# Seeing Is No Longer Believing: Teaching Visual Literacy in the Age of AI

**Kali Plkl** , Western Oregon University
Faculty Sponsor: **Dr. Ethan McMahan**

Media literacy and critical analysis of multiple formats of information are both critical skills with the rate at which Artificial Intelligence models are advancing. The current study sampled 87 university students and sought to evaluate participants' abil ity to detect AI -altered/generated images. This was measured via survey and was followed by a training module designed to teach participants about some of the key artifacts in AI -generated/altered images. Participants completed a second survey following th e training portion to evaluate for change. It was hypothesized that there would be a significant difference between the participant's pre -treatment and post -treatment scores following the training module. A paired-samples t-test showed that post-treatment survey scores ($M = 11.66, SD = 2.25$) were significantly higher than pre-treatment scores ($M = 9.05, SD = 2.61$), $t(86) = -10.19, p < .001, 95\%\ CI\ [-3.12, -2.10]$. The training produced a substantial improvement in scores, after controlling for relevant individual differences, with a large effect size ($d = 1.09$). These results show a strong ability to teach audiences how to recognize images that have been generated or altered by AI. The applicability to the generalized population of the specific tips provided in this study is questionable as image-generating AI models will likely phase out many of the included AI-artifacts.

*Keywords*: Artificial intelligence, media literacy, misinformation

## Introduction

With the invention of photography and videography, the phrase "Seeing is believing" used to have some validity to it. But in today's age, can we trust that images or videos we see are things we should believe? Exposure to artificially generated images or c ontent becomes a regular occurrence with the increasing use of artificial intelligence (AI) and large language models (LLMs). In 2017, the transformer architecture of large language models was introduced (Vaswani et al., 2017) and became the backbone of models such as ChatGPT.

Companies are now using AI-generated images to advertise their products, social media companies are incorporating AI into many of their apps, and schools are having to combat a new form of plagiarism as students attempt to pass off AI -written work as their own. As AI becomes increasingly present and more advanced while being cheap for the consumer, it is important to evaluate sources of content for AI generated material and to avoid falling for

 ©2026

misinformation, scams, or other exploitative content. The current research looks the participant's baseline ability to discern between real and AI-altered/generated images, and whether or not participants can be trained to identify some of the common artifacts in AI-generated images.

*Fake news* is described as the production, consumption, and distribution of false information through digital channels, often disguised as credible content (Giordano et al., 2025). While the origin of false or misleading stories predates our historical records, humans could not instantly relay images or video to mass audiences until the invention of the television, which became a common household item in the late 1940's. When the internet started to make its way into American homes in the mid-to-late 1990's, it gave anyone with internet access the ability to quickly disseminate and consume news. Now that we have access to LLMs and AI programs, disseminating news and information to large audiences happens at an unprecedented speed and scale. These generative systems can create entirely fabricated stories, images, and even videos without building in a way for people to easily discern between what's real and what's not. This can lead to a rapid dissemination of misinformation online by bad-faith actors, making it especially difficult to determine what's truth and what's fiction.

Otgaar et al. (2022) described false memories as recollections of events that never occurred or are distorted versions of actual events, often formed through suggestive influences or misinformation. The two main processes that factor into false memories are belief and recollection, though they are not mutually exclusive. False memories stemming from fake news are suggestion-induced, as they are being provided to influence the audience to believe what the authors of the fake news are claiming. It is important to also factor repetition into the increase in belief of false memories, as research shows that the more frequently someone has been shown a false story, the more likely they are to believe it is true (Grinfeld et al., 2025). A person may be exposed to false or misleading information and initially dismiss it, but if it is widely shared via social media and they see it repeatedly, they will be more likely to believe the information is true.

