Human Perception of Art in the Age of Artificial Intelligence

Jules van Hees[1], Tijl Grootswagers[1,2], Genevieve L. Quek[1], Manuel Varlet[1,3]

[1] The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia

[2] School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia

[3] School of Psychology, Western Sydney University, Australia

Corresponding author:
Associate Professor Manuel Varlet, MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia.
Email: m.varlet@westernsydney.edu.au

Abstract

Recent advancements in Artificial Intelligence (AI) have rendered image-synthesis models capable of producing complex artworks that are nearly indistinguishable from human-made works. Here we present the first quantitative assessment of human perception and preference for art generated by OpenAI's DALL·E 2, a leading AI tool for art creation. Participants were presented with pairs of artworks, one human-made and one AI-generated, in either a preference task or a discrimination task. Results revealed a significant preference for AI-generated artworks. At the same time, a separate group of participants were above-chance at detecting the AI-generated work within each pair, indicating a perceptible distinction between human and artificial creative works. This shift in art preference to favour synthetic creations is poised to revolutionise the way we think about art and its value to human society, prompting reflections on authorship, authenticity, and human creativity in the era of generative AI.

Artificial Intelligence (AI) is rapidly becoming prevalent in our everyday lives. With each iteration in technological capabilities, the gap between AI and human ability is narrowing. One such advancement has been the recent wave of image-synthesis models; AI image-generation tools that have evolved to a level of sophistication such that it is nearly impossible to distinguish between photographs of real human faces and those generated by a computer [1,2]. The fact that AI is able to fool the human visual system's perception of faces – one of our most deeply-rooted and evolutionarily relevant brain functions – is certainly cause for concern [3], but how does it fare against what is arguably the pinnacle of human creativity: Art?

Addressing this question has become increasingly important for understanding the changing landscape of the art world and the role of technology in shaping artistic production and consumption. Aside from growing concerns about intellectual property and privacy violations [4], AI-generated art raises fundamental questions about how we might (re)define creativity [5], an ability considered until now to be human-specific. Here we provide the first objective and quantitative assessment of the human perception of artificial art made using OpenAI's DALL·E 2, one of the most advanced AI tools for art generation (https://openai.com/dall-e-2). We compared human observers' appreciation of AI- and human-generated art and tested observers' ability to distinguish between the two. With high-level performance for representational image generation, DALL·E 2 represents a step change in the field, as technology and synthetic representational artworks publicly accessible before its release were often much more rudimentary in complexity of composition and general verisimilitude [6]. As such, prior research on the perception of AI-generated art has been restricted to abstract productions, which, while akin to human artwork, tend to be readily construed as artificial, especially when placed in the context of studying AI art [7].

To assess human observers' appreciation and discrimination of non-abstract artworks generated by both humans and AI, we paired 50 lesser-known real artworks by famous representational artists with 50 artificial artworks generated in a similar style using OpenAI's DALL·E 2, as depicted in Figure 1. Online observers viewed these image pairs in either a preference judgement task (Experiment 1, 'Which artwork do you like the most?', 127 participants) or a real-artificial discrimination task (Experiment 2, 'Which artwork was generated by a computer?', 137 participants) (see Figure 2A). To minimise bias and conceal the true purpose of each experiment, in both cases the 50 matched image pairs appeared randomly intermingled with random pairs drawn from the full image set.

