the theoretical contributions of both significant results. Ultimately, this work aims to inform the development of evidence-based strategies that explore AI's unique multimodal capabilities to combat health-related misinformation in an era defined by rapid technological change.

## Literature Review

### *Visuals in Misinformation Correction*

According to past literature on visual persuasion, visuals enhance persuasion by making information more concrete, emotionally engaging, and memorable (Zillmann & Brosius, 2012). This unique capacity of visuals has already been exploited by creators of misinformation (Yang et al., 2023). Much research has unpacked different ways visuals can be incorporated into misinformation (Brennen et al., 2021; Hameleers et al., 2020; Hemsley & Snyder, 2018; Khan et al., 2023), including visual recontextualization, visual manipulation, and visual fabrication (Heley et al., 2022; King & Lazard, 2020; Yang et al., 2023). These strategies range from altering the context of an image to mislead viewers, to digitally manipulating content to distort reality, and even fabricating entirely new visuals to deceive audiences. The development of advanced multimodal LLMs further amplifies concerns about the proliferation of deepfakes, as these AI-generated multimodal falsehoods are becoming increasingly difficult to detect (C. Chen & Shu, 2023). Moreover, deepfakes often appear highly convincing, making them easier for wider acceptance and spreading (J. Lee & Hameleers, 2024).

Despite the documented concern that visuals might fuel the creation, acceptance, and spreading of falsehoods, visuals also hold significant potential as corrective enhancements by making corrections more engaging, credible, and easier to process (Dan & Coleman, 2024; Dixon et al., 2015; Domgaard & Park, 2021; Mena, 2023; Nyhan & Reifler, 2018). The development of multimodal LLMs offers the potential to accelerate and scale up context-specific visual corrections to combat the spread of multimodal misinformation. That said, empirical findings on the effectiveness of multimodal correction have been inconsistent. While some research suggests that adding visuals enhances correction messages (Sundar et al., 2021; Young et al., 2018), others found that multimodal fact-checking efforts are not significantly more effective in debunking textual disinformation than they are for multimodal disinformation (Hameleers et al., 2023). One possible explanation for these mixed results is that many studies focus on comparing different modalities without accounting for the distinct effects of specific visual types, which can play a crucial role in shaping outcomes.

Visuals commonly used in online (mis)information can be presented in different ways, including photographs, videos, memes, and visualizations, as summarized in the agenda for studying visual (mis)information (Y. Peng et al., 2023). This study focuses on visual exemplars and data visualization presented as infographics, as these formats are both widely used and extensively studied in the context of health communication (H. S. Kim et al., 2012; Mena, 2023).

Data visualizations play a crucial role in simplifying and presenting quantitative information in an easily digestible format. Journalists frequently use infographics in news stories to provide readers with accurate information and enhance their understanding (Hoffman, 2019; Porter & Russell, 2018). Empirical studies have demonstrated that infographics increase engagement, improve information processing (Comello et al., 2016), enhance message credibility (E. J. Lee & Kim, 2015; Sundar, 2008), and help correct misperceptions (Domgaard & Park, 2021; Mena, 2023; Nyhan & Reifler, 2010). Visual exemplars, on the other hand, are a visual presentation of vivid and relatable examples to represent a broader category or support a more generalizable claim or argument. For example, showing a patient receiving vaccination at a clinic serves as a concrete example to reinforce the pro-vaccination argument. They function similarly to testimonials, drawing their effectiveness from the principles proposed in exemplification theory, which posits that people tend to rely on concrete, case-based heuristics when forming judgments about broader issues (Zillmann & Brosius, 2000). Furthermore, visual exemplars can also increase emotional processing and overall message engagement, which in turn improves message persuasiveness (Ophir et al., 2019). While not limited to visual formats, a meta-analysis by Bigsby et al. (2019) found that individuals exposed to messages featuring exemplars reported stronger persuasion outcomes, highlighting their powerful role in shaping attitudes.

Creating effective visuals requires more production efforts than crafting text alone. However, advancements in AI, particularly the multimodal LLMs, have significantly streamlined the process of visual production. For example, multiple data visualization companies (e.g., Tableau, Visme, and Cognos Analytics by IBM) have incorporated AI

in data analysis and visualization. Using ChatGPT, one can generate data visualizations by automatically writing scripts based on input data, which the user can copy-paste and execute in their preferred programming language (e.g., through *R*'s *ggplot2* package, *Python*'s *Matplotlib* library) to produce the desired data visualizations. This workflow significantly reduces the time and effort required for data processing and script writing from scratch, although code refinement and human supervision are still necessary. Additionally, these multimodal LLMs can produce highly customizable images based on textual prompts (e.g., MidJourney), which can help produce visual exemplars tailored to specific corrective contexts. Despite these technical capabilities, to our knowledge, no research has empirically examined whether such multimodal LLMs can be leveraged to facilitate misinformation correction. As AI continues to transform the media landscape, such investigation is crucial to optimize the use of AI-generated visuals for combating misinformation effectively.

Thus, based on theoretical and empirical insights, we pre-registered the following hypotheses:

**H1:** Compared to text-only corrections, corrections with AI-generated visual exemplars and data visualizations will further reduce misperceptions.

### *Credibility Boosters: Partisan Neutrality and Expert Endorsement*

AI-generated visuals show potential for reducing misbeliefs, but their effectiveness might be contingent upon the presence of other message features. In response to growing concerns about misinformation and the authenticity of online content, social media platforms have started emphasizing AI transparency. Both Meta and TikTok, for example, have implemented measures to label AI-generated content, aiming to help users distinguish between human- and machine-created materials. The role of AI extends beyond content creation to the way it is perceived as a source of correction. Source credibility plays a critical role in shaping how audiences evaluate and trust corrective information (Pennycook & Rand, 2019; Vraga & Bode, 2017). We examine two potential credibility cues: partisan neutrality in posting history and AI labeling with expert endorsement.

Partisan neutrality in posting history is a promising credibility cue to help enhance the effectiveness of misinformation correction, particularly in the U.S. context of high political polarization (Gawronski et al., 2023). The machine heuristic, as described in the MAIN (Modality, Agency, Interactivity, and Navigability) model (Sundar, 2008), suggests that people perceive AI systems as more objective than humans because machines are seen as devoid of personal motives or affiliations. This perception makes AI-generated fact-checking less likely to trigger psychological reactance, thereby increasing credibility for AI-generated corrective content (Chung et al., 2023; Horne et al., 2019; Huang & Sundar, 2020; Sundar, 2020; Sundar & Kim, 2019). However, with the rise of generative AI and increased interactions with these AI tools, there is growing concern about the potential biases in AI (Wahbeh et al., 2023).

These expressed concerns are not groundless, as algorithms often reflect the biases embedded in their training data and the social contexts in which they operate (Quadflieg et al., 2022). Evidence indicates that AI exhibits bias related to race (Z. Chen, 2023), gender (L. Sun et al., 2024), and political orientation (Motoki et al., 2024; Rutinowski et al., 2024). For instance, both Motoki et al. and Rutinowski et al. found that ChatGPT exhibits a systematic political bias favoring the Democratic Party in the United States, as its responses align more closely with Democratic positions than Republican ones. Despite claims of neutrality by AI developers, these findings suggest that generative AI models may reinforce particular political narratives, raising concerns about their influence on public discourse and decision-making.

As a result, for fact-checkers interested in incorporating AI-produced visuals, it is crucial to add credibility cues that can demonstrate their partisan neutrality to build trust and credibility. This need is particularly pressing in health communication, where topics like vaccination have become highly politicized (Zhou et al., 2024). Prior research on misinformation correction has used cues such as addressing misclaims from political candidates across different parties to signal partisan neutrality (Jia & Liu, 2021; Li & Wagner, 2020). The presence of such partisan cues has been shown to influence individuals' beliefs. Based on these insights, we pre-registered the following hypotheses:

**H2a:** Partisan neutrality positively moderates the effects of AI-generated data visualization, such that the correction effects of AI-generated data visualization (vs. textual correction only) are enhanced when the AI fact-checker demonstrates a history of partisan neutrality (vs. no history of neutrality).

**H2b:** Partisan neutrality positively moderates the effects of AI-generated visual exemplars, such that the correction effects of AI-generated visual exemplars (vs. textual correction only) are enhanced when the AI fact-checker demonstrates a history of partisan neutrality (vs. no history of neutrality).

Beyond concerns about AI biases, the public is rightly suspicious of the credibility of AI-generated messages due to documented cases of AI hallucination, which can result in factual inaccuracies or misleading information, particularly in the context of misinformation. Thus, human validation becomes essential, and cues indicating expert validation may enhance the credibility and trustworthiness of AI-generated content, particularly in correction messages. Research consistently shows that reinforcing expert endorsement helps improve the credibility of correction messages (Vraga & Bode, 2017; Walter et al., 2020; J. Zhang et al., 2021), as they draw on authority heuristic to signal reliability and trustworthiness (Sundar, 2008). Therefore, based on this rationale, we pre-registered the following hypotheses:

**H3a:** Expert validation positively moderates the effects of AI-generated data visualization, such that the correction effects of AI-generated data visualization (vs. textual correction only) are enhanced when validated by experts (vs. AI-only).

**H3b:** Expert validation positively moderates the effects of AI-generated visual exemplars, such that the correction effects of AI-generated visual exemplars (vs. textual correction only) are enhanced when validated by experts (vs. AI-only).

