Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2025, Vol. 69(1) 2000–2005


DOI: 10.1177/10711813251363209
journals.sagepub.com/home/pro
Sage

Detecting Midjourney-Generated Images: An Eye-Tracking Study

Joost C. F. de Winter[1], Jenna Pfeifer[1], Dimitra Dodou[1], and Yke Bauke Eisma[1]

Abstract

This study investigated human performance in identifying AI-generated images. In a speeded forced-choice task, 255 participants viewed paired images (one real, one AI-generated by Midjourney) of standard or futuristic cars and buildings and had to identify the AI-generated one, while eye movements were recorded using an eye-tracker. Results revealed a powerful "futurism-as-artificiality" heuristic. Specifically, participants performed poorly (55% correct) when an AI-generated standard image was paired with a real futuristic image. Conversely, accuracy was high (91% correct) when the AI-generated futuristic image was paired with a real standard image. Participants' gaze landed first on the AI-generated image more often when it depicted a futuristic design than when it depicted a standard one. The demonstrated heuristic presents a double-edged sword for information veracity: it may lead to the uncritical acceptance of AI-generated misinformation that appears conventional, while simultaneously causing real forward-thinking designs to be dismissed as fake.

Keywords

AI-generated images, visual attention, futurism-as-artificiality heuristic

Introduction

As AI becomes increasingly capable of generating text, sound, and images, a fundamental question emerges: How can we determine whether what we read, hear, or see is real or artificially generated? Originally, this challenge was addressed in the Turing Test, which examines whether humans can distinguish computer-generated output from that of a human.

In texts written by students, academics, and others, there is increasing evidence that these texts are often (partially) generated by a large language model (LLM; De Winter et al., 2023; Liang et al., 2024). These suspicions are supported by evidence such as the prevalence of certain writing styles or words compared to fully human-written text (Soto et al., 2024). Similar issues are also emerging with photos, videos, products, and interfaces, where there is increasing debate as to whether the material is real or fake (e.g., Cooke et al., 2025; Lu et al., 2023).

When assessing the authenticity of images, viewers may apply a variety of strategies. One strategy is to evaluate features such as texture and lighting. Another is to identify implausible objects or situations or to detect artifacts (Kamali et al., 2024; Mathys et al., 2024). Beyond artifact detection, individuals may assess authenticity by comparing images to established mental schemas and prototypes of "realness". Consequently, highly novel real-world depictions, such as futuristic designs, might appear unconvincing because they deviate from familiar prototypes.

The current study aimed to gain insight into how people determine what is real or fake. Participants were presented with two images at a time, one real and one AI-generated. The participants' task was to determine as quickly as possible which of the two images was the fake one. We used images that may frequently appear in the real world (standard cars and buildings) as well as images of a "novel" nature (futuristic cars and buildings). Our hypothesis was that participants would struggle to distinguish real from fake when a real futuristic object was paired with a fake standard object. We reasoned that the inherent unfamiliarity of the real futuristic object might make it appear less convincing compared to the AI-generated, but familiar, standard object, potentially overriding subtle cues that might otherwise indicate the standard object was fake. Apart from measuring response accuracy and response times, eye-tracking was included to explore how participants allocated their visual attention.

[1]Delft University of Technology, The Netherlands

Corresponding Author:
J. C. F. de Winter, Faculty of Mechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands.
Email: j.c.f.dewinter@tudelft.nl