Image generating artificial intelligence models include generative adversarial networks (GAN), variational autoencoders (VAE), or diffusion models. A literature review conducted by Chen et al. (2024) described VAEs as generative models that are trained on images by reducing them down to unique characteristics and then building a new image based off the combination of these unique image characteristics. GANs are image generators that use two models that compete between the image that the first model is generating and the model's perceived

realness of the image being generated. Chen et al. (2024) described diffusional models image generating process as beginning with an image that is entirely comprised of the “noise” of various images it has been trained on and with each step, it reduces the noise to produce a clear image. They describe diffusion models as having surpassed the ability of both GANs or VAEs. Both the article by Chen et al. (2024), as well as an experiment conducted by Ho et al. (2020), detail that all types of image-generating artificial intelligence models are incapable of replicating images exactly, as they create images based off of the amalgamation of the thousands or millions of images that the model is trained on. Even when provided a copy of an image to be replicated with text prompts to copy the image, the image generation models begin with alternative materials and edit the image until it shares similarities or features with the original image.

It is important to consider the ethical implications of the use of large language models or image-generating models and the potential data they’re trained on whenever discussing the use of AI. Illia et al. (2022) described many of the ethical challenges with generative AI, such as its ability to amplify biases and learn implicit patterns from within the data related to sexism, racism, gender inequalities, and other types of discrimination. The research further details the lack of information regarding the type of data models are trained with, and a lack of accountability for the products of AI-generated content. Zhou et al. (2024) collected online comments regarding Sora, OpenAI’s advanced text-to-video generative AI model. Many of the comments showed public concern for Sora’s ability to create content that blurs the lines between real and fake content. Both the studies by Illia et al. (2022) and Zhou et al. (2024) discussed the ethical considerations of data privacy and copyright issues as additional aspects of the ethical challenges of generative AI and the data they are trained with.

Research conducted by Pataranutaporn et al. (2025) utilized AI to edit real images or generate fake images and used a control group of legitimate images that were not created or edited by AI. In addition to showing participants the images to assess their memory in recalling them, they also asked participants to score their level of confidence in the accuracy of their memory. The research showed that AI-edited content not only boosts the likelihood of false memories, but participants also had a high degree of confidence in the accuracy of their recollection of false stories.

Relatedly, Motoki et al. (2025) conducted multiple experiments to assess political bias of the AI applications ChatGPT-3.5, ChatGPT-4, and DALL-E 3. When impersonating the political values of the “average American”, both ChatGPT-3.5 and ChatGPT-4 leaned further left than

the known population distribution. For DALL-E 3 image generation, ChatGPT and Google's Gemini were asked to analyze and compare generated images and then create prompts that were fed into DALL-E 3, which also showed more left-leaning results. The researchers noted that ChatGPT refused to generate images for negative aspects of certain themes such as racial equality, but only for the right-wing perspective, stating that it was refusing because creating an image from that perspective could "propagate stereotypes, misinformation, or bias". The authors indicated that users must exercise caution when using any version of GPT with politically charged content as users are unlikely to receive entirely unbiased responses.

Vividness is an important aspect to consider in evaluating a person's ability to discern what's real and fake. Lee and Shin (2022) conducted two experiments to study the effect of vividness. The first study evaluated the perceived vividness of the content and the results showed that the perceived vividness of the source was the highest with the deepfake video news condition versus an AI generated fake story that had only images and text. By conducting a second experiment that included tags to indicate to participants whether the news was fake, they found that including the tags was an effective way to weaken the strength of the validity which led to less engagement with the fake stories. Less engagement meant that the fake stories were not being shared with broader audiences, mitigating the spread of misinformation.

Guo et al. (2025) conducted an experiment aimed at providing specific media literacy tips regarding how to stop or slow the rapid dissemination of AI generated images with the goal to reduce the susceptibility to AI-generated visual misinformation (AIVM). They randomly assigned participants to one of three groups: a control group that was provided with no media literacy tips, a treatment group that was provided general tips on spotting misinformation, and a second treatment group that was provided specific media literacy tips aimed at detecting AIVM. Their research showed that both media literacy treatments reduced belief in AIVM compared to the control, with specific tips reducing belief in AIVM more than the general tips. They contrasted these positive results with results that showed that the inclusion of both specific and general media literacy tips reduced belief in real headlines compared to the control. This implies that in the present study, participants ability to correctly identify the AI-altered/generated images may lead to an increased sense of skepticism and lower overall scores when identifying the real images.