### *Exploring Mechanisms of Visual Correction*

At the heart of the challenge of correcting misinformation lies a psychological tension between accuracy- versus identity-preserving motivations (Kunda, 1990; Li, 2025): on the one hand, individuals are driven by a motivation to seek accurate information, relying on credible evidence to guide belief updating; on the other hand, they are influenced by a strong desire to maintain existing beliefs, often resisting corrections that threaten their autonomy in decision-making and identity protection. Drawing on the concept of credibility (Flanagin et al., 2020; Huang & Wang, 2022; S. C. Kim et al., 2021) and psychological reactance theory (PRT) (Brehm & Brehm, 1981; Dillard & Shen, 2005; Rains & Turner, 2007; Y. Sun & Lu, 2022), we explored two complementary pathways that might account for the effects of AI-produced visual corrections: enhancing the credibility of corrective information and/or reducing resistance to it.

**Through Credibility.** The credibility literature distinguishes between multiple types of credibility, including the credibility of the *source* (Ecker et al., 2022) and the credibility of the *message* (Metzger et al., 2003). In this study, we focus on message credibility toward the correction, as our primary interest lies in how people perceive the correction of AI-generated visuals, rather than evaluating the broader credibility of AI as a source.

Credibility is widely studied in misinformation research, frequently examined either as a dependent variable (Flanagin et al., 2020) or as a mediator (Huang & Wang, 2022; S. C. Kim et al., 2021) influencing attitudinal changes and behavioral outcomes (Ha & Ahn, 2011). Visuals, in general, are multidimensional information that offers more nuanced and diverse content than text. Y. Peng et al. (2023) set an agenda for studying credibility perceptions of visual misinformation, emphasizing three objective visual features: color, composition, and content. While color and composition relate more to the aesthetic level of images (Lu & Shen, 2023; Y. Peng, 2022), content relates more to the topic, narrative style, contextual setting, and symbolic meaning conveyed through the visuals (Joo et al., 2014; Y. Sun et al., 2025).

For data visualization, they represent evidence in a visual format. Research has shown that when data are presented through visualizations, it enhances perceptions of vividness and credibility, fostering greater trust in the message (Henke et al., 2020). Visual exemplars, such as images of individuals receiving vaccinations, may serve as “seeing is believing” cues, enhancing the perceived credibility of the message. According to exemplification theory (Zillmann, 1999), the use of specific, concrete examples helps to make abstract or complex ideas, such as the safety and efficacy of vaccines, more tangible and relatable to the audience. Among textual, audio, and visual modalities, Sundar et al. (2021) found that video-based misinformation, serving

as a visual exemplar, was perceived by participants as more credible and was more likely to be shared.

*Through Psychological Reactance.* PRT suggests that in situations where individuals sense that their autonomy or freedom of choice is being restricted, they may experience a psychological state known as reactance, which motivates them to restore their perceived control or independence (Brehm & Brehm, 1981). Although Brehm and Brehm (1981) assert that reactance cannot be measured, Dillard and Shen (2005) reconceptualized it as a combination of negative cognitions and anger. Anger refers to the emotional response to feeling coerced or controlled, whereas negative cognitions refer to counter-arguing and other forms of mental resistance aimed at nullifying the perceived pressure (Rains & Turner, 2007).

Initially, PRT was widely applied in health campaign messages, such as anti-smoking messages (Jebai et al., 2023; LaVoie et al., 2017; Miller et al., 2006) and responsible alcohol consumption campaigns (Ringold, 2002). The core mechanism of reactance theory is that when individuals sense they are being told what to do or believe, they may push back to reassert their freedom of choice (Quick & Stephenson, 2008). This dynamic is particularly salient in public health contexts where messages may adopt a directive tone, which can unintentionally raise individuals' defenses (Miller et al., 2007).

More recently, scholars have recognized that misinformation correction efforts often encounter the same barrier: attempts to "correct" someone's belief can be perceived as an attack on freedom, activating reactance (Y. Sun & Lu, 2022). Misinformation correction inherently challenges the validity of individuals' existing beliefs, regardless of whether they pertain to health practices, political issues, or social topics (Lewandowsky et al., 2017). If the correction is presented in a way that feels dogmatic or threatening, recipients may experience anger (affective) and devise counterarguments to defend their position (cognitive), ultimately rejecting the corrective information (Featherstone & Zhang, 2020).

PRT offers a valuable theoretical framework for understanding both *why* corrective messages sometimes fail and *how* they might be adjusted to succeed. Previous research has identified various factors associated with reactance, such as using language that could elicit empathy (L. Shen, 2010) and incorporating narratives and other-referencing strategies (Gardner & Leshner, 2016). For this study, we are interested in the role of AI-generated visuals, more specifically data visualization and visual exemplars.

Data visualizations may reduce psychological reactance by addressing the negative cognition component, such as counterarguing. By presenting information in a clear, structured, and evidence-based manner, data visualizations minimize opportunities for individuals to generate counterarguments, as the logical and objective nature of the visuals makes the message harder to dismiss. This aligns with findings from Braverman's (2008) study, which demonstrated that informational persuasive messages are more effective for individuals who are highly involved and possess a strong need for cognition. For visual exemplars, their ability to reduce psychological reactance may stem from their narrative power and capacity to obscure their persuasive

intent, thereby reducing audience reactance and enhancing receptivity to the message (Dal Cin et al., 2004). Research has shown that narrative messages, whether in text format (Gardner & Leshner, 2016) or video format (Occa et al., 2015), can positively influence attitudes and intentions, suggesting that visual exemplars may similarly engage audiences by creating relatable and emotionally compelling contexts. Building on these two theoretical mechanisms, we pose this research question:

**RQ1:** Do psychological reactance or perceived credibility mediate the effects of AI-generated visuals on misinformation correction?

## Method

### Sample

We recruited 1,751 participants from the national panel maintained by the survey company Lucid in the summer of 2024 to participate in an online experiment administered using Qualtrics. A post hoc power analysis using G\*Power (Faul et al., 2007) confirmed that our sample size exceeded the threshold required to detect small effects with 95% power. To be eligible, participants needed to be parents, living in the United States, and can read and write in English. We further employed quota sampling to recruit two groups of parents of equal size ( $n < 990$  for each group): those with children younger than 9 years old (not eligible for HPV vaccination) and those with children aged 9 to 17 years who had not completed the HPV vaccination series. In other words, we excluded parents whose children had completed the HPV vaccination series given the focus on misinformation correction related to vaccine hesitancy. The collected sample included a higher proportion of female parents ( $n = 1,194$ ) compared to male parents ( $n = 550$ ). These parents and guardians reported having 1,035 boys and 735 girls in total. Detailed characteristics of the sample, including demographic variables, are presented in the Supplemental Materials. This study was preregistered on the Open Science Framework (OSF). All planned hypotheses, design details, and analysis procedures were specified in advance. The full preregistration document is available online: [https://osf.io/tyjra?view\\_only=65b87a0d22954fd08dd561e3a4d298ed](https://osf.io/tyjra?view_only=65b87a0d22954fd08dd561e3a4d298ed)

### Procedure

After providing consent, eligible participants were asked to complete a pre-treatment questionnaire measuring demographics and other pre-exposure covariates. Next, participants were randomized into 1 of the 14 conditions: 3 (AI-created visual enhancements: data visualization vs. visual exemplar vs. textual correction only)  $\times$  2 (partisan neutrality: balanced history vs. no history)  $\times$  2 (source tagging: AI only vs. AI with expert endorsement) + 1 (misinformation-only control) + 1 (questionnaire-only control). The core treatment stimuli took the form of a three-post Twitter/X feed, starting with a misinformation post, followed by a correction post, with neutral, non-vaccination-related posts interspersed between them. We fixated this Twitter/X feed on the

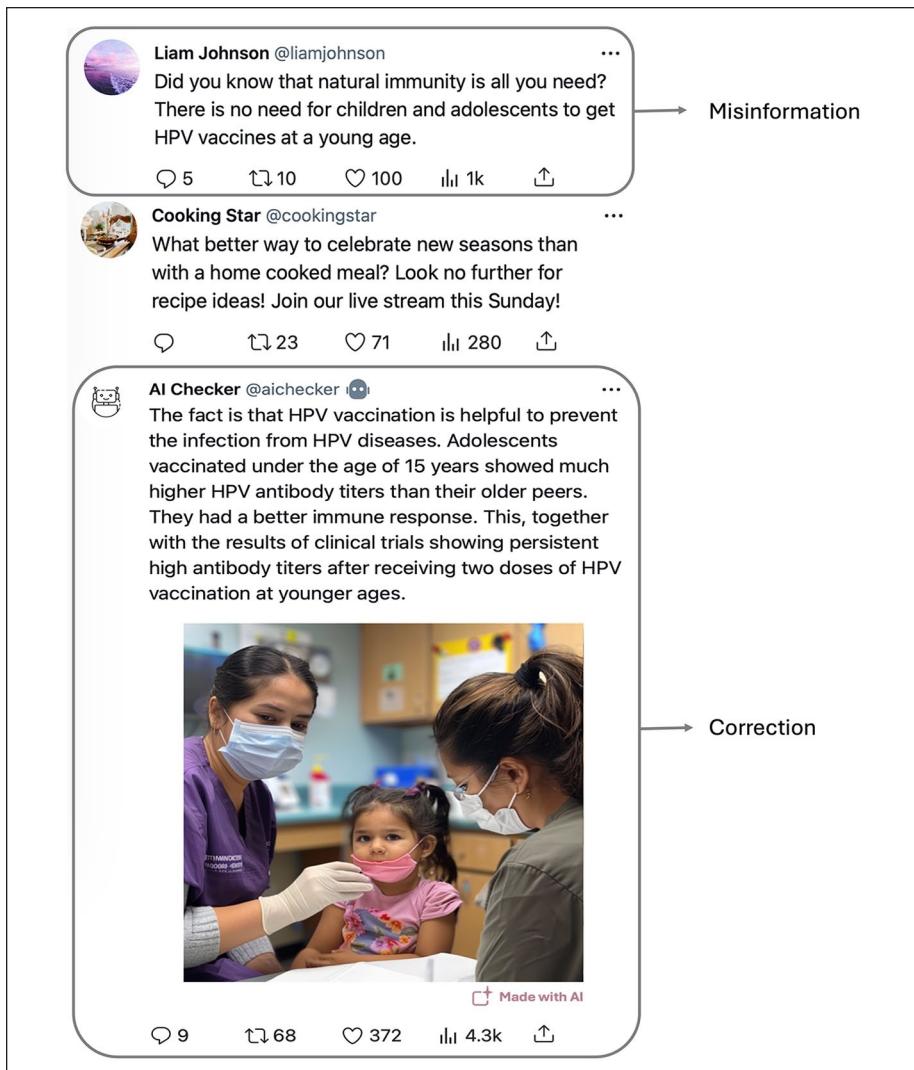
screen for at least 10 seconds before displaying the “next” button for participants to proceed to complete questions measuring key mediators and outcomes. The correction post varied by the presence and type of incorporated AI-generated images as well as source tagging (more details in the following section). There was no manipulation on either the misinformation post or the filler post. Importantly, to systematically manipulate partisan neutrality, we showed participants randomized to the *balanced history* conditions an additional pre-feed representing the AI account’s “history” *prior to* the core treatment stimuli. This pre-feed consisted of four posts (see Figures 1 and 2a for an example): two correcting misleading claims made by Joe Biden and two by Donald Trump, demonstrating the source’s neutrality.

