Title: Additional methodological details appendix for “People are more susceptible to misinformation with realistic AI-synthesized images that provide strong evidence to headlines”
Authors: Sean Guo (1), Yiwen Zhong (2), Xiaoqing Hu (1) (3)
Date: November 10th, 2025
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix E: Additional methodological details

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

Pretesting

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

Procedure

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

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

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

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

Delayed post-correction memory (phase 5): We invited participants to complete the second part of the experiment approximately seven days after completion of phase 4 (*M* interval between sessions = 165 hours, 58 minutes [6.9 days]). After receiving the invitation, they had 24 hours to complete the experiment. We told participants: "The following section will be a memory test. For each headline, please try to remember whether it was true or false, as indicated in the corrections and confirmations presented one week ago. Please adjust your rating according to your confidence. For example, if you are very confident that the headline was true, rate "10." If you are slightly less confident, rate "9," and so on. If you have no recollection at all of whether it was true or false, rate "5." Please try your best to remember whether each headline was true or false". Participants then viewed all 40 headlines again individually in a randomized order and completed memory ratings using a discrete scale from 0 (*definitely false*) to 10 (*definitely true*). These memory ratings were subsequently reverse-coded such that higher values indicated greater memory for corrections.