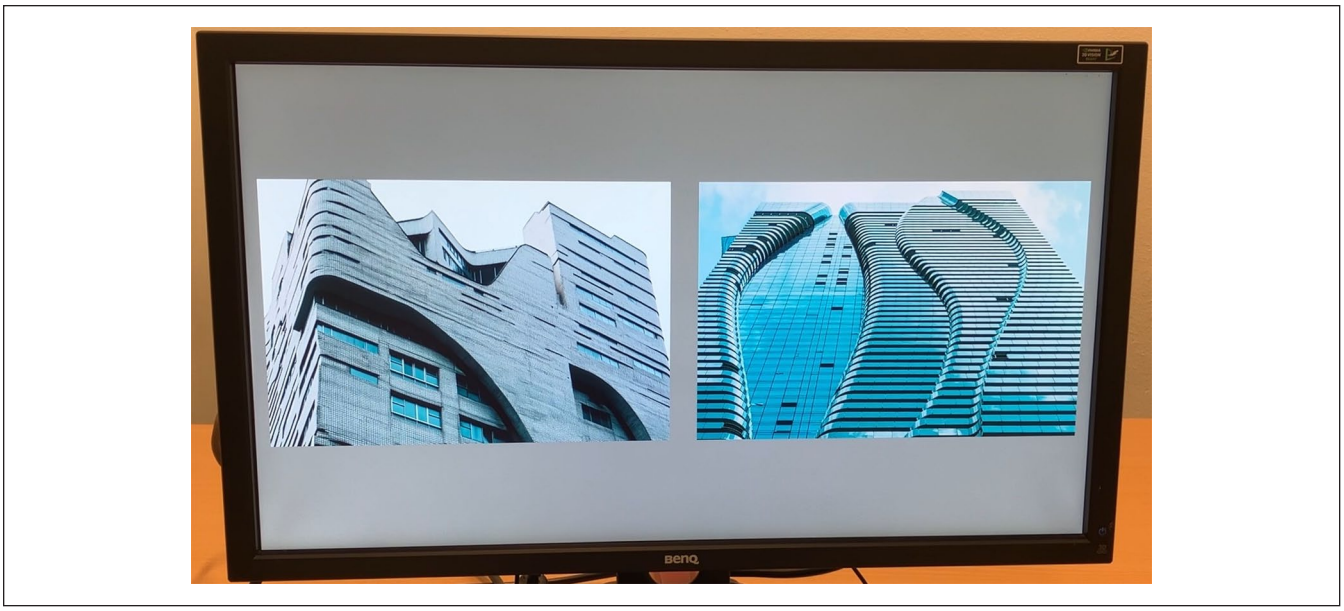

Figure 1. Image pair displayed on the computer screen. In this image pair, the AI-generated standard image is shown on the left, while the right image is a real futuristic building. 42.6% of participants correctly pressed the left shift key to indicate that the left image was AI-generated.
Source. Right image: Photo by John Salvino on Unsplash).

Methods

This experiment constituted Task 2 of a three-task battery; the design and results of Task 1 (free-viewing task) are reported elsewhere (Pfeifer et al., 2025). Participants were MSc engineering students who were enrolled in the TU Delft course Human-Robot Interaction. Recruitment procedures, testing environment, and ethical clearance (HREC #4742) were identical to those described by Pfeifer et al.

A total of 255 MSc students participated. They had a mean age of 23.5 years ($SD = 1.9$); one participant did not report a valid age. The participants consisted of 198 males (77.6%), 56 females (22.0%), and one individual (0.4%) who preferred not to disclose their gender.

A 24-inch BenQ XL2420 monitor (1,920×1,080 px) was used for stimulus presentation. Participants were seated 97 cm from the screen with their heads placed in a support. Eye movements were recorded using an EyeLink 1000 Plus. Eye-tracking data were available for 241 participants.

Participants completed 40 trials. In each trial, they viewed a pair of images, one real, one AI-generated, presented side-by-side (see Figure 1 for an example). Participants had to identify the AI-generated image as quickly as possible by pressing the corresponding left or right shift key. 129 participants viewed 40 image pairs of one image set, while the remaining 126 participants viewed 40 different image pairs of a second image set. Because two image sets were used, each containing 40 pairs, the study comprised 80 unique image pairs in total.

Each trial began with a fixation cross, followed by the image pair for 5,000 ms. Feedback ("Correct!" in green, "Incorrect!" in red, or "No response detected. Go faster next time!" in black) was shown for 2,250 ms after each trial. Each participant viewed 20 building and 20 car image pairs. The pairs were distributed as follows: (1) 10 pairs: real standard versus AI-generated futuristic, (2) 10 pairs: real futuristic versus AI-generated futuristic, (3) 10 pairs: real standard versus AI-generated standard, (4) 10 pairs: real futuristic versus AI-generated standard. The presentation order of the 40 image pairs was randomized for each participant. The on-screen position of the images (left vs. right) was fixed to a single predetermined sequence for the first eight participants and randomized for all subsequent participants.

The majority of the real photos were retrieved from Unsplash or Wikimedia Commons. The AI counterparts were generated through Midjourney version 6.1, using the "Imagine" function, for example, "Generate a futuristic variant of this car."