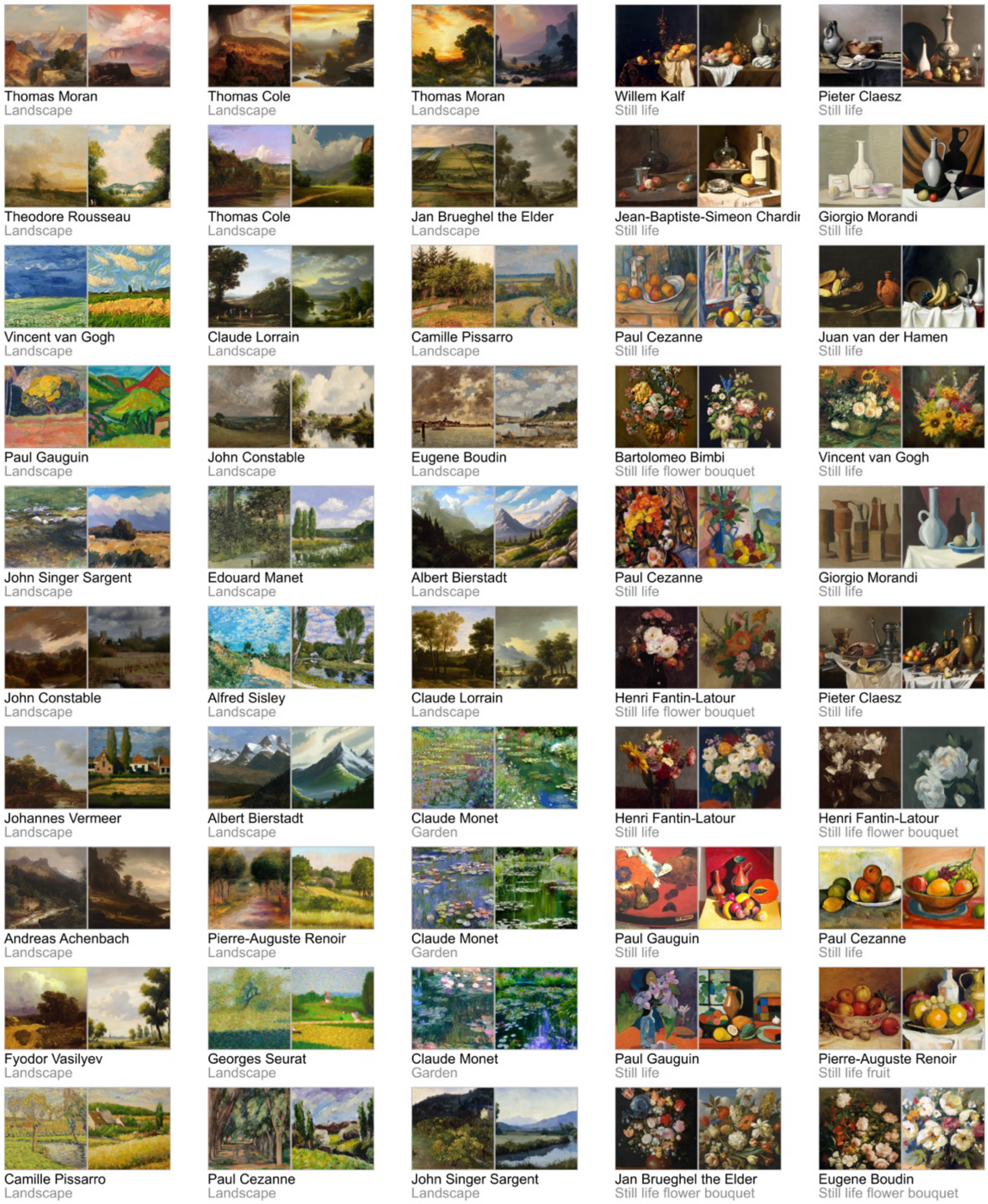

Figure 1. The 50 pairs of human-made (left image) and AI-generated (right image) artworks in Experiments 1 and 2. Corresponding author and style used as prompts in DALL·E 2 appear below each pair. Human-made images were sourced from Wikimedia Commons (https://commons.wikimedia.org) and WikiArt (https://www.wikiart.org); AI-generated images were obtained from DALL·E 2 (https://openai.com/dall-e-2).

Results

Experiment 1 revealed a significant preference for AI-generated artworks. Without being provided with any information about the origin/authorship of the artworks, participants in the preference task tended to prefer the AI-generated artworks significantly more often than the human-created artworks (AI-preference scores significantly above 50% chance-level, $t(126) = 5.39$, $p < 0.001$, $d = 0.48$).