The present study seeks to build on previously mentioned research by evaluating participant's ability to discern between real, unedited photographs and AI-generated or AI-

altered images. Participants completed a survey consisting of a mix of 15 real, unedited photographs, AI-altered images, or fully AI-generated images to obtain a baseline score. Afterwards, participants completed a treatment involving training to spot some of the key differences between real, unedited photographs and AI-altered images. After completion of the training portion, the participants complete a second survey with another set of 15 images to evaluate changes in scores. The study used a repeated-measures design, with exposure to AI-images acting as the independent variable, and the baseline survey as the dependent variable. We hypothesized that there will be a statistically significant difference between participant's baseline survey scores and their post-treatment scores to examine for a causal effect from the training module.

## Method

### Participants

Participants were recruited through the psychology department at a mid-sized public university. The sample consisted of 90 individuals over the age of 18. Three participant responses were excluded due to incomplete data, yielding a total of 87 participants (76 females, 6 males, and 5 non-binary individuals). Participants' ages were measured in predefined ranges, with the most common age range being 18-21 at 75.9%, followed by age 22-25 at 10.3%, and ranges 30-33 and 38-41 at 3.4%. Participation in the study took approximately 30 minutes, and participants earned points in their respective psychology courses by completing the survey. Confidentiality was maintained using random five-digit identification codes assigned to each participant, and all other identifying information was stored separate from the data.

### Materials and Procedure

A total of 45 unedited, unfiltered images were selected from the principal investigator's personal collection. Images were chosen based on the strengths and weaknesses of image generating AI models and their limitations in replicating real photographs. Images that contained animals, plants, and people were selected to play into the strengths of AI image generation, and images containing text, patterns and textures were selected to play into weaknesses. Copies were made of the original photographs, and then all 45 of the images were fed into OpenArt's diffusion-based Juggernaut XL model. To closely replicate the reference photographs with key AI artifacts despite the model's inability to create exact

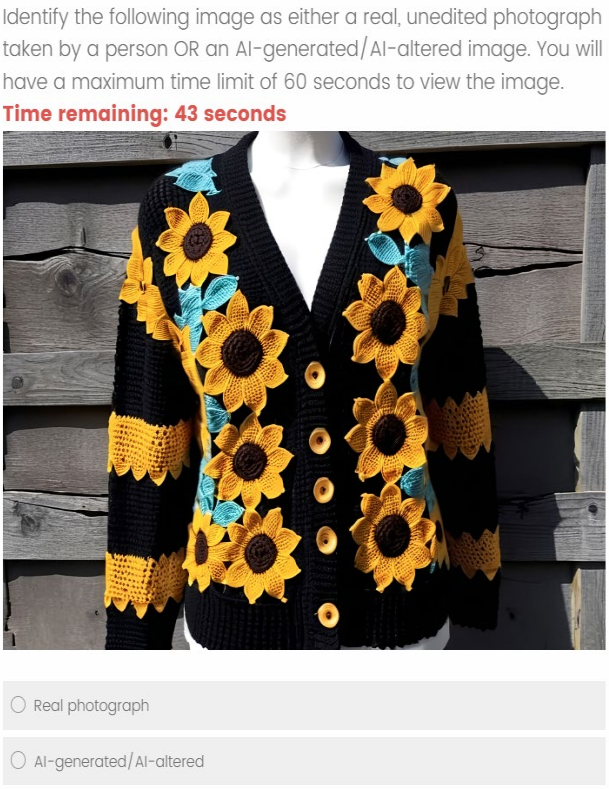
replicas of images, a text prompt stating, “Create an exact replica of this image, don’t change a thing” was included. The reference image was included in the “Image to image” guidance, with the creativity level set to an interval between 0.1-0.3 (range: 0.1-1.0). Higher numbers indicate increased creativity and less similarity to the original image. Prompt adherence was set to 15 (range: 1-15). This indicates how strictly the model will stick to the prompt. Lower numbers let AI be more creative, while higher numbers force it to stick to the text prompt. All images were set to 25 steps, which tells the AI image generator how many times to denoise the image before revealing the final product. They were also upscaled to the highest image resolution. An additional 8 images from the original set of 45 were generated into text prompts via OpenArt’s “Image to Prompt” feature creating a detailed prompt using GPT-4o-mini. The generated prompt was input into the text prompt box with the previous prompt to replicate the image removed, and the “Image to image” guidance box was intentionally left blank. All other settings remained the same. Figure 1 shows an example of a survey question with an AI-generated image that participants may have been randomly assigned during the first survey.