We calculated the following measures for each of the 80 image pairs: (1) Response accuracy (%), where non-responses (2.39% of all trials) were marked as incorrect; (2) Response time (ms). Non-responses were assigned a response time of 5,000 ms. (3) Gaze entry (% of trials in which the image was glanced at first). Using a coordinate system with the origin at the top left (0, 0), the left image spanned pixels 32 to 927 horizontally, while the right image spanned pixels 992 to 1,887. Gaze entry was defined as the first moment after trial onset at which the horizontal gaze coordinate

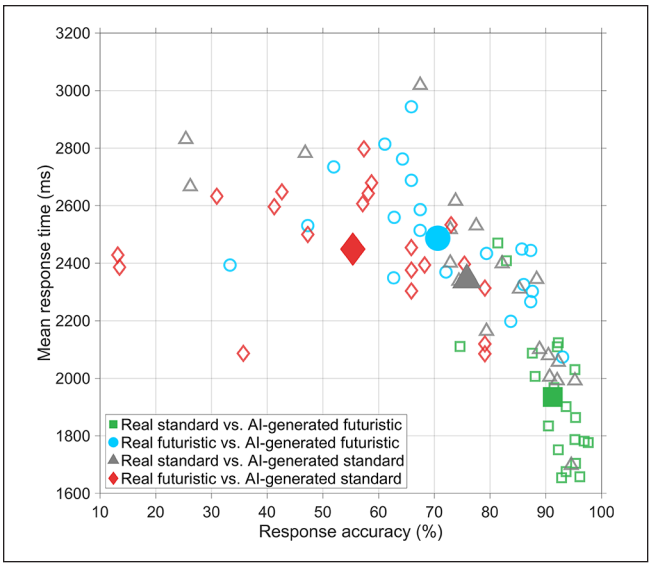

Figure 2. Mean response time versus percentage correct for the 80 image pairs. Open markers represent image pairs, while filled markers represent means of 20 image pairs.

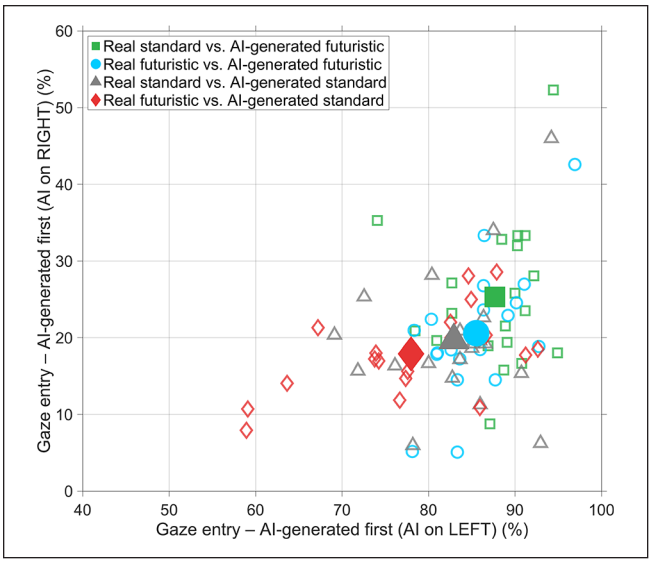

Figure 3. First gaze entry percentage on the AI-generated image as a function of its presentation side (right vs. left). Open markers represent image pairs, while filled markers represent means of 20 image pairs.

reached a threshold value of $x \leq 900$ pixel units for the left image or $x \geq 1{,}020$ pixel units for the right image.

In a post-experiment questionnaire, participants were asked "What information or strategies did you use to determine which image was authentic and which was AI-generated?" Ten categories were defined based on the participants' responses, using ChatGPT. Subsequently, the responses were automatically coded into these predefined categories using OpenAI's o3 (o3-2025-04-16, via the API service, with reasoning effort set to high). The participants' responses were presented to o3 in 17 randomly permuted batches of 15. This entire process was repeated three times, and a majority vote across the three rounds determined the final score (0 or 1) for each of the 255 comments and 10 categories. The prompt used was as follows: "Categorize the following 15 comments into the above 10 categories. A comment can be placed in more than one category, but be conservative. Produce a comma-separated 15×10 matrix consisting of 0s and 1s, nothing else."

Results

Figure 2 shows the mean response time versus accuracy in detecting the AI-generated image. Accuracy was higher when the AI-generated image was futuristic and the real image standard (green; $M = 91.3\%$, $SD = 5.8\%$, $n = 20$ image pairs) than when the AI-generated image was standard and the real image futuristic (red; $M = 55.4\%$, $SD = 20.3\%$, $n = 20$ image pairs), $t(38) = 7.58$, $p = 4.06 \times 10^{-9}$. Similarly, response times were faster when the AI-generated image was futuristic and the real image standard ($M = 1{,}935$ ms, $SD = 234$ ms, $n = 20$ image pairs) compared to when the AI-generated image was standard and the real image was futuristic ($M = 2{,}449$ ms, $SD = 201$ ms, $n = 20$ image pairs), $t(38) = -7.44$, $p = 6.26 \times 10^{-9}$.