Interestingly, when a separate group of participants in Experiment 2 were asked to detect which one of the two artworks was made by a computer, they could do so significantly better than chance. A one-sample t-test indicated AI detection accuracy was significantly above 50%, $t(136) = 3.62$, $p < 0.001$, $d = 0.31$, although a smaller effect size was observed compared to Experiment 1.

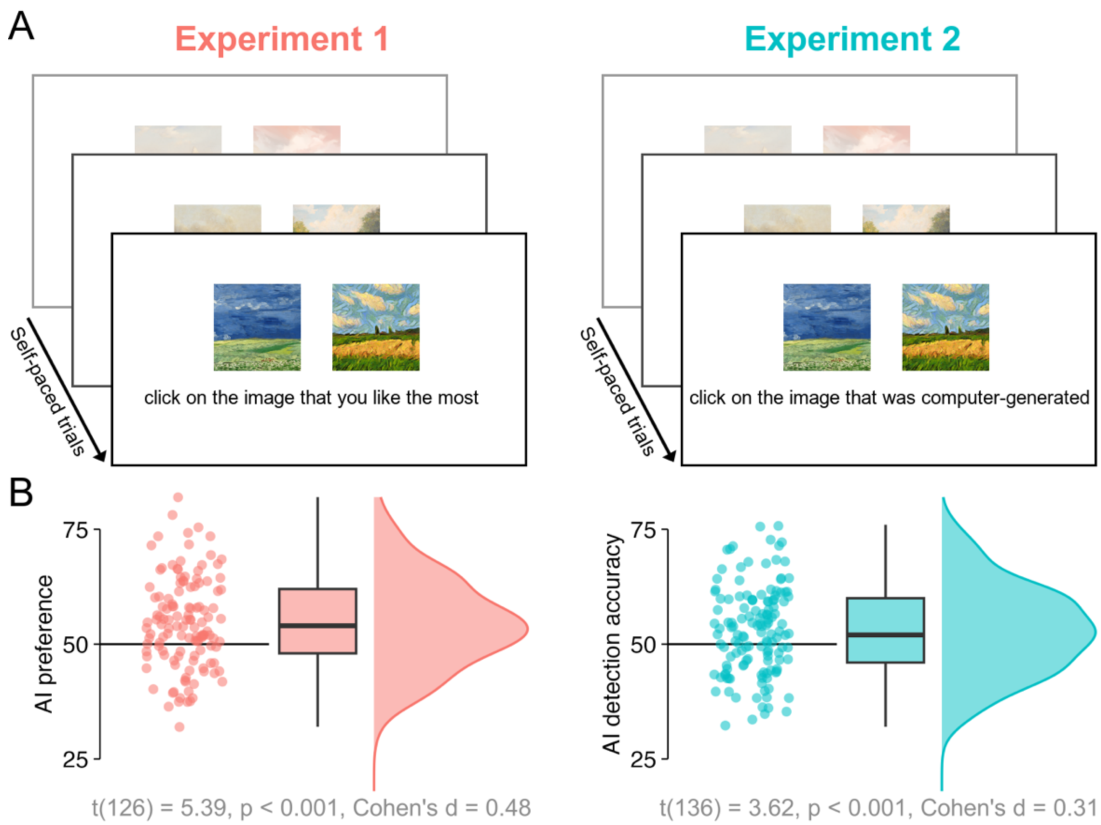

Figure 2. (A) Illustration of task displays used in Experiments 1 and 2 to respectively examine AI preference and AI-detection accuracy. There were 100 trials in each experiment, 50 of which were the critical human-AI pairs (B) Individual participants' AI preference scores (Experiment 1, at left) and AI detection accuracy scores (Experiment 2, at right) averaged across all pairs of artworks. Corresponding boxplots and distributions appear at right.