Furthermore, participants in the *misinformation-only* condition did not view any correction posts or pre-feed. Those randomized to the *questionnaire-only* condition were not shown any Twitter/X feed at all, although they still needed to complete the questionnaire measuring key outcomes while skipping questions measuring mediators related to message evaluation. Before participants left the survey, we showed everyone a debriefing page with factual information about the safety and efficacy of HPV vaccines as well as links to credible sources. This step was taken to address ethical concerns associated with misinformation studies, ensuring that participants were provided with accurate and reliable information. All study materials, including the replication dataset, questionnaire, and experimental stimuli, are publicly available online: [https://osf.io/36j84/files/osfstorage?view\\_only=8dc191a59c3c4b4ebca12415b2fd5d7f](https://osf.io/36j84/files/osfstorage?view_only=8dc191a59c3c4b4ebca12415b2fd5d7f).

### *Stimuli Preparation*

In this study, we focused on the following misconception: “Children and adolescents are not sexually active so there is no need for them to get HPV vaccines at a young age.” We selected this misperception because in our pilot study, this misperception was highly predictive of parents’ vaccination hesitancy based on Hornik and Woolf’s (1999) approach, and it was relatively understudied in the literature on HPV misinformation and correction. We then developed the textual part of the correction message, with the core factual basis adapted from an existing fact-checking article (Taumberger et al., 2022). We provided ChatGPT with both the misinformation claim and the fact-checking article, and instructed it to summarize the argument and correct the misinformation in a format suitable for social media platforms while adhering to Twitter’s 280-character limit. This textual correction was held constant across all experimental conditions, while the accompanying static images and source-tagging were systematically varied.

The three image conditions included two treatment conditions with potential visual enhancements, AI-generated data visualization or visual exemplar on top of a baseline textual correction, and one control condition with textual correction only. Examples of stimuli for each condition are presented in Figure 2b. Data visualization was based on data from Rosenblum et al. (2022), which documented the impact of the HPV vaccine following its introduction in the United States. The visualizations were created using ChatGPT, where the data were uploaded, which then generated the R scripts that we



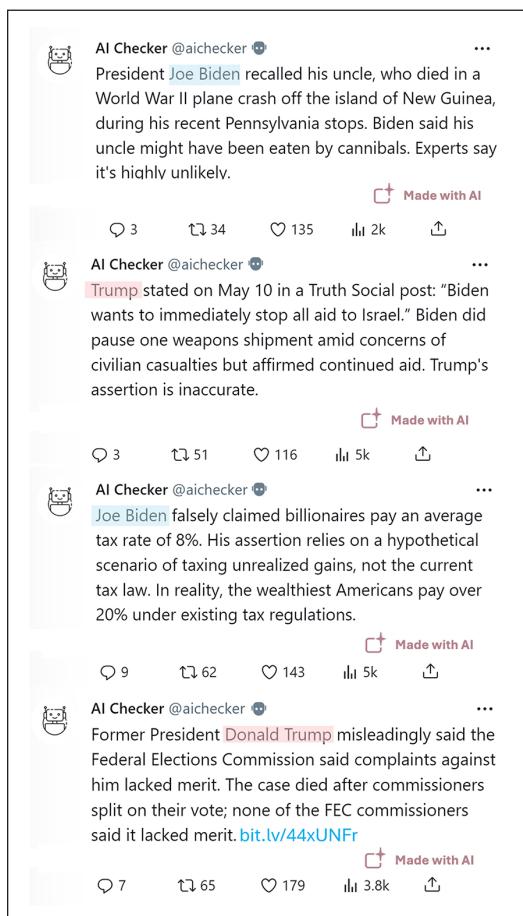
**Figure 1.** Example Stimuli.

ran to produce the visualizations. Visual exemplars were created using Midjourney, a text-to-image generative AI tool. The prompt instructed the AI to generate an image depicting three people in a clinical setting, featuring a child receiving a vaccination, accompanied by a healthcare practitioner and a parent. In line with calls to use multiple stimuli per condition to address message variability and enhance validity in experimental studies (Reeves et al., 2016; Slater et al., 2015), we generated a pool of eight exemplars varying the child's gender (male, female) and race (Asian, Black, Latino,

White), given that demographic characteristics can meaningfully influence how visual messages are perceived (Ratcliff et al., 2025). Each participant was shown one randomly selected visual exemplar from the pool.

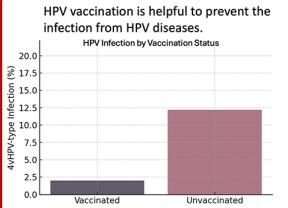
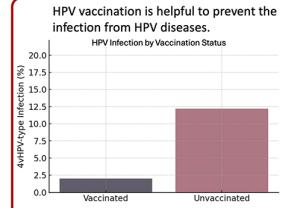
Regarding the construction of the pre-feed used to manipulate the AI account's *partisan neutrality* in posting history, we created a pool of misinformation correction posts explicitly debunking claims made by either Donald Trump ( $n=5$ ) or Joe Biden ( $n=5$ ). Two posts from the Trump pool and two posts from the Biden pool were randomly selected and ordered for display, ensuring that every participant in the *balanced history* condition was exposed to a balanced set of corrections across political parties. In the *no history* condition, no pre-feed was displayed.

Lastly, regarding the factor of source tagging, each correction post was labeled as either AI only ("Made with AI") or AI with expert endorsement ("Made with AI and



(a)

**Figure 2.** (continued)

	AI	AI with expert endorsement												
Data Visualization	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p>  <p>HPV Infection by Vaccination Status</p> <table border="1"> <thead> <tr> <th>Vaccination Status</th> <th>Infection (%)</th> </tr> </thead> <tbody> <tr> <td>Vaccinated</td> <td>~2.5</td> </tr> <tr> <td>Unvaccinated</td> <td>~12.5</td> </tr> </tbody> </table> <p>Notes: The data is from National Health and Nutrition Examination Survey in 2011 to 2014 showing 4+HPV-type infection among sexually experienced 14-24-year-old females by vaccination status.</p> <p>Made with AI</p>	Vaccination Status	Infection (%)	Vaccinated	~2.5	Unvaccinated	~12.5	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p>  <p>HPV Infection by Vaccination Status</p> <table border="1"> <thead> <tr> <th>Vaccination Status</th> <th>Infection (%)</th> </tr> </thead> <tbody> <tr> <td>Vaccinated</td> <td>~2.5</td> </tr> <tr> <td>Unvaccinated</td> <td>~12.5</td> </tr> </tbody> </table> <p>Notes: The data is from National Health and Nutrition Examination Survey in 2011 to 2014 showing 4+HPV-type infection among sexually experienced 14-24-year-old females by vaccination status.</p> <p>Made with AI and validated by the American Academy of Pediatrics</p>	Vaccination Status	Infection (%)	Vaccinated	~2.5	Unvaccinated	~12.5
Vaccination Status	Infection (%)													
Vaccinated	~2.5													
Unvaccinated	~12.5													
Vaccination Status	Infection (%)													
Vaccinated	~2.5													
Unvaccinated	~12.5													
Visual exemplars	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p>  <p>Made with AI</p>	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p>  <p>Made with AI and validated by the American Academy of Pediatrics</p>												
No image	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p> <p>Made with AI</p>	<p>AI Checker @aichecker</p> <p>The fact is that HPV vaccination is helpful to prevent the infection from HPV diseases. Adolescents vaccinated under the age of 15 years showed much higher HPV antibody titers than their older peers. They had a better immune response. This, together with the results of clinical trials showing persistent high antibody titers after receiving two doses of HPV vaccination at younger ages.</p> <p>Made with AI and validated by the American Academy of Pediatrics</p>												

**Figure 2.** Example Stimuli by Condition: (a) Example of Added Account History of Partisan Neutrality; (b) Examples of Corrections Varying AI-Created Visual Enhancements and Source Tagging.