The survey was conducted online using Qualtrics with the same instructions and answers but different images for each survey question. Each question began with: “*Identify the following image as either a real, unedited photograph taken by a person OR an AI-generated/AI-altered image. You will have a maximum time limit of 60 seconds to view the image.*” JavaScript and HTML coding were utilized to give participants a maximum time limit of 60 seconds to analyze each image, with a countdown timer displayed above the image to inform participants of the time remaining. After 60 seconds, the image disappeared, but the survey question and answers remained. They were then prompted to pick between two answer options: “*Real photograph*” or “*AI-generated/AI-altered*”.

Of the original 45 images that were run through OpenArt to create artificially reprocessed images, 12 sets of images (original and reprocessed versions) were turned into a training module to be completed after Survey 1. To ensure participants were exposed to only one image from each set (original, reprocessed, or prompt generated), each image was assigned a number. To balance the likelihood of a real photograph or AI-reprocessed/generated image being selected, images 1-8 were given 4 entries for the real image being included in the survey, 2 entries for the AI-reprocessed images, and 2 entries for the AI-generated images, which were shuffled and randomly selected. For images 9-45, 4 entries for the real image and 4 entries for the AI-reprocessed image were shuffled and randomly selected. Survey 1 included some of the training images and was comprised of a total of 24 images consisting of 8 real photographs, 15 AI-reprocessed images, and 1 AI-prompt generated image. The Qualtrics survey was

programmed to randomly select 15 of the 24 total questions for each participant. Survey 2 included a total of 19 images consisting of 8 real photographs, 8 AI -reprocessed images, and 3 AI-prompt generated images, 15 of which were randomly selected for the second survey. Images included in Survey 1 were not included in Survey 2 and vice versa.

**Figure 1.** Sample Question




After completing Survey 1, participants then completed the treatment portion of the study that began with an introduction to the training module as well as a list of common AI mistakes or artifacts including background elements, objects, and texture/lighti ng. The tips were adapted from the Reddit forum "r/RealOrAI" page titled "Common AI Mistakes to Watch For" (2025). Following the general tips, participants continued to the interactive "Spot the Differences" game portion of the training. The training was d esigned within Qualtrics to show one set of images on the screen simultaneously with clear labels, the real photograph appearing first, and the AI-reprocessed image below it, stating: *"Click on one difference in the image below before continuing*" (see Figures 2 and 3).

**Figure 2**: Unedited photo included in the training portion of the survey.


**Figure 3.** AI-reprocessed image included in the training portion of the survey.


For the AI-reprocessed image, the Qualtrics “Hot spot” question type was utilized to allow key AI artifacts selected by the researcher to be highlighted in green when interacted with by the participants. After interacting with the AI-reprocessed image and navigating to the next page, they were shown a side-by-side comparison of the two images with differences magnified and provided a text section with “Key indicators” describing the AI-artifacts (Figure 4). After completing the training module, participants were prompted to complete Survey 2 with a second set of 15 images to determine if each image was real or AI-generated or altered. Each question was scored out of 15 to provide pre- and post-training scores.

Deception was not used for this study. At the end of the second survey, participants were debriefed and given information on the purpose of the study. Following this, participants exited the study administration system.

**Figure 4** : Comparison between real and AI-altered image.


## Results

Participants' post-treatment survey scores significantly increased following the training portion of the study compared to their pre-treatment scores. A paired-samples t-test showed that post-treatment survey scores ($M = 11.66, SD = 2.25$) were significantly higher than pre-treatment scores ($M = 9.05, SD = 2.61$), $t(86) = -10.19, p < .001, 95\%\ CI\ [-3.12, -2.10]$. The training produced a substantial improvement in scores, demonstrating a large effect size with Cohen's $d = 1.09$. A one-sample t-test of the difference between pre-treatment scores and post-treatment scores revealed that the mean difference in the number of correct answers between the pre- and post-training surveys was significant, $t(86) = 10.17, p < .001, 95\%\ CI\ [2.10, 3.12]$. The average improvement after the training was 2.61 correct answers ($SD = 2.39$). A repeated measures ANOVA further confirmed a significant effect of the training on the post-treatment scores, $F(1,86) = 103.38, p < .001$, partial $\eta^2 = .546$, indicating that over half of the variance in scores was explained

by the training portion of the study (see Tables 1 and 2). Figures 5 and 6 provide visual representations of these results.