Image-wise correlational analysis revealed a positive relationship between AI preference and AI detection accuracy scores associated with each pair, $t(48) = 3.23$, $p = 0.002$, $r = 0.42$. As shown in Figure 3, the AI-generated artworks that Experiment 1 participants tended to prefer were also those that Experiment 2 participants were better able to detect, suggesting there may be features in the artworks driving both preference for and detection of AI-generated art. No significant correlations between participants' experience in art (i.e., interest and knowledge in art, see methods for further details) and AI preference and detection accuracy were observed in either experiment ($r(125) = 0.02$, $p = 0.79$, and $r(135) = 0.16$, $p = 0.06$ for Experiments 1 and 2, respectively).

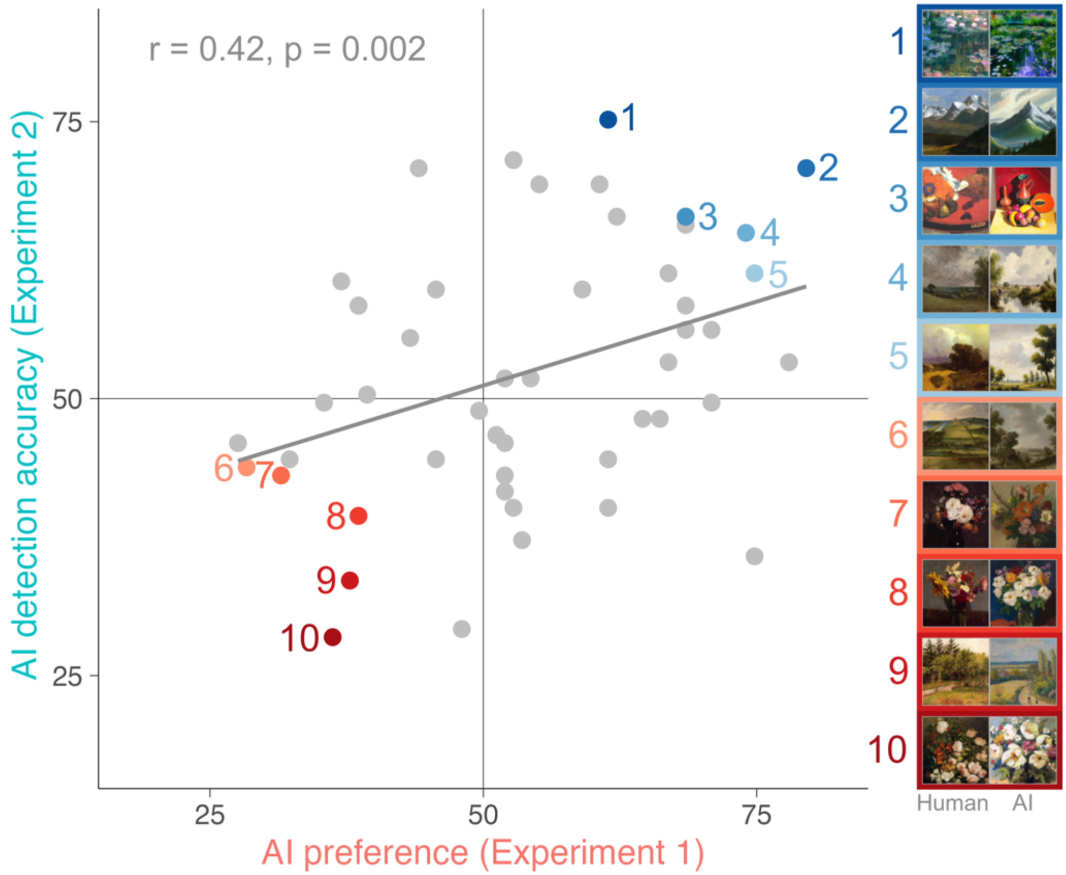

Figure 3. Correlation between AI preference scores in Experiment 1 and AI detection accuracy scores in Experiment 2. Each dot represents one human-AI artwork pair (50 in total). Coloured dots highlight the five pairs of artworks at both ends of the spectrum. The solid line represents the line of best fit.