Note. We generated a pool of eight visual exemplars. Participants assigned to the visual exemplar condition were shown one randomly selected exemplar from this pool.

validated by the American Academy of Pediatrics"). To ensure external validity, we used a labeling approach similar to that of Meta, incorporating icons and descriptive text.

## **Measurements**

**Misinformation Belief.** Participants' misbeliefs were assessed using a statement similar to the false claim in the stimuli that there is no need for children and adolescents to get HPV vaccines at a young age. Participants were asked to indicate the extent to which they believed the statement to be true or false, using a 5-point Likert scale ranging from "*Definitely false*" to "*Definitely true*." Higher scores indicate greater misbeliefs about the HPV vaccine. Group means and standard deviations by experimental condition are provided in the Supplemental Materials.

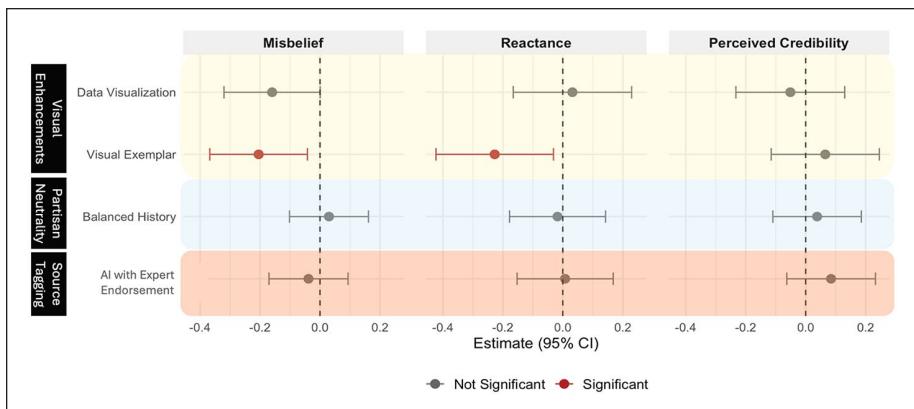
**Psychological Reactance.** Psychological reactance was operationalized as a latent construct indicated by anger and counterarguing. Anger was measured by asking participants to recall their feelings while reading the correction post. They rated four items—annoyed, irritated, angry, and aggravated—on a 7-point Likert scale ( $\alpha=.94$ ,  $M=2.69$ ,  $SD=1.68$ ). Higher scores indicate more anger toward the content. Counterarguing was measured through three items adapted from Silvia (2006) and Gardner & Leshner (2016), such as "Did you criticize the tweet while reading it?" Participants responded on a 7-point Likert scale ( $\alpha=.86$ ,  $M=3.34$ ,  $SD=1.73$ ), where higher scores indicate a greater tendency to counterargue against the post.

**Credibility.** Participants were asked to evaluate the post on six bipolar semantic differential items on a 7-point Likert scale ( $\alpha=.91$ ,  $M=4.53$ ,  $SD=1.44$ ): fair/unfair, inaccurate/accurate, biased/unbiased, doesn't tell the whole story/tells the whole story, imbalanced/balanced, and cannot be trusted/can be trusted higher scores represent greater perceptions of the post's credibility.

## **Analytical Strategies**

To examine whether there was any systematic imbalance of covariates (e.g., demographic variables) across conditions, we performed a multinomial logistic regression analysis, in which all covariates were included as independent variables to predict assignment to conditions. A likelihood ratio test was then conducted to compare the full model (with all covariates) to a baseline model (without predictors). The result was not statistically significant ( $\chi^2 (234)=240.46$ ,  $p=.37$ ), confirming that randomization was successful.

H1 and H2 were pre-registered prior to data collection to ensure transparency and rigor in testing our theoretical predictions. To test H1, we conducted linear regression analyses with the textual correction only condition set as the reference group. The model also included dummies representing the factors of partisan neutrality and AI source tagging, with the no history condition and the AI with human endorsement condition set as the reference groups, respectively. Including all three factors ensures



**Figure 3.** Summary of Estimated Main Effects.

Note. The reference groups used are textual-only for visual type, no history provided for neutrality, and AI only for source tagging. The table presents unstandardized coefficients. Psychological reactance was measured as a composite score indicated by anger and counterarguing. Alternative model specification including covariates is presented in the Supplemental Materials. Results are similar hence omitted for the simplicity of presentation.

that the analysis evaluates the main effects of visual conditions while accounting for any potential influence that the other factors may have on misbeliefs. Similarly, to test H2 and H3, which focus on interaction effects, we further incorporated interaction terms into the models.

Lastly, we examined the two mechanisms proposed in RQ1, acknowledging that this analysis was exploratory in nature and not pre-registered. We employed structural equation modeling (SEM) with 5,000 bootstrap samples to estimate indirect effects and their confidence intervals. The visual conditions were dummy coded into visual exemplar, data visualization, and no visual groups, and participants' misbelief was entered as the dependent variable. The mediators, psychological reactance and credibility, were modeled in parallel to examine their simultaneous indirect effects while allowing covariance between themselves. Following established literature on psychological reactance (Rains, 2013), we adopted the intertwined model, in which psychological reactance is modeled as a latent variable composed of two observed indicators: anger and counterarguing. We included covariates in the SEM analyses, since mediators are not randomized and can be influenced by confounding factors (Imai et al., 2010). Details on covariate selection and descriptive statistics are provided in the Supplemental Materials.

## Results

Regarding H1, we conducted linear regressions to compare the two treatment conditions (i.e., AI-generated visual exemplars or data visualization) with the textual correction only control. As detailed in Figure 3, the results show that exposure to AI-generated visual exemplars significantly reduced misbeliefs ( $b = -0.20$ ,  $p = .013$ , CI: [-0.37,

**Table 1.** Summary of Estimated Interaction Effects on Misbeliefs.

Independent Variables	Image × Source		Image × History		Source × History	
	Estimate	p	Estimate	p	Estimate	p
Data visualization (vs. text only)	<b>-0.30**</b> [-0.53, -0.08]	.009	-0.01 [-0.24, 0.21]	.900 -0.17		
Visual exemplar (vs. text only)	<b>-0.31**</b> [-0.54, -0.09]	.006				
AI with expert endorsement (vs. AI only)	0.13 [-0.10, 0.36]	.263				
Balanced history (vs. no history)			0.15 [-0.08, 0.38]	.189		
Data visualization × AI with expert endorsement	-0.28 [-0.61, 0.03]	.079				
Visual Exemplar × AI with expert endorsement	-0.22 [-0.55, 0.11]	.185				
Data Visualization × Balanced History			-0.29 [-0.61, 0.03]	.078		
Visual Exemplar × Balanced History			-0.08 [-0.40, 0.25]	.637		
AI with Expert Endorsement × Balanced history					0.01 [-0.25, 0.28]	.927

Note. The table presents unstandardized coefficients. Levels of significance: · p < .10; \* p < .05; \*\* p < .01; \*\*\* p < .001. Bolded coefficients indicate statistically significant effects at the level of p < .05.

$-0.04]$ ). The effect of exposure to AI-generated data visualization approached significance, though it was marginal ( $b=-0.16$ ,  $p=.051$ , CI:  $[-0.32, 0.00]$ ). These results suggest that AI-generated visual exemplars, but not the data visualization, had incremental persuasive benefits beyond textual correction alone. Thus, H1 was partially supported.

Both H2 and H3 concerned potential moderation effects from incorporating credibility boosters including partisan neutrality in posting history (H2) and source tagging (H3) interaction effects. As shown in Table 1, these pre-registered interaction effects largely did not reach conventional statistical significance. That said, we observed only marginally significant effect: incorporating source tagging with expert endorsement showed a trend toward enhancing the debunking effectiveness of AI-generated data visualization ( $b=-0.28$ ,  $p=.079$ , CI:  $[-0.61, 0.03]$ ), and the interaction between data visualization and balanced history yielded a marginally negative effect ( $b=-0.29$ ,  $p=.078$ , CI:  $[-0.61, 0.03]$ ).

To explore potential mediation effects through psychological reactance and/or credibility perceptions, we conducted SEM with maximum likelihood estimation with standard errors. The model statistics suggested an acceptable model fit ( $CFI=0.99$ ,  $RMSEA=0.03$ ,  $SRMR=0.01$ ). Following conventional thresholds (Hu & Bentler, 1999), values of  $CFI \geq 0.95$ ,  $RMSEA \leq 0.06$ , and  $SRMR \leq 0.08$  are considered indicators of good model fit. Path coefficients were summarized in Table 2 and visualized in Figure 4. The results suggest that visual exemplars significantly reduced psychological reactance ( $\beta=-.09$ ,  $p=.011$ , CI:  $[-0.16, -0.02]$ ). Moreover, psychological reactance significantly predicted misbeliefs ( $\beta=.59$ ,  $p<.001$ , CI:  $[0.51, 0.67]$ ). The indirect effect of visual exemplars on misbeliefs through psychological reactance was statistically significant ( $\beta=-.05$ ,  $p=.011$ , CI:  $[-0.09, -0.01]$ ), indicating that reductions in reactance partially explained the effectiveness of visual exemplars in lowering misbeliefs.

In contrast, the indirect effects through credibility were not significant for either visual exemplars ( $\beta=-.01$ ,  $p=.123$ , CI:  $[-0.02, 0.00]$ ) or data visualization ( $\beta=.00$ ,  $p=.432$ , CI:  $[-0.01, 0.01]$ ), suggesting that the perceived credibility of AI-generated visuals did not significantly contribute to misbelief correction. The findings answer RQ1, suggesting that the mechanism of psychological reactance played a more substantial role compared to credibility. Visual exemplars can reduce misbeliefs primarily by mitigating reactance.