**Figure 5**. Distribution of Pre- & Post-treatment Scores and Their Difference

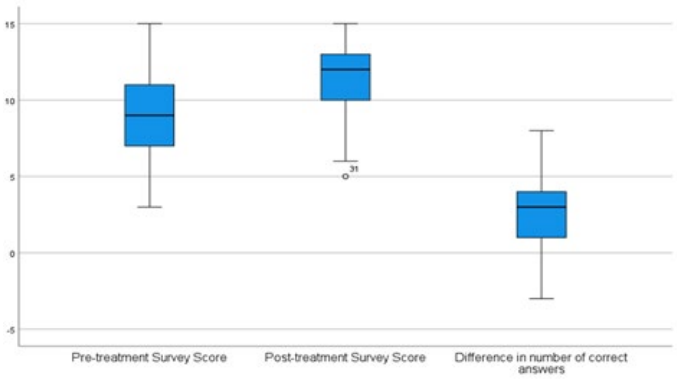



**Figure 6**. Pre- and Post-Treatment Scores

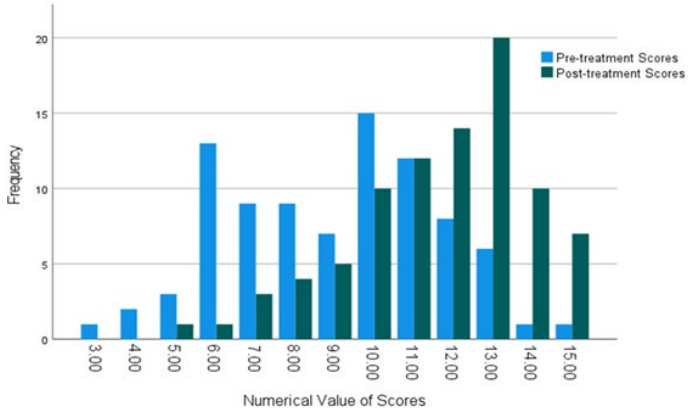


**Table 1.** Tests of Within-Subject Effects

a. Computed using alpha = .05

| Source | Procedure | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared | Noncent. Parameter | Observed Power[a] |
|---|---|---|---|---|---|---|---|---|---|
| Survey | Sphericity Assumed | 296.144 | 1 | 296.144 | 103.380 | <.000 | .546 | 103.380 | 1.000 |
| | Greenhouse-Geisser | 296.144 | 1.000 | 296.144 | 103.380 | <.000 | .546 | 103.380 | 1.000 |
| | Huynh-Feldt | 296.144 | 1.000 | 296.144 | 103.380 | <.000 | .546 | 103.380 | 1.000 |
| | Lower-bound | 296.144 | 1.000 | 296.144 | 103.380 | <.000 | .546 | 103.380 | 1.000 |
| Error (Survey) | Sphericity Assumed | 246.356 | 86 | 2.865 | | | | | |
| | Greenhouse-Geisser | 246.356 | 86.000 | 2.865 | | | | | |
| | Huynh-Feldt | 246.356 | 86.000 | 2.865 | | | | | |
| | Lower-bound | 246.356 | 86.000 | 2.865 | | | | | |

**Table 2.** Tests of Between-Subjects Effects (Measure: MEASURE_1; Transformed Variable: Average)

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared | Noncent. Parameter | Observed Power[a] |
|---|---|---|---|---|---|---|---|---|
| Intercept | 18641.385 | 1 | 18641.385 | 2068.286 | <.000 | .960 | 2068.286 | 1.000 |
| Error | 775.115 | 86 | 9.013 | | | | | |

a. Computed using alpha = .05

A repeated measures ANOVA was conducted to examine changes in survey scores with age, gender, and ethnicity included as between-subjects factors. Results revealed a significant improvement from the training, $F(1,64) = 24.36, p < .001,$ partial $\eta^2 = .276$. The between-subjects effects of age ($p = .142$), ethnicity ($p = .074$), and gender ($p = .993$) were not significant, suggesting that improvement was consistent across demographic groups. Table 3 displays the demographic frequencies of the individuals that participated in this study.