The gaze-first percentages in Figure 3 revealed a strong overall left bias (e.g., Ossandón et al., 2014). In 9.9% of all trials, the participant was already looking at the left image at the start of the trial, while in 1.4% of all trials, the participant was already looking at the right image (these were also counted as the first gaze). If the AI-generated image was on the left, it was glanced at first in 87.7% of trials ($SD = 5.3\%$, $n = 20$) when it was futuristic and paired with a standard real image, compared with 78.0% ($SD = 9.9\%$, $n = 20$) when it was standard and paired with a futuristic real image, $t(38) = 3.86$, $p = 4.32 \times 10^{-4}$. If the AI-generated image was on the right, it was glanced at first in 25.3% of trials ($SD = 9.5\%$, $n = 20$) when it was futuristic and paired with a standard real image, compared with 17.9% ($SD = 5.5\%$, $n = 20$) when it was standard and paired with a futuristic real image, $t(38) = 3.03$, $p = 4.40 \times 10^{-3}$.

The results of the LLM-based analysis of the free-response question (Table 1) indicated that a large proportion of participants referred to the background, environment, or sky ($n = 94$), as well as lighting, reflections, and shadows ($n = 135$). Some participants reported that unrealistic shapes or futurism itself was a reason for believing the image was AI-generated ($n = 79$). In a substantial number of cases, participants referred to existing knowledge ($n = 46$) or more specific aspects such as texts, logos, brand names, or license plates ($n = 43$). For example, participants believed a car was fake because of a non-existent car logo, or they identified a

Table 1. Frequency of Self-Reported Cues and Strategies for Distinguishing Real and AI-Generated Images ($n = 255$ Participants; Multiple Categories Per Response Allowed).

Category	*n* (sorted in descending order)
Lighting, reflections, and shadows	135
Background/environment/sky	94
Too perfect/no imperfections (smoothness, blur)	91
Unrealistic shapes/futuristic design	79
Prior knowledge (recognizing real objects/places)	46
Text, logos, brand names, license plates	43
Color saturation/glow	35
Gut feeling/intuition	31
Human presence or depiction	19
Other/none	3

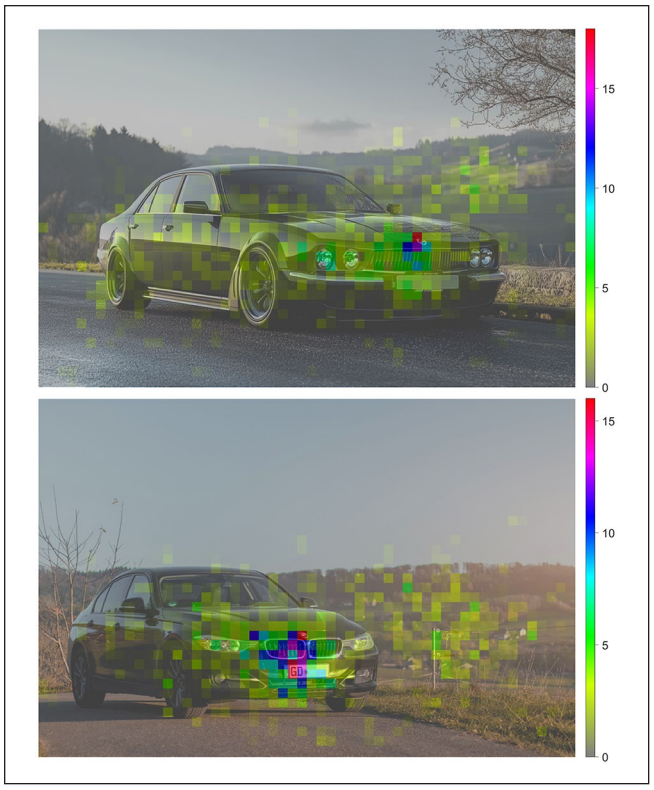

Figure 4. Heatmap of all gaze points directed to an AI-generated standard image (top) and its real standard counterpart (bottom). The heatmaps were created by dividing the images into 16 × 16 squares and summing the number of gaze points within these squares from all participants, and then linearly scaling these counts to an arbitrary unit.

building known to exist in reality, meaning the other image must have been fake.