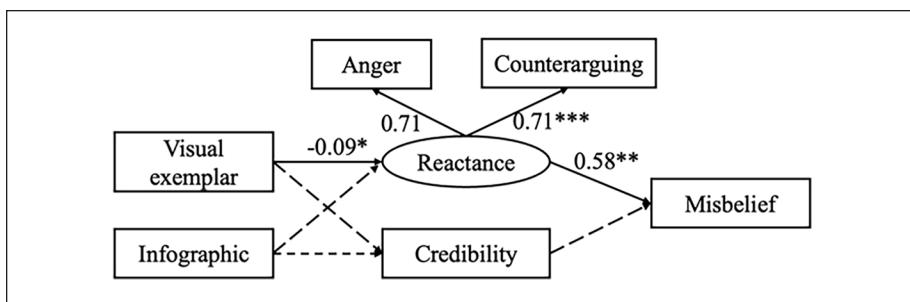
## Discussion

This study examined the role of AI-generated visuals and potential credibility boosters in enhancing the effectiveness of misinformation correction within the context of health communication. Our analysis indicated small but significant effects of AI-created visual exemplars compared to text-only corrections. The effects of visual exemplars were largely independent of the two types of credibility cues, showing the robustness of this type of AI-generated visuals. Furthermore, the exploratory mediation analysis found that AI-generated visual exemplars reduce misbeliefs primarily by

**Table 2.** Statistical Summary of Mediation Analysis.

Paths	$\beta$	Bootstrapped 95% CI
Total effects		
c1-path: Visual exemplar → Misbelief	-.07*	[-0.13, -0.02]
c2-path: Data visualization → Misbelief	-.05	[-0.10, 0.00]
Through reactance		
a1-path: Visual exemplar → Reactance	-.09*	[-0.16, -0.02]
a2-path: Data visualization → Reactance	.02	[-0.05, 0.09]
b1-path: Reactance → Misbelief	.59**	[0.51, 0.67]
a1b1-path: Visual exemplar → Reactance → Misbelief	-.05*	[-0.09, -0.01]
a2b1-path: Data visualization → Reactance → Misbelief	.01	[-0.03, 0.05]
c1'-path (direct): Visual exemplar → Misbelief	-.02	[-0.01, 0.04]
c2'-path (direct): Data visualization → Misbelief	-.06*	[-0.12, -0.00]
Through perceived credibility		
a3-path: Visual exemplar → Perceived credibility	.04	[-0.01, 0.09]
a4-path: Data visualization → Perceived credibility	-.02	[-0.07, 0.03]
b2-path: Perceived credibility → Misbelief	.18***	[-0.24, -0.12]
a3b2-path: Visual exemplar → Perceived credibility → Misbelief	-.01	[-0.02, 0.00]
a4b2-path: Data visualization → Perceived credibility → Misbelief	.00	[-0.01, 0.01]
c1'-path (direct): Visual exemplar → Misbelief	-.07*	[-0.12, -0.01]
c2'-path (direct): Data visualization → Misbelief	-.05*	[-0.11, -0.00]

Note. The coefficients are fully standardized. Covariates were added in the SEM analyses. The covariates were selected from the LASSO model based on their non-zero coefficients. More details about the covariates could be found in the Supplemental Materials. SEM = structural equation modeling. Levels of significance:  
 ·  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$



**Figure 4.** Path Diagram of Mediation Effects.

Note. The coefficients are fully standardized. The same set of covariates was used in the model. Anger was fixed for identification purposes; thus, no *p*-value is provided.

mitigating psychological reactance rather than boosting credibility perceptions, highlighting the relevance of PRT in understanding the effects of AI-generated multimodal corrections.

This study offers preliminary evidence that not all AI-generated visuals are equally effective—our results supported the efficacy of AI-generated visual exemplars but not the data visualization in enhancing the effects of text-only corrections. This finding underscores the need to take heterogeneity seriously in research on the roles of visuals in misinformation correction, regardless of whether they are produced by AI. The tendency to neglect distinctions between different types of visuals might account for the mixed findings in the current literature (Hameleers et al., 2020, 2023; Sundar et al., 2021; Young et al., 2018). We recommend that future studies expand both the types of visuals, and specific visual message instances per type to advance the current line of research in terms of both internal and external validity. Future research should also build on these findings by disentangling whether the observed effects are primarily driven by the content of AI-generated visuals or by audience perceptions of the AI generation process itself, such as manipulations on provided explanations of AI's working mechanisms (Xu & Shi, 2024). Because the current design does not include human-generated image controls, we cannot determine whether participants responded to the persuasive features of AI-generated content itself or to the knowledge (or assumptions) about the generation process.

Practically, given the challenge of efficiently producing high-quality visuals at scale, our results demonstrated that multimodal LLMs can expedite production, paving the way for wider adoption of effective visuals in corrective messages, particularly for individuals and organizations understaffed for media production. However, it is equally important to emphasize that generative AI cannot entirely replace human efforts, as there are many potential risks and ethical concerns related to AI-generated information (J. J. Kim et al., 2024). For infographics, it is important to ensure that the data source used for visualization is reliable and that there are no inaccurate portrayals of data in the visualization procedure. For visual exemplars, attention must be paid to potential biases

and legal implications, particularly to ensure the ethical depiction of individuals and the avoidance of harmful stereotypes or unintended misrepresentations. Educational programs focused on AI literacy are necessary (Long & Magerko, 2020), especially for designers, to promote responsible and informed visual message creation.

Despite the finding that AI-generated visual exemplars lowered misperceptions compared with textual correction alone, visual exemplars plus the base textual correction combined did not significantly differ from the misinformation-only condition (see Supplemental Materials for details). Notably, the misinformation-only condition did not significantly elevate misbeliefs compared to the questionnaire-only condition in this context. This unexpected null result might be due to a weak, though undetectable, boomerang effect associated with text-only corrections. As suggested by DeVerna et al. (2024), textual AI-generated fact checks may fail to improve correction and readers' discernment. Given the observed indirect pathway through psychological reactance when we compared AI-generated visual exemplars and text-only corrections, we speculate that the non-significant effect of text-only correction as compared to the misinformation-only condition might be attributable to induced reactance. In other words, our findings may also be interpreted as visual exemplars mitigating the potential backfire effects of textual correction alone. Further research is encouraged to empirically test this possibility. Therefore, we interpret this null finding as an important qualification of the main effects observed when comparing AI-generated visual exemplars to text-only conditions.

To better understand the theoretical mechanisms underlying the differential effects of AI-generated visual exemplars versus the data visualization, we conducted exploratory analyses testing two potential indirect pathways—credibility perceptions and psychological reactance. Reactance, as a latent construct constituted by anger and counterarguing, is often invoked to explain the (in)effectiveness of health persuasive messages (Jebai et al., 2023; LaVoie et al., 2017; Miller et al., 2006; Ringold, 2002). Consistent with a recent meta-analysis that showed narrative health messages can reduce different forms of resistance including psychological reactance (Ratcliff & Sun, 2020), we extended this literature to misinformation correction, where the theoretical framework of PRT is less commonly applied. Since misinformation correction represents an explicit attempt at changing one's preexisting (mis)beliefs, psychological reactance is arguably a key mediating mechanism. Our results suggest that employing AI-generated visual exemplars, even as a subtle and less sophisticated form of narrative, is sufficient to mitigate reactance. This is probably because a visual exemplar might have made the persuasive attempt more unobtrusive. Future research should empirically verify this assumption by directly measuring perceived freedom threat, which is often theorized as a precedent to the experience of reactance (Quick et al., 2013). As AI continues to evolve, it will likely generate more complex narrative formats, such as video, with models such as Sora from OpenAI already demonstrating the ability to produce videos from text. As these capabilities expand, understanding and addressing reactance will become even more crucial for successful corrective messaging.

It is worth noting that credibility seems to play a less important role in accounting for the correction efficacy of different types of AI-generated visuals. We found no

evidence for indirect effects through credibility perceptions, nor did we observe any moderation effects from either partisan neutrality in posting history or expert endorsement as credibility cues. It is useful to clarify that our interest lies primarily in contrasting different types of AI-generated visual correction, whereas in prior literature researchers often compare AI versus human-generated messages (Moon et al., 2022). Perhaps credibility perceptions play a more prominent role when machine heuristic is invoked in the AI versus human contrast. Another possibility is that our operationalization of credibility cues did not suffice in enhancing perceived credibility. Regarding partisan neutrality in posting history, any inclusion of political cues may invite accusations of bias at a time characterized by severe political polarization in the United States. When it comes to expert endorsement, similar to current social media practices, we included a source tagline at the bottom of the correction post, though this presentation of source may not have been visually prominent enough to capture attention. Therefore, our null findings should be interpreted with regard to our operationalization of credibility cues as objective message features, per O'Keefe's conceptual framework (O'Keefe, 2003). Future research is encouraged to explore alternative operationalizations before reaching a conclusion about credibility cues as a category in the context of AI-generated visual correction.

Lastly, our results also suggest that in the research on AI-generated visuals, alternative theoretical frameworks, such as the recent theorization of the roles of authenticity in computer-mediated communication (E. J. Lee, 2020), deserve more scholarly attention. As generative AI technologies are increasingly adopted to usher in more human-AI hybrid communication, concerns about whether the presented source identity matches with its real identity (i.e., source authenticity), how individuals conceptualize the degree of "realism" of AI-generated corrections (i.e., message authenticity), and how humans judge the veracity of interactive exchanges with the AI fact-checker (i.e., interaction authenticity), may loom large.