**Table 3.** Demographic Frequencies of Study Participants

| Category | Demographic | Frequency | Percent | Valid Percent |
|---|---|---|---|---|
| Ethnicity | Asian | 2 | 2.3 | 2.4 |
| | Black or African American | 4 | 4.6 | 4.7 |
| | Hispanic or Latino | 17 | 19.5 | 20.0 |
| | Native Hawaiian or other Pacific Islander | 2 | 2.3 | 2.4 |
| | White/Caucasian | 57 | 65.5 | 67.1 |
| | Other | 3 | 3.4 | 3.5 |
| | Missing | 2 | 2.3 | |
| | Total | 87 | 100.0 | 100.0 |
| Gender | Female | 76 | 87.4 | 87.4 |
| | Male | 6 | 6.9 | 6.9 |
| | Non-binary | 5 | 5.7 | 5.7 |

| Category | Demographic | Frequency | Percent | Valid Percent |
|---|---|---|---|---|
| | Total | 87 | 100.0 | 100.0 |
| Age | 18-21 | 66 | 75.9 | 75.9 |
| | 22-25 | 9 | 10.3 | 10.3 |
| | 26-29 | 1 | 1.1 | 1.1 |
| | 30-33 | 3 | 3.4 | 3.4 |
| | 34-37 | 2 | 2.3 | 2.3 |
| | 38-41 | 3 | 3.4 | 3.4 |
| | 46-49 | 2 | 2.3 | 2.3 |
| | 58-61 | 1 | 1.1 | 1.1 |
| | Total | 87 | 100.0 | 100.0 |

*Note*. $N = 87$

To evaluate the pre and post-training accuracy by image type, a 2x3 repeated measures ANOVA showed significant effects between surveys, $F(1,56) = 46.403, [< .001, \text{partial } \eta^2 = .453$, and image type, $F(2,55) = 8.305, p < .001, \text{partial } \eta^2 = .232$, indicating that accuracy varied across image types and that accuracy improved overall. The interaction between image types and surveys was also significant, $F(2,55) = 32.822, p < .001, \text{partial } \eta^2 = .544$. The mean accuracies for the sums of the three image types across all questions can be seen in Table 4. Accuracy for AI-generated and AI-altered images increased from Survey 1 to Survey 2 while accuracy for real images slightly decreased, as displayed in Figure 7.

To ensure the appropriate use of parametric analyses of the data, visual analysis of histograms with normal curves and Q-Q plots showed that the data were approximately normally distributed with mild deviations. To further support the use of parametric analyses, a multiple regression analysis was conducted to predict improvement scores from pre-treatment survey score and demographic variables (See Figure 8). The overall model was significant, $F(4,80) = 12.495, p < .001$, accounting for approximately 39% of variance

in improvement scores $R^2 = .385$). Only pre-treatment survey score was a significant predictor ($p < .001$), while age ($p = .286$), gender ($p = .269$), and ethnicity ($p = .240$) were not significant.

**Table 4.** Descriptive Statistics

| Statistic | Mean | Std. Deviation | N |
|---|---|---|---|
| Accuracy mean for real images | .7445 | .16338 | 87 |
| Accuracy mean for AI generated images | .7289 | .27820 | 87 |
| Accuracy mean for AI altered images | .6411 | .19223 | 87 |