Discussion

There is burgeoning sentiment that AI-image generation technology has reached a point of refinement that challenges our traditional understanding of the human perception and appreciation of art [5]. Our results evidence this claim, revealing that human observers prefer AI-generated artworks over stylistically similar artworks painted by real people. This paradigm shift in art appreciation, favouring synthetic works over those created by human artists, has the potential to revolutionize the art world, while also raising new questions about authorship, authenticity, and the role of human creativity in the age of generative AI.

Our findings stand in contrast to prior research on subjective evaluations of computer-generated artwork, which have largely reported a negative bias towards AI art [8–12]. This work has primarily examined the role of authorship attribution in AI art perception, rather than the aesthetic value of the artworks themselves. Thus, the observed negative bias in these studies appears to relate to our explicit prejudice against artificially-generated content (i.e., if an artwork is labelled as computer-generated, we tend not to like it). In contrast, here we obtained observer preference decisions in the absence of any authorship label – a neutral presentation format that encourages observers to judge the inherent aesthetic qualities of the artworks – and assessed authorship discrimination in a separate experiment. This approach allowed us to obtain a quantitative assessment of the current art-generating capabilities of AI image-generation models, free from external biases.

Although observers in the first experiment consistently preferred artworks generated by DALL·E 2 over those made by human artists, it was not the case that these AI artworks were *indistinguishable* from human creations. In Experiment 2, a separate group of observers were asked to explicitly judge which of the two artworks in each pair was generated by a computer. We found they could reliably do so above chance-level. Moreover, there was a positive correlation between the image-pairs' AI-preference and AI-detection scores, suggesting that the same visual features that made the AI-generated artworks more detectable to participants in Experiment 2 also made those artworks more appealing to participants in Experiment 1 [13]. This intriguing pattern underscores the role that explicit bias against artificial creations has likely played in prior investigations [8–12] of the aesthetic appeal of AI-generated artworks: When participants do not know the artworks are computer-generated, they freely prefer them. Interestingly, we found no evidence that these effects were moderated by observers' art expertise, suggesting that the features in question are broadly accessible; a possibility which future research will no doubt explore in detail.

More generally, these results suggest that GANs, the technology behind DALL·E 2, in striving for stronger verisimilitude in computer-generated art, have evolved to do so by extrapolating (or exploiting) existing known biases in human cognition. On this thinking, DALL·E's capacity to produce works that observers tend to prefer over human artworks can be explained by the fact that its training dataset comprises images of artworks that are broadly considered to be aesthetically pleasing. This is in line with recent research on 'deepfakes', wherein AI-generated faces not only fool observers with their hyper-realistic nature, but are also associated with enhanced perceptions of trustworthiness [2,14,15]. These findings raise critical concerns about the exact nature of the cognitive processes that could be targeted and manipulated using generative-AI, and therefore, about its large-scale deployment without detailed investigation.

In a world increasingly shaped by the algorithms around us, the current findings suggest that AI has not only caught up with human-generated art, but is redefining our understanding of creative expression altogether. If AI-generated content has reached or surpassed aesthetic equivalence with human creation, the question of whether something can truly be considered 'art' if it has no human architect becomes more complicated. Our results are an initial step towards untangling the complex interaction between generative AI and human aesthetic preference; clearly, systematic examinations of AI-generated artworks' features are needed to fully understand the mechanisms and implications of AI preferences. As the field of generative-AI continues to accelerate – spurring equal parts concern and excitement – there can be no doubt as to the urgency in this challenge. DALL·E 2 will soon be superseded by the next generation of algorithms with as-yet unknown capabilities. Understanding how the human experience intersects with this technology will be critical to ensuring its positive impact in our society.