Regarding our finding that the AI-generated data visualization only marginally enhanced correction in the unconditional main effects models, another possible explanation lies in the cognitive demands such visualizations impose on the audience. Data visualizations can be seen as credible, but they require active cognitive engagement to interpret. As Braverman (2008) notes, informational messages work best for individuals who are highly involved and enjoy thinking critically. Additionally, the veneer of scientific legitimacy of visualizing data does not ensure veracity, as both truthful and misleading social media posts are equally likely to incorporate data visualizations (Lo et al., 2022). Misleading or poorly designed visualizations in misinformation posts can erode trust in this format, leading to a "spillover effect" where exposure to deceptive visuals undermines the credibility of all data visualizations, regardless of their accuracy or source. Future studies should further examine which segments of the population might be more receptive to AI-generated data visualizations (e.g., those higher on need for cognition). Overall, this largely null finding about the AI-generated data visualization serves as a reminder that technological capacities do not necessarily translate into impact. As AI grows increasingly more powerful and multifunctional, systematic empirical research is needed to identify which AI-afforded functionalities are worth incorporating into practices of misinformation correction.

The study is not free from limitations. Due to challenges in quota implementation, our sample includes more male (59.11%) than female children. This modest sample imbalance is unlikely to bias our key conclusions, as we have tested both unconditional models without any covariates and conditional models with covariates (see Supplemental Materials) to demonstrate robustness and consistency. Furthermore, although we conceptualize the two types of source credibility cues as objective message features, and hence saw no need for manipulation checks (O'Keefe, 2003), we recognize the concern that our specific operationalization might not suffice in improving perceived credibility. To enhance external validity, we designed the messages to resemble realistic social media feeds, particularly in terms of source tagging. Indeed, participants failed to accurately recognize both partisan neutrality in posting history and AI labeling cues. Most participants in both the balanced history (33.92%) and no-history conditions (33.33%) perceived the source as Democrat-leaning rather than neutral (32.16% and 32.40%, respectively). Similarly, participants in both AI conditions predominantly assumed expert validation was involved (48.22% and 50.66%), regardless of whether this was actually indicated. These findings limit the generalizability of our conclusions to the specific operationalizations we employed, which we encourage future research to expand. We adopted the operationalization of source tagging to align with current practices in industry, where social media platforms have added tags to mark content made by AI to "enhance transparency and literacy" (TikTok, 2023). However, our results indicate that these labels may not effectively achieve their intended purpose, which highlights the urgent need to move beyond simply labeling content as "made with AI." and consider alternative practices (e.g., increasing the visibility of labels, focusing on AI literacy training instead).

Despite these limitations, this study is among the first to examine the role of AI-generated visuals in combating health misinformation, offering insights into the psychological mechanisms that shape their effectiveness. Our findings suggest that the AI-generated visual exemplars serve as a modest yet promising enhancement to textual correction, particularly with regards to reducing psychological reactance, an important mediating pathway to the success of misinformation correction efforts. While the misinformation-only condition did not produce strong disbelief reinforcement, visual exemplars appeared to move audiences further toward more accurate perceptions. As AI continues to shape the information landscape, understanding how to optimize AI-driven interventions remains critical for developing effective, scalable strategies to counter misinformation.

### **Acknowledgements**

We are grateful to Nathan Kalmoe, Emily K. Vraga, and Letitia Bode, along with research assistants Xiaohui Cao, Rongwei Tang, and Yuwei Ma, for their generous assistance and contributions to this work.

### **Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the National Science Foundation through the Convergence Accelerator Track F (Agency Tracking Number: 2230692; Award Number: MSN 266268) and the John S. and James L. Knight Foundation (Award Number: MSN231314). This research was also supported by the University of Wisconsin—Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (Award Number: MSN275781).

## ORCID IDs

- Yibing Sun  <https://orcid.org/0000-0003-1754-9895>
- Liwei Shen  <https://orcid.org/0000-0003-0160-2882>
- Ji Soo Choi  <https://orcid.org/0009-0009-8480-990X>
- Porislmita Borah  <https://orcid.org/0000-0002-1140-4233>
- Michael W. Wagner  <https://orcid.org/0000-0003-4590-5033>
- Dhavan V. Shah  <https://orcid.org/0000-0001-5034-2816>
- Sijia Yang  <https://orcid.org/0000-0003-4209-9881>

## Supplemental Material

Supplemental material for this article is available online.

## References

- Barari, S., Lucas, C., & Munger, K. (2025). Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics*, 87(2), 510–526. <https://doi.org/10.1086/732990>
- Barman, D., Guo, Z., & Conlan, O. (2024). The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 16, Article 100545. <https://doi.org/10.1016/j.mlwa.2024.100545>
- Bigsby, E., Bigman, C. A., & Gonzalez, A. M. (2019). Exemplification theory: A review and meta-analysis of exemplar messages. *Annals of the International Communication Association*, 43(4), 273–296. <https://doi.org/10.1080/23808985.2019.1681903>
- Braverman, J. (2008). Testimonials versus informational persuasive messages: The moderating effect of delivery mode and personal involvement. *Communication Research*, 35(5), 666–694. <https://doi.org/10.1177/0093650208321785>
- Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance: A theory of freedom and control*. Academic Press.
- Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (mis)representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics*, 26(1), 277–299. <https://doi.org/10.1177/1940161220964780>
- Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., . . . Viale, R. (2024). The impact of generative artificial intelligence on socio-economic inequalities and policy making. *PNAS Nexus*, 3(6), Article pgae191. <https://doi.org/10.1093/pnasnexus/pgae191>

- Chen, C., & Shu, K. (2023). *Can LLM-generated misinformation be detected?* arXiv preprint arXiv:2309.13788.
- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1), 1–12. <https://doi.org/10.1057/s41599-023-02079-x>
- Chung, M., Moon, W. K., & Jones-Jang, S. M. (2023). AI as an apolitical referee: Using alternative sources to decrease partisan biases in the processing of fact-checking messages. *Digital Journalism*, 12(10), 1548–1569. <https://doi.org/10.1080/21670811.2023.2254820>
- Comello, M. L. G., Qian, X., Deal, A. M., Ribisl, K. M., Linnan, L. A., & Tate, D. F. (2016). Impact of game-inspired infographics on user engagement and information processing in an eHealth program. *Journal of Medical Internet Research*, 18(9), e237. <https://doi.org/10.2196/jmir.5976>
- Dal Cin, S., Zanna, M. P., & Fong, G. T. (2004). Narrative persuasion and overcoming resistance. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and persuasion* (pp. 175–192). Psychology Press.
- Dan, V., & Coleman, R. (2024). “I’ll change my beliefs when I see it”: Video fact checks outperform text fact checks in correcting misperceptions among those holding false or uncertain pre-existing beliefs. *Communication Research*, 52(6), 778–802. <https://doi.org/10.1177/00936502241287870>
- DeVerna, M. R., Yan, H. Y., Yang, K. C., & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50), e2322823121. <https://doi.org/10.1073/pnas.2322823121>
- Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72(2), 144–168. <https://doi.org/10.1080/03637750500111815>
- Dixon, G. N., McKeever, B. W., Holton, A. E., Clarke, C., & Eosco, G. (2015). The power of a picture: Overcoming scientific misinformation by communicating weight-of-evidence information with visual exemplars. *Journal of Communication*, 65(4), 639–659. <https://doi.org/10.1111/jcom.12159>
- Domgaard, S., & Park, M. (2021). Combating misinformation: The effects of infographics in verifying false vaccine news. *Health Education Journal*, 80(8), 974–986. <https://doi.org/10.1177/00178969211038750>
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), Article 1. <https://doi.org/10.1038/s44159-021-00006-y>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G\* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Featherstone, J. D., & Zhang, J. (2020). Feeling angry: The effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9), 692–702. <https://doi.org/10.1080/10810730.2020.1838671>
- Flanagan, A. J., Winter, S., & Metzger, M. J. (2020). Making sense of credibility in complex information environments: The role of message sidedness, information source, and thinking styles in credibility evaluation online. *Information, Communication & Society*, 23(7), 1038–1056. <https://doi.org/10.1080/1369118X.2018.1547411>
- Gardner, L., & Leshner, G. (2016). The role of narrative and other-referencing in attenuating psychological reactance to diabetes self-care messages. *Health Communication*, 31(6), 738–751. <https://doi.org/10.1080/10410236.2014.993498>