The heatmaps for the image pair in Figure 4 show that participants' gazes were often concentrated on the front of the car, particularly its logo. This focus may reflect a deliberate strategy to find artifacts, which would align with self-reports where participants mentioned checking logos and brand names (Table 1). However, an alternative explanation is that these central and well-defined features are visually salient, and drew attention regardless of the participant's specific strategy.

Discussion

Generative AI for image creation offers various opportunities, including cost-effective production of advertising (Hartmann et al., 2025), entertainment material (Schatten, 2024), or product designs (Paliwal et al., 2024). However, it also introduces deception risks. This study examined how well participants could distinguish between real and AI-generated images.

Literature indicates that humans often struggle to distinguish AI-generated from real content, with many studies reporting performance that is poor or even near chance level (e.g., Cooke et al., 2025; Diel et al., 2024; Frank et al., 2024; Nightingale & Farid, 2022; Partadiredja et al., 2020). Our two-alternative forced-choice design, which allows for direct comparison and may be inherently easier than tasks requiring evaluation of a single image, likely contributed to a relatively high overall accuracy of 73.3%.

However, this aggregate figure conceals a performance discrepancy driven by a "futurism-as-artificiality" heuristic. Accuracy was high (91.3%) for pairings congruent with this mental shortcut (real standard vs. AI-generated futuristic), but dropped to just 55.4% for incongruent pairings that challenged it (real futuristic vs. AI-generated standard), with intermediate results for the neutral conditions (75.8% for real standard vs. AI-generated standard and 70.6% for real futuristic vs. AI-generated futuristic). Our findings point to a heuristic more specific than the often-cited "seeing-is-believing" tendency (cf. Köbis et al., 2021; Tahir et al., 2021). That is, participants were not inclined to trust AI-generated images in general, but rather to associate futuristic esthetics with artificiality.

The judgment pattern can be explained using the representativeness heuristic (Kahneman & Tversky, 1972). A standard image easily matches a mental prototype for "real", whereas a futuristic image fits the "AI-generated" prototype, partly because its esthetic conventions, such as perfect surfaces and occasional staged/unusual lighting, may resemble common AI artifacts. This cognitive association offers one perspective on why participants exhibited slower responses and lower accuracy when faced with this specific pairing.

When a real standard image was paired with an AI-generated futuristic one, participants' gaze was often drawn to the AI-generated image first. It is possible that participants could make a quick judgment using their peripheral vision, even before looking directly at the image. The AI-generated images, particularly the futuristic ones, tended to have higher contrast and more detailed textures. These

visual qualities might have made them "pop out" and capture attention.

Some participants implicitly confirmed using novelty as a cue, citing "*Unrealistic shapes / futuristic design*" as a reason for identifying an image as AI-generated (Table 1), which indicates how this heuristic can be misleading. A future is conceivable in which people incorrectly dismiss genuine innovations or designs, unfamiliar real-world scenes, or unconventional artistic expressions as artificial, simply because they deviate from the norm.

Instead of relying on misleading heuristics such as novelty or futurism, a more effective approach may be to train individuals on indicators of the AI generation process (Chen et al., 2025; Diel et al., 2024; Tahir et al., 2021). The strategies reported by our participants, focusing on lighting, backgrounds, and overly perfect shapes, mirror findings from, for example, Bozkir et al. (2025) and Huang et al. (2024), whose research also showed that people scrutinize images for common artifacts and inconsistencies typical of generative models.

Our findings may have limited generalizability due to the specific sample of young, predominantly male engineering students, whose AI detection abilities might differ from those of the broader population. This consideration is important given the findings of Cooke et al. (2025) and Lüdemann et al. (2025), who showed that older individuals perform significantly worse than their younger counterparts in detecting AI-generated content. Another limitation of our work is that we used standard and futuristic cars and buildings generated solely by Midjourney. The "futurism-as-artificiality" heuristic might not apply equally to other image categories or outputs from different AI models.

Data Availability

MATLAB scripts reproducing the results, a demonstration video of the experiment, and the Experiment Builder files are available at https://doi.org/10.4121/96b5eb59-b9a4-4094-8c47-04c14d57af1a.