**Figure 7.** Mean Accuracy by Image Type and Survey.

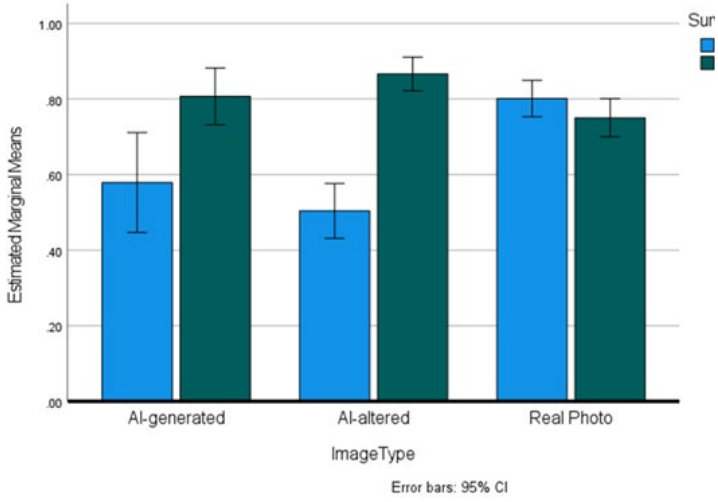
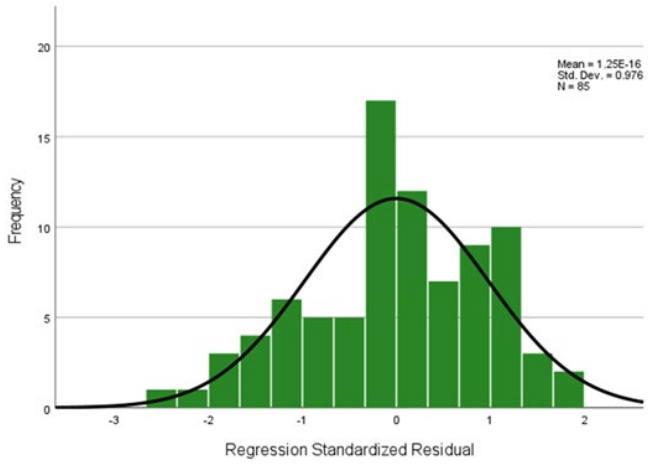



**Figure 8.** Difference in Number of Correct Answers.




## Discussion

The current study examined participant's ability to discern between real, unedited photographs and AI -altered/generated images via a 15-question survey. Additionally, we sought to provide a treatment portion in the form of training on some of the key artif acts included in images that have been reprocessed by AI, followed by a second survey to evaluate the effectiveness of the training and its impact on participant's

scores. We hypothesized that there would be a significant improvement in accuracy between pre-treatment and post-treatment survey scores, which was supported by multiple statistical methods of analysis. The analyses showed a significant improvement in accuracy following the treatment portion of the study, regardless of demographic variables. The most significant improvement was participants' ability to correctly identify images that were AI-altered, while their ability to correctly identify real, unedited photographs slightly decreased.

While we sought to analyze the effect of the treatment on the survey score that followed it, meaningful effect size on the improvement of the scores provided a clear benefit to those participants whose scores improved. One limitation, however, is that participants did not receive feedback for the accuracy of their answers and unfortunately, would not have had confirmation if their own ability to detect images correctly improved. Future research should include feedback for the benefit of the participants. While it is not ideal that the treatment portion led to a slight reduction in the detection of real images, these results are in line with the results from the study completed by Guo et al. (2025) surrounding media literacy tips for detecting AI-generated misinformation. The researchers cited a common side effect of media literacy interventions as reducing the belief in real information due to an increased level of skepticism. The higher accuracy in detecting AI-altered/generated images from enhanced skepticism is seen as a benefit of the research. The research conducted by Lee and Shin (2022) that involved providing a "false-flag" indicator on fake news to evaluate engagement levels had a small-to-medium effect size, and they explained possible reasoning for this as being explained by the emotional impact of stories or possible reputational gains that drive content engagement. Lee and Shin (2022) offered the reinforcement that future research should aim to develop literacy interventions specifically designed to make audiences pause and think about the validity of the message and encourage scrutinized processing of multimodal information such as images and video.

Some of the additional potential limitations to the current study included an unbalanced set of images due to both a scarcity of fully AI-generated images as well as the random assignment of each set of 15 survey questions. Some participants may have had more images with text (a common area where image-generating models make mistakes) in either of their surveys, leading to higher scores than participants who had fewer survey questions that included images with text. Additionally, a sample image at the beginning of the pre-treatment survey should have been included to provide participants with an idea of what an AI-reprocessed image looks like, although the places where audiences are most likely to be exposed to AI content are typically via online sources which are not typically flagged as being AI.