- Gawronski, B., Ng, N. L., & Luke, D. M. (2023). Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General*, 152(8), 2205–2236. <https://doi.org/10.1037/xge0001381>
- Ha, S., & Ahn, J. (2011). *Why are you sharing others' tweets?: The impact of argument quality and source credibility on information sharing behavior*. ICIS 2011 Proceedings, Shanghai China, 4. <https://aisel.aisnet.org/icis2011/proceedings/humanbehavior/4>
- Hameleers, M., Harff, D., & Schmuck, D. (2023). The alternative truth kept hidden from us: The effects of multimodal disinformation disseminated by ordinary citizens and alternative hyper-partisan media: Evidence from the US and India. *Digital Journalism*, 1–22. <https://doi.org/10.1080/21670811.2023.2210616>
- Hameleers, M., Powell, T. E., Van Der Meer, T. G., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2), 281–301. <https://doi.org/10.1080/10584609.2019.1674979>
- Heley, K., Gaysinsky, 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>
- Hemsley, J., & Snyder, J. (2018). FIVE dimensions of visual misinformation in the emerging media landscape. In B. G. Southwell, E. A. Thorson, & L. Sheble (Eds.), *Misinformation and mass audiences* (pp. 91–106). University of Texas Press.
- Henke, J., Leissner, L., & Möhring, W. (2020). How can journalists promote news credibility? Effects of evidences on trust and credibility. *Journalism Practice*, 14(3), 299–318. <https://doi.org/10.1080/17512786.2019.1605839>
- Hoffman, J. (2019). One more time, with big data: Measles vaccine doesn't cause autism. *The New York Times*, 7.
- Horne, B. D., Nevo, D., O'Donovan, J., Cho, J. H., & Adali, S. (2019, July). Rating reliability and bias in news articles: Does AI assistance help everyone? *Proceedings of the international AAAI conference on web and social media*, 13(1), 247–256. <https://doi.org/10.1609/icwsm.v13i01.3226>
- Hornik, R., & Woolf, K. D. (1999). Using cross-sectional surveys to plan message strategies. *Social Marketing Quarterly*, 5(2), 34–41. <https://doi.org/10.1080/15245004.1999.9961044>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huang, Y., & Sundar, S. S. (2022). Do we trust the crowd? Effects of crowdsourcing on perceived credibility of online health information. *Health Communication*, 37(1), 93–102. <https://doi.org/10.1080/10410236.2020.1824662>
- Huang, Y., & Wang, W. (2022). When a story contradicts: Correcting health misinformation on social media through different message formats and mechanisms. *Information, Communication & Society*, 25, 1192–1209. <https://doi.org/10.1080/1369118X.2020.1851390>
- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1), 51–71. <https://doi.org/10.1214/10-STS321>
- Jebai, R., Asfar, T., Nakkash, R., Chehab, S., Wu, W., Bursac, Z., & Maziak, W. (2023). Impact of pictorial health warning labels on smoking beliefs and perceptions among waterpipe smokers: An online randomised cross-over experimental study. *Tobacco Control*, 32(6), 715–722. <https://doi.org/10.1136/tobaccocontrol-2021-057202>

- Jia, C., & Liu, R. (2021). Algorithmic or human source? Examining relative hostile media effect with a transformer-based framework. *Media and Communication*, 9(4), 170–181. <https://doi.org/10.17645/mac.v9i4.4164>
- Joo, J., Li, W., Steen, F. F., & Zhu, S.-C. (2014). Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)* (pp. 216–223). [https://openaccess.thecvf.com/content\\_cvpr\\_2014/html/Joo\\_Visual\\_Persuasion\\_Inferring\\_2014\\_CVPR\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2014/html/Joo_Visual_Persuasion_Inferring_2014_CVPR_paper.html)
- Khan, S. A., Sheikhi, G., Opdahl, A. L., Rabbi, F., Stoppel, S., Trattner, C., & Dang-Nguyen, D. T. (2023). Visual user-generated content verification in journalism: An overview. *IEEE Access*, 11, 6748–6769.
- Kim, H. S., Bigman, C. A., Leader, A. E., Lerman, C., & Cappella, J. N. (2012). Narrative health communication and behavior change: The influence of exemplars in the news on intention to quit smoking. *Journal of Communication*, 62(3), 473–492. <https://doi.org/10.1111/j.1460-2466.2012.01644.x>
- Kim, J. J., Um, R. S., Lee, J. W., & Ajilore, O. (2024). Generative AI can fabricate advanced scientific visualizations: Ethical implications and strategic mitigation framework. *AI and Ethics*, 1–13. <https://doi.org/10.1007/s43681-024-00439-0>
- Kim, S. C., Vraga, E. K., & Cook, J. (2021). An eye tracking approach to understanding misinformation. *Health Communication*, 36(13), 1687–1696. <https://doi.org/10.1080/10410236.2020.1787933>
- King, A. J., & Lazard, A. J. (2020). Advancing visual health communication research to improve infodemic response. *Health Communication*, 35(14), 1723–1728. <https://doi.org/10.1080/10410236.2020.1838094>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- LaVoie, N. R., Quick, B. L., Riles, J. M., & Lambert, N. J. (2017). Are graphic cigarette warning labels an effective message strategy? A test of psychological reactance theory and source appraisal. *Communication Research*, 44(3), 416–436. <https://doi.org/10.1177/0093650215609669>
- Lee, E. J. (2020). Authenticity model of (mass-oriented) computer-mediated communication: Conceptual explorations and testable propositions. *Journal of Computer-Mediated Communication*, 25(1), 60–73. <https://doi.org/10.1093/jcmc/zmz025>
- Lee, E. J., & Kim, Y. W. (2016). Effects of infographics on news elaboration, acquisition, and evaluation: Prior knowledge and issue involvement as moderators. *New Media & Society*, 18(8), 1579–1598. <https://doi.org/10.1177/1461444814567982>
- Lee, J., & Hameleers, M. (2024). Effects of health-related deepfakes on misperceptions: Moderating effects of issue relevance and accuracy motivation. *Media Psychology*, 1–30. <https://doi.org/10.1080/15213269.2024.2401539>
- Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of applied research in memory and cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Li, J. (2025). Not all skepticism is “healthy” skepticism: Theorizing accuracy-and identity-motivated skepticism toward social media misinformation. *New Media & Society*, 27(1), 522–544. <https://doi.org/10.1177/14614448231179941>
- Li, J., & Wagner, M. W. (2020). The value of not knowing: Partisan cue-taking and belief updating of the uninformed, the ambiguous, and the misinformed. *Journal of Communication*, 70(5), 646–669. <https://doi.org/10.1093/joc/jqaa022>
- Liu, A., Sheng, Q., & Hu, X. (2024, July). *Preventing and detecting misinformation generated by large language models*. Proceedings of the 47th International ACM SIGIR Conference

- on Research and Development in Information Retrieval (pp. 3001–3004). ACM. <https://doi.org/10.1145/3626772.3661377>
- Lo, L. Y. H., Gupta, A., Shigyo, K., Wu, A., Bertini, E., & Qu, H. (2022, June). Misinformed by visualization: What do we learn from misinformative visualizations?. *Computer Graphics Forum*, 41(3), 515–525.
- Long, D., & Magerko, B. (2020, April). *What is AI literacy? Competencies and design considerations*. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). ACM. <https://doi.org/10.1145/3313831.3376727>
- Lu, Y., & Shen, C. (2023). Unpacking multimodal fact-checking: Features and engagement of fact-checking videos on Chinese TikTok (Douyin). *Social Media + Society*, 9(1). <https://doi.org/10.1177/20563051221150406>
- Mena, P. (2023). Reducing misperceptions through news stories with data visualization: The role of readers' prior knowledge and prior beliefs. *Journalism*, 24(4), 729–748. <https://doi.org/10.1177/14648849211028762>
- Metzger, M. J., Flanagan, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association*, 27(1), 293–335. <https://doi.org/10.1080/23808985.2003.11679029>
- Miller, C. H., Burgoon, M., Grandpre, J., & Alvaro, E. (2006). Identifying principal risk factors for the initiation of adolescent smoking behaviors: The significance of psychological reactance. *Health Communication*, 19, 241–252. [https://doi.org/10.1207/s15327027hc1903\\_6](https://doi.org/10.1207/s15327027hc1903_6)
- Miller, C. H., Lane, L. T., Deatrick, L. M., Young, A. M., & Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, 33(2), 219–240. <https://doi.org/10.1111/j.1468-2958.2007.00297.x>
- Moon, W.-K., Chung, M., & Jones-Jang, S. Mo. (2022). How can we fight partisan biases in the COVID-19 pandemic? AI source labels on fact-checking messages reduce motivated reasoning. *Mass Communication and Society*, 26(4), 646–670. <https://doi.org/10.1080/15205436.2022.2097926>
- Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1), 3–23. <https://doi.org/10.1007/s11127-023-01097-2>
- Neumann, M., Moore, S. T., Baum, L. M., Oleinikov, P., Xu, Y., Niederdeppe, J., Gollust, S. E., & Fowler, E. F. (2024). Politicizing masks? Examining the volume and content of local news coverage of face coverings in the US through the COVID-19 pandemic. *Political Communication*, 41(1), 66–106. <https://doi.org/10.1080/10584609.2023.2239181>
- Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 29(2), 222–244. <https://doi.org/10.1080/17457289.2018.1465061>
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Occa, A., & Suggs, L. S. (2015). Communicating breast cancer screening with young women: An experimental test of didactic and narrative messages using video and infographics. *Journal of Health Communication*, 21(1), 1–11. <https://doi.org/10.1080/10810730.2015.1018611>
- O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory*, 13(3), 251–274. <https://doi.org/10.1111/j.1468-2885.2003.tb00292.x>