Methods

Participants

Online participants from Western Sydney University were recruited via the university (SONA) participant management system in exchange for course credits. We recruited 127 participants in Experiment 1, including 31 males, 95 females, and 1 non-binary with mean age of 22.27 (SD = 5.89), and 137 participants in Experiment 2, including 26 males, 109 females, and 2 non-binary with a mean age of 21.76 (SD = 6.99). Our participants reported on average a medium level of expertise in art, with subjective ratings of interest in art of 63.32 (SD = 29.68) and 62.22 (SD = 28.12), knowledge of art history of 30.81 (SD = 25.41) and 28.41 (SD = 26.46), artistic personality of 53.32 (SD = 28.35) and 52.49 (SD = 29.00) on a scale of 0 to 100, in Experiment 1 and 2, respectively. All participants provided written informed consent prior to the study, which was approved by the Human Research Ethics Committee of Western Sydney University.

Stimuli

Stimuli in Experiments 1 and 2 were 50 images of real artworks and 50 images of AI-generated artworks representative of various artistic styles (impressionism, early expressionism, baroque and romanticism). Each image was presented at 200×200 pixels, which, assuming a standard laptop screen, corresponds to approximately 6×6 degrees visual angle (note this varies depending on the participant's own device). Real and synthetic artworks were matched across artistic styles to form 50 pairs of images shown in Figure 1.

The AI-generated stimuli were created with DALL·E 2, an image diffusion model that generates high-quality, complex images based on textual prompts input by the user (https://openai.com/dall-e-2). Briefly described, this process relies on a text encoding model (Contrastive Language-Image Pre-training; CLIP) to link textual input to visual output by use of a two-stage model involving a 'prior' image caption embedder and an 'encoder', which work in tandem to extract information relevant to the desired visual output [16]. After sufficient training, the CLIP model is frozen and the now-embedded semantic information it produced is used to train a diffusion 'decoder' that allows for the process to be inverted. DALL·E 2 employs a diffusion model named Guided Language to Image Diffusion for Generation and Editing, which after training, allows for text-conditional image generation. This is achieved by training a Markov chain to make certain inferences using a set of sample images, which are iteratively provided with more Gaussian noise until it is able to reverse the generation process [17]. This model is then trained using a generative adversarial network (GAN), where two networks, a generator and a discriminator, are locked in a zero-sum game and continually pushed to greater levels of image generation refinement [18]. The result is a highly accessible and versatile AI image-generation tool that can convert textual prompts into detailed realistic images.

DALL·E 2 was used with 36 unique prompts that included both an artist's name and the type of artwork (e.g., “Paul Cezanne style still life painting”, see Figure 1). DALL-E generates several images in response to each prompt. We selected a single image per prompt and cropped the image to remove the DALL·E logo in the bottom-right corner. Several prompts were used multiple times (e.g., Claude Monet style garden painting). To minimize bias in the selection process, the generated images were manually compared to famous artists’ works found on Wikimedia Commons (https://commons.wikimedia.org) and WikiArt (https://www.wikiart.org), focusing on comparable visual features (e.g., colour, style, composition). The experiment ran online in participants' web browsers [19], was coded using the javascript framework jsPsych, version 7.3 [20], and ran on Pavlovia [21].

Procedure

Experiments 1 and 2 had the same experimental procedure and design, differing only in terms of participant instruction. At the start of the experiment, participants reported their demographic information, along with their art expertise, operationalised through three questions: (1) “*Rate your interest in art*”, (2), “*Rate your knowledge of art history*”, and (3) “*How artistic are you?*”. Participants indicated their response using a slider coded to a value between 0 and 100. Next, in the main part of the experiment, the 50 pairs of artworks shown in Figure 1 were presented once in a random order with the human-made and AI-generated stimuli being randomly presented either on the left or right side. Participants in Experiment 1 were not aware of the true aim of the study. They were not informed of the origin of the artworks and were simply instructed to select which one of the two images they preferred (see Figure 2A). Participants in Experiment 2 were told that one in each pair was AI-generated and instructed to click on it (see Figure 2A). Fifty additional trials with unique 50 pairs randomly drawn from the 100 (human and AI) artworks were included in each experiment (but not analysed) to ensure that participants remained naïve to the experimental manipulation. The 100 trials in total were performed by participants at a self-selected pace. Each pair of images remained onscreen until a selection was made. The total duration of the experiment was about 5 minutes. Participants could only participate in either Experiment 1 or Experiment 2 to ensure participants in Experiment 1 were not aware of the presence of AI-generated stimuli.