For this research to be generalized and applied to larger populations, future research should sample larger, more diverse populations, as all of the participants in the current study were recruited through university professors and may have higher levels of education than the general population at large. The

sample also had significantly more female participants than male, and the majority of participants were in the 18–21-year age group, as displayed in Table 1.

The most important limitation and consideration for future research would revolve around the rate at which AI models are advancing. Liu et al. (2024) conducted a thorough analysis of image-generating AI models, specifically focusing on OpenAI's image-generating model called Sora, which was introduced in February 2024. Their research was published in April of 2024 and emphasized the need for interdisciplinary collaboration involving the fields of law and psychology to ensure safety and define appropriate norms before the roll out of Sora to the general public. Roughly six months after the publication of their research, Sora was made available to the general public with minimal safety measures in place. As stated by Guo et al. (2025), "Specific tips about AIVM (AI-generated visual misinformation) may quickly become obsolete due to the ever-changing landscape of AI, which may pose challenges to implementing real-world media literacy interventions." Many of the key indicators and artifacts included in the current study such as illegible text, inconsistent textures, and other irregularities will likely be phased out of future image-generating AI models. This means that future research should not focus on specific media literacy tips and it's generalizability for the general population but should instead focus on increasing skepticism and critical thinking skills.

## References

Chen, M., Mei, S., Fan, J., & Wang, M. (2024). Opportunities and challenges of diffusion models for generative AI. *National Science Review*, *11*(12). https://doi.org/10.1093/nsr/nwae348

*Common AI mistakes to watch for: R/realorai*. (2025, July 14). Reddit - the heart of the internet. https://web.archive.org/web/20251201044802/https://www.reddit.com/r/RealOrAI/wiki/common-ai-mistakes/#wiki_common_ai_mistakes_to_watch_for

Giordano, G., Catone, M., & Primerano, I. (2025). The fake news phenomenon in the scientific debate: Evidence from a bibliometric analysis. *Social Indicators Research, 177*, 31–52. https://doi.org/10.1007/s11205 -024-03485-7

Grinfeld, G., De Keersmaecker, J., Roets, A., & Unkelbach, C. (2025). Registered report: Does repetition increase the credibility of AI-generated images? *Journal of Experimental Psychology Learning Memory and Cognition*. https://doi.org/10.1037/xlm0001505

Guo, S., Swire-Thompson, B., & Hu, X. (2025). Specific media literacy tips improve AI-generated visual misinformation discernment. *Cognitive Research Principles and Implications*, *10*(1), 38. https://doi.org/10.1186/s41235 -025-00648-z

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models* . arXiv.org. https://doi.org/10.48550/arXiv.2006.11239

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2022). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics the Environment & Responsibility, 32*(1), 201–210. https://doi.org/10.1111/beer.12479

Lee, J., & Shin, S. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology*, *25*(4), 531–546. https://doi.org/10.1080/15213269.2021.2007489

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., He, L., & Sun, L. (2024). SORA: a review on background, technology, limitations, and opportunities of large vision models. *arXiv (Cornell University)* https://doi.org/10.48550/arxiv.2402.17177

Motoki, F., Neto, V., & Rangel, V. (2025). Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, *234*, 106904. https://doi.org/10.1016/j.jebo.2025.106904

Otgaar, H., Howe, M., & Patihis, L. (2022). What science tells us about false and repressed memories. *Memory, 30*(1), 16–21. https://doi.org/10.1080/09658211.2020.1870699

Pataranutaporn, P., Archiwaranguprok, C., Chan, S., Loftus, E., & Maes, P. (2025). Synthetic human memories: AI-Edited images and videos can implant false memories and distort recollection. *Association for Computing Machinery* , 1–20. https://doi.org/10.1145/3706598.3713697

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://arxiv.org/abs/1706.03762

Zhou, K., Choudhry, A., Gumusel, E., & Sanfilippo, M. (2024). "Sora is incredible and scary": Emerging governance challenges of text-to-video generative AI models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.11859