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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Joost C. F. de Winter https://orcid.org/0000-0002-1281-8200
Dimitra Dodou https://orcid.org/0000-0002-9428-3261

References

Bozkir, E., Riedmiller, C., Skodras, A. N., Kasneci, G., & Kasneci, E. (2025). Can you tell real from fake face images? Perception of computer-generated faces by humans. *ACM Transactions on Applied Perception*, *22*(2), 6. https://doi.org/10.1145/3696667

Chen, E., Seo, H., Ruffin, M., Lee, D., Wang, G., & Xiong, A. (2025). A study of training strategies on enhancing human detection of AI-synthesized faces. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 372–384. https://doi.org/10.1609/icwsm.v19i1.35821

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2025). *As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli*. arXiv. https://doi.org/10.48550/arXiv.2403.16760

De Winter, J. C. F., Dodou, D., & Stienen, A. H. A. (2023). ChatGPT in education: Empowering educators through methods for recognition and assessment. *Informatics*, *10*(4), 87. https://doi.org/10.3390/informatics10040087

Diel, A., Bäuerle, A., & Teufel, M. (2024). *Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy*. OSF. https://doi.org/10.31219/osf.io/muwnj

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., & Holz, T. (2024). *A representative study on human detection of artificially generated media across countries* [Conference session]. Proceedings of the 2024 IEEE Symposium on Security and Privacy, San Francisco, CA, 55–73. https://doi.org/10.1109/SP54263.2024.00159

Hartmann, J., Exner, Y., & Domdey, S. (2025). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing*, *42*(1), 13–31. https://doi.org/10.1016/j.ijresmar.2024.09.002

Huang, J., Gopalakrishnan, S., Mittal, T., Zuena, J., & Pytlarz, J. (2024). *Analysis of human perception in distinguishing real and AI-generated faces: An eye-tracking based study*. arXiv. https://doi.org/10.48550/arXiv.2409.15498

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kamali, N., Nakamura, K., Chatzimparmpas, A., Hullman, J., & Groh, M. (2024). *How to distinguish AI-generated images from authentic photographs*. arXiv. https://doi.org/10.48550/arXiv.2406.08651

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews [Conference session]. *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, 29575–29620. https://proceedings.mlr.press/v235/liang24b

Lüdemann, R., Schulz, A., & Kuhl, U. (2025). Generation gap or diffusion trap? How age affects the detection of personalized AI-generated images. In H. Plácido & P. Silva da Cipresso (Eds.), *Computer-human interaction research and applications* (pp. 359–381). Springer. https://doi.org/10.1007/978-3-031-83845-3_22

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In A.

Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 25435–25447). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/505df5ea30f630661074145149274af0-Paper-Datasets_and_Benchmarks.pdf

Mathys, M., Willi, M., & Meier, R. (2024). *Synthetic photography detection: A visual guidance for identifying synthetic images created by AI*. arXiv. https://doi.org/10.48550/arXiv.2408.06398

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Ossandón, J. P., Onat, S., & König, P. (2014). Spatial biases in viewing behavior. *Journal of Vision*, *14*(2), 20. https://doi.org/10.1167/14.2.20

Paliwal, G., Donvir, A., Gujar, P., & Panyam, S. (2024). *Accelerating time-to-market: The role of generative AI in product development* [Conference session]. Proceedings of the 2024 IEEE Colombian Conference on Communications and Computing, Barranquilla, Colombia. https://doi.org/10.1109/COLCOM62950.2024.10720255

Partadiredja, R. A., Entrena-Serrano, C., & Ljubenkov, D. (2020). *AI or human: The socio-ethical implications of AI-generated media content* [Conference session]. Proceedings of the 13th CMI Conference on Cybersecurity and Privacy-Digital Transformation-Potentials and Challenges, Copenhagen, Denmark. https://doi.org/10.1109/CMI51275.2020.9322673

Pfeifer, J., De Winter, J. C. F., Dodou, D., & Eisma, Y. B. (2025). *Loneliness, personality, and attention to AI-generated images depicting social threat: An eye tracking study*. ResearchGate. https://www.researchgate.net/publication/390426605_Loneliness_personality_and_attention_to_AI-generated_images_depicting_social_threat_An_eye_tracking_study

Schatten, M. (2024). *AI and the future of entertainment technology*. HAL. https://hal.science/hal-04637685/document

Soto, R. A. R., Koch, K., Khan, A., Chen, B. Y., Bishop, M., & Andrews, N. (2024). *Few-shot detection of machine-generated text using style representations* [Conference session]. Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria. https://openreview.net/pdf?id=cWiEN1plhJ

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M. A., & Zaffar, M. F. (2021). *Seeing is believing: Exploring perceptual differences in deepfake videos* [Conference session]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, Article 174. https://doi.org/10.1145/3411764.3445699