Data and statistical analysis

AI preference scores from Experiment 1 and AI detection accuracy scores from Experiment 2 of each participant were averaged across the 50 pairs of stimuli and then submitted to one-sample t-tests to examine deviations from the 50% chance level. The scores for each pair of images were also averaged across all participants within each experiment separately to test the image-wise correlation between the two experiments using Pearson correlations. A principal component analysis was conducted on the three expertise scores and data on the first dimension were used to test the effect of expertise on AI preference and AI detection accuracy in Experiment 1 and 2 using Pearson correlations.

Data availability

Stimuli and data used in this article are publicly available on the open science framework: https://osf.io/n7w32/

Acknowledgment

This research was supported by the Australian Research Council (DP220103047, DE230100380).

References

1. Moshel, M. L., Robinson, A. K., Carlson, T. A. & Grootswagers, T. Are you for real? Decoding realistic AI-generated faces from neural activity. *Vision Res.* **199**, 108079 (2022).
2. Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci.* **119**, (2022).
3. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **9**, (2019).
4. Zhou, K.-Q. & Nabus, H. The Ethical Implications of DALL-E: Opportunities and Challenges. *Mesopotamian J. Comput. Sci.* **2023**, 17–23 (2023).
5. Epstein, Z. *et al.* Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
6. Pan, Z. *et al.* Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access* **7**, 36322–36333 (2019).
7. Chamberlain, R., Mullin, C., Scheerlinck, B. & Wagemans, J. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychol. Aesthet. Creat. Arts* **12**, 177–192 (2018).
8. Agudo, U., Arrese, M., Liberal, K. G. & Matute, H. Assessing Emotion and Sensitivity of AI Artwork. *Front. Psychol.* **13**, 879088 (2022).

9. Caporusso, N. *et al.* User Discrimination of Content Produced by Generative Adversarial Networks. in *Human Interaction and Emerging Technologies* (eds. Ahram, T., Taiar, R., Colson, S. & Choplin, A.) vol. 1018 725–730 (Springer International Publishing, 2020).
10. Chiarella, S. G. *et al.* Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Comput. Hum. Behav.* **137**, 107406 (2022).
11. Gangadharbatla, H. The Role of AI Attribution Knowledge in the Evaluation of Artwork. *Empir. Stud. Arts* **40**, 125–142 (2022).
12. Kirk, U., Skov, M., Hulme, O., Christensen, M. S. & Zeki, S. Modulation of aesthetic value by semantic context: An fMRI study. *Neuroimage* **44**, 1125–1132 (2009).
13. Van Geert, E. & Wagemans, J. Order, complexity, and aesthetic appreciation. *Psychol. Aesthet. Creat. Arts* **14**, 135 (2020).
14. Boháček, M. & Farid, H. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proc. Natl. Acad. Sci.* **119**, e2216035119 (2022).
15. Groh, M., Epstein, Z., Firestone, C. & Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci.* **119**, e2110013119 (2022).
16. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. *ArXiv Prepr. ArXiv220406125* **1**, 3 (2022).
17. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
18. Poltronieri, F. A. & Hänska, M. Technical Images and Visual Art in the Era of Artificial Intelligence: From GOFAI to GANs. in *Proceedings of the 9th International Conference on Digital and Interactive Arts* 1–8 (ACM, 2019). doi:10.1145/3359852.3359865.

19. Grootswagers, T. A primer on running human behavioural experiments online. *Behav. Res. Methods* **52**, 2283–2286 (2020).

20. De Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* **47**, 1–12 (2015).

21. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).