- Ophir, Y., Brennan, E., Maloney, E. K., & Cappella, J. N. (2019). The effects of graphic warning labels' vividness on message engagement and intentions to quit smoking. *Communication Research*, 46(5), 619–638. <https://doi.org/10.1177/0093650217700226>
- Peng, Q., Lu, Y., Peng, Y., Qian, S., Liu, X., & Shen, C. (2024). *Crafting synthetic realities: Examining visual realism and misinformation potential of photorealistic AI-generated images*. arXiv preprint arXiv:2409.17484.
- Peng, Y. (2022). Athec: A Python Library for computational aesthetic analysis of visual media in social science research. *Computational Communication Research*, 4(1), 323–349. <https://doi.org/10.5117/CCR2022.1.009.PENG>
- Peng, Y., Lu, Y., & Shen, C. (2023). An Agenda for studying credibility perceptions of visual misinformation. *Political Communication*, 40(2), 225–237. <https://doi.org/10.1080/10584609.2023.2175398>
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowd-sourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526. <https://doi.org/10.1073/pnas.1806781116>
- Porter, E., & Russell, K. (2018). Migrants are on the rise around the world, and myths about them are shaping attitudes. *New York Times*, 20.
- Quadflieg, S., Neuburg, K., & Nestler, S. (2022). *(Dis)Obedience in digital societies: Perspectives on the power of algorithms and data*. Transcript Verlag. <https://directory.doa-books.org/handle/20.500.12854/80788>
- Quick, B. L., Shen, L., & Dillard, J. P. (2013). Reactance theory and persuasion. In J. P. Dillard & L. Shen (Eds.), *The SAGE handbook of persuasion: Developments in theory and practice* (2nd ed., pp. 167–183). SAGE Publications.
- Quick, B. L., & Stephenson, M. T. (2008). Examining the role of trait reactance and sensation seeking on perceived threat, state reactance, and reactance restoration. *Human Communication Research*, 34(3), 448–476. <https://doi.org/10.1111/j.1468-2958.2008.00328.x>
- Rains, S. A. (2013). The nature of psychological reactance revisited: A meta-analytic review. *Human Communication Research*, 39(1), 47–73. <https://doi.org/10.1111/j.1468-2958.2012.01443.x>
- Rains, S. A., & Turner, M. M. (2007). Psychological reactance and persuasive health communication: A test and extension of the intertwined model. *Human Communication Research*, 33(2), 241–269. <https://doi.org/10.1111/j.1468-2958.2007.00298.x>
- Ratcliff, C. L., & Sun, Y. (2020). Overcoming resistance through narratives: Findings from a meta-analytic review. *Human Communication Research*, 46(4), 412–443. <https://doi.org/10.1093/hcr/hqz017>
- Ratcliff, C. L., King, A. J., Wicke, R., Pokharel, M., Adams, D. R., & Ratcliff, C. L., King, A. J., Wicke, R., Pokharel, M., Adams, D. R., & Jensen, J. D. (2025). Examining Reactance to Visual and Verbal Features of Mask Promotion PSAs. *Journal of Health Communication*, 30(sup 1), 28–38. <https://doi.org/10.1080/10810730.2024.2437039>
- Reeves, B., Yeykelis, L., & Cummings, J. J. (2016). The use of media in media psychology. *Media Psychology*, 19(1), 49–71. <https://doi.org/10.1080/15213269.2015.1030083>
- Ringold, D. J. (2002). Boomerang effects in response to public health interventions: Some unintended consequences in the alcoholic beverage market. *Journal of Consumer Policy*, 25(1), 27–63. <https://doi.org/10.1023/A:1014588126336>
- Rosenblum, H. G., Lewis, R. M., Gargano, J. W., Querec, T. D., Unger, E. R., & Markowitz, L. E. (2022). Human papillomavirus vaccine impact and effectiveness through 12 years after vaccine introduction in the United States, 2003 to 2018. *Annals of Internal Medicine*, 175(7), 918–926. <https://doi.org/10.7326/M21-3798>

- Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024(1), Article 7115633. <https://doi.org/10.1155/2024/7115633>
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 4787. <https://doi.org/10.1038/s41467-018-06930-7>
- Shen, C., Kasra, M., Wenjing, P., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society*, 21(2), 438–463. <https://doi.org/10.1177/1461444818799526>
- Shen, L. (2010). Mitigating psychological reactance: The role of message-induced empathy in persuasion. *Human Communication Research*, 36(3), 397–422. <https://doi.org/10.1111/j.1468-2958.2010.01381.x>
- Silvia, P. J. (2006). Reactance and the dynamics of disagreement: Multiple paths from threatened freedom to resistance to persuasion. *European Journal of Social Psychology*, 36, 673–685. doi:10.1002/ejsp.v36:5v
- Slater, D. M., Peter, J., & Valkenburg, P. M. (2015). Message variability and heterogeneity: A core challenge for communication research. *Annals of the International Communication Association*, 39(1), 3–31. <https://doi.org/10.1080/23808985.2015.11679170>
- Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2024). Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication*, 29(1), Article zmad045. <https://doi.org/10.1093/jcmc/zmad045>
- Sun, Y., & Lu, F. (2022). How misinformation and rebuttals in online comments affect people's intention to receive COVID-19 vaccines: The roles of psychological reactance and misperceptions. *Journalism & Mass Communication Quarterly*, 100(1), 145–171. <https://doi.org/10.1177/10776990221084606>
- Sun, Y., Pendyala, V., Lian, R., Xin, H., Patel, P., Bucy, E. P., & Shah, D. V. (2025). From Internet meme to the mainstream: Using computer vision to track "Pepe the Frog" across news platforms. *Visual Communication Quarterly*, 32(1), 33–55. <https://doi.org/10.1080/15551393.2025.2455495>
- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Sundar, S. S., & Kim, J. (2019, May). *Machine heuristic: When we trust computers more than humans with our personal information*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–9). ACM. <https://doi.org/10.1145/3290605.3300768>
- Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6), 301–319. <https://doi.org/10.1093/jcmc/zmab010>
- Taumberger, N., Joura, E. A., Arbyn, M., Kyrgiou, M., Sehouli, J., & Gultekin, M. (2022). Myths and fake messages about human papillomavirus (HPV) vaccination: Answers from the ESGO Prevention Committee. *International Journal of Gynecological Cancer*, 32(10), 1316–1320. <https://doi.org/10.1136/ijgc-2022-003685>

- TikTok. (2023, September 19). New labels for disclosing AI-generated content. *TikTok*. <https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content>
- Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5), 621–645. <https://doi.org/10.1177/1075547017731776>
- Wahbeh, A., Al-Ramahi, M., El-Gayar, O., Elnoshokaty, A., & Nasralah, T. (2023, December). *Perception of bias in ChatGPT: Analysis of social media data*. 2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) (pp. 34–39). IEEE. <https://doi.org/10.1109/GCAIoT61060.2023.10385099>
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), 350–375. <https://doi.org/10.1080/10584609.2019.1668894>
- Xu, K., & Shi, J. (2024). Visioning a two-level human-machine communication framework: initiating conversations between explainable AI and communication. *Communication Theory*, 34(4), 216–229. <https://doi.org/10.1093/ct/qtae016>
- Yang, Y., Davis, T., & Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*, 73(4), 316–328. <https://doi.org/10.1093/joc/jqac051>
- Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck.org and FlackCheck.org. *Journalism & Mass Communication Quarterly*, 95(1), 49–75. <https://doi.org/10.1177/1077699017710453>
- Zhang, J., Featherstone, J. D., Calabrese, C., & Wojcieszak, M. (2021). Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine*, 145, Article 106408. <https://doi.org/10.1016/j.ypmed.2020.106408>
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's song in the AI ocean: A survey on hallucination in large language models*. arXiv preprint arXiv:2309.01219.
- Zhou, A., Liu, W., & Yang, A. (2024). Politicization of science in COVID-19 vaccine communication: Comparing US politicians, medical experts, and government agencies. *Political Communication*, 41(4), 649–671. <https://doi.org/10.1080/10584609.2023.2201184>
- Zillmann, D. (1999). Exemplification theory: Judging the whole by some of its parts. *Media Psychology*, 1(1), 69–94. [https://doi.org/10.1207/s1532785xmep0101\\_5](https://doi.org/10.1207/s1532785xmep0101_5)
- Zillmann, D., & Brosius, H.-B. (2000). *Exemplification in communication: The influence of case reports on the perception of issues*. Lawrence Erlbaum Associates Publishers.
- Zillmann, D., & Brosius, H. B. (2012). *Exemplification in communication: The influence of case reports on the perception of issues*. Routledge.

## Author Biographies

**Yibing Sun** is a Ph.D. candidate in the School of Journalism and Mass Communication at the University of Wisconsin-Madison. Her research examines the production and effects of multimodal content in networked communication environments, with a focus on visual communication, computational methods, and emerging technologies such as artificial intelligence in shaping public understanding.

**Liwei Shen** is a Ph.D. candidate in Communication Science at the Department of Communication Arts at the University of Wisconsin-Madison. Her research interests include computational communication, misinformation, and group identity.

**Ji Soo Choi** is a Ph.D. student in the School of Journalism and Mass Communication at the University of Wisconsin–Madison.

**Porismita Borah** is Professor in the Edward R. Murrow College of Communication at Washington State University and a graduate faculty member in the Prevention Science program. Her research focuses on the intersection of emerging technology, politics, and health communication.

**Michael W. Wagner** is the William T. Evjue Distinguished Chair for the Wisconsin Idea and Professor of Journalism and Mass Communication at the University of Wisconsin–Madison, where he directs the Center for Communication and Civic Renewal and serves as Director of Graduate Studies. His research examines how people's experiences in the information environment shape their political beliefs, policy preferences, and civic and political participation.

**Dhavan V. Shah** is the Jack M. McLeod Professor of Communication Research and the Louis A. & Mary E. Maier-Bascom Chair at the University of Wisconsin–Madison. He directs the Mass Communication Research Center and serves as Research Director of the Center for Communication and Civic Renewal. His interdisciplinary research focuses on how mass media and interpersonal communication shape social judgments, civic engagement, and health-related support and behavior.

**Sijia Yang** is Associate Professor in the School of Journalism and Mass Communication at the University of Wisconsin–Madison. His research focuses on computational communication science, message effects and persuasion in digital media, especially in the areas of health and science communication.