AMERICAN PSYCHOLOGICAL ASSOCIATION
DIVISION 10
Psychology of Aesthetics, Creativity, and the Arts

ISSN: 1931-3896
https://doi.org/10.1037/aca0000570

Artificial Intelligence and Art: Identifying the Aesthetic Judgment Factors That Distinguish Human- and Machine-Generated Artwork

Andrew Samo and Scott Highhouse
Department of Psychology, Bowling Green State University

Artistic creation has traditionally been thought to be a uniquely human ability. Recent advancements in artificial intelligence (AI), however, have enabled algorithms to create art that is nearly indistinguishable from human artwork. Existing research suggests that people have a bias against AI artwork but cannot accurately identify it in blind comparisons. The current study extends this investigation to examine the aesthetic judgment factors differentiating human and machine art. Results indicate that people are unable to accurately identify artwork source but prefer human art and experience more positive emotions in response to human artwork. The aesthetic judgment factors differentiating human- and machine-generated art were all related to positive emotionality. This finding has several implications for this research area and limitation and avenues for future research are discussed.

Keywords: aesthetics, aesthetic judgment, computational creativity, artificial intelligence art

The originality, creativity, and emotions that are involved in art are often thought of as uniquely human (Coeckelbergh, 2017; Collingwood, 1938; Ramachandran & Hirstein, 1999) because machines cannot yet experience originality, creativity, or emotions as humans do (Boden, 2007; Llano et al., 2022; Mazzone & Elgammal, 2019; McCormack et al., 2019; Ren & Bao, 2020). In a world increasingly run by software, these creative pursuits were thought to be safe from automation (Bakhshi et al., 2015; Gunkel, 2021). Recent advances in artificial intelligence (AI), however, have allowed algorithms to generate images, music, and prose that are indistinguishable from human made artwork (Carnovalini & Rodà, 2020; Köbis & Mossink, 2021; Ramesh et al., 2021, 2022; Saharia et al., 2022; cf., Raji et al., 2021). These algorithms have been criticized for being “stochastic parrots” (Bender et al., 2021, p. 614) creating “zombie art” (Hassine & Neeman, 2019, p. 29), but they are also challenging the idea that art is solely a human enterprise (Gunkel, 2017).

Little is known about how people perceive and appreciate machine-generated artwork because this is an emerging area of research based around new technologies creating new forms of human–computer interaction (Cetinic & She, 2021; Chamberlain et al., 2018; Ren & Bao, 2020). So far, empirical work suggests that when presented with artwork that has not been identified as human- or machine-generated, people seem to prefer the artwork created by humans—and this effect persists even when people cannot identify the art or when they are told that the human art was made by machines (Chamberlain et al., 2018; Elgammal et al., 2017; Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020). This is a surprising finding because it means that people may *feel* differently about machine-generated art without *knowing* what the difference is. The purpose of the current research was to identify, using discriminant function analysis, which aesthetic judgment factors differentiate human- and machine-generated artwork. All stimuli, data, and syntax are available on the Open Science Framework (OSF; osf.io/z4cnj/).

Human Aesthetic Judgment

Human aesthetic judgment involves the psychological processes of perceiving, judging, and evaluating an aesthetic stimuli—a work of art, a building’s architecture, or sunsets (Leder et al., 2004; Lindell & Mueller, 2011; Pelowski et al., 2016).1 The mechanisms underlying our appreciation of aesthetic experiences are studied by empirical aesthetics. The foundations for empirical aesthetics were set by Gustav Fechner in the 19th century. The field has grown to understand empirical aesthetics as the psychological study of the mechanisms allowing humans to experience and appreciate a variety of phenomena (i.e., objects, design, nature, people, etc.) aesthetically (i.e., beautiful, ugly, nostalgic, exciting, etc.) (Leder & Nadal, 2014). This empirical work has established that there are three broad factors influencing aesthetic judgments, including objective, personal, and contextual factors (Nadal & Vartanian, 2021). Objective factors reflect the statistical properties of art (i.e., spacing, symmetry, color, complexity; see

Andrew Samo https://orcid.org/0000-0003-2225-184X
There are no conflicts of interest regarding the authorship or production of this article.
All experimental stimuli, data, and syntax are available on the Open Science Framework: https://osf.io/z4cnj/.
Correspondence concerning this article should be addressed to Andrew Samo, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403, United States. Email: asamo@bgsu.edu

[1] Here, it is worth noting that a work of art and an aesthetic are distinct phenomena, such that all art is aesthetic, but not everything aesthetic is art (see Skov & Nadal, 2020). Therefore, the distinction between making and appreciating can be reframed as a distinction between making art and appreciating aesthetic experiences.

Graham & Redies, 2010). Personal factors include individual differences in personality, ability, knowledge that influence perceptions, emotions, and understanding of the objective factors. Context factors capture the current discourse and history around an artwork (Jacobsen, 2010; Lindell & Mueller, 2011; Palmer et al., 2013; Pearce et al., 2016). The way these factors interact have been formalized in process models of aesthetic judgment.

Process models of aesthetic appreciation and judgment provide integrative descriptions of the processes involved in the judgment and appreciation of aesthetic objects—including art (Leder et al., 2012). The models are integrative because they frame the emotion, meaning, and evaluative outcomes of aesthetic judgment as the result of person–situation interactions, or interactions between the personal, object, and contextual factors (Pearce et al., 2016). One of the most recent models of aesthetic judgment, The Vienna Integrated Model of top-down and bottom-up processes in Art Perception (VIMAP; Pelowski et al., 2017), builds on earlier models (e.g., Leder et al., 2004; Leder & Nadal, 2014; Pelowski et al., 2016) to propose that there are a series of stages and processing checks through which people form aesthetic judgments and experiences. The stages involve sequential perceptual, affective, and cognitive processes that integrate the person, object, and contextual features of empirical aesthetics. The processing checks include schema congruency and self-relevance checks. Schema congruency captures a process where the audience ensures they have reached a proper understanding of the artwork by checking their understanding against existing schemas and the fluency of their judgment. Self-relevance captures a process where the audience ensures that the content, meaning, or context of the artwork is personally relevant to them.

The processing stages and checks dynamically interact to result in five unique outcomes, ranging from little emotion felt to intense positive (i.e., flow, chills, being moved), negative (i.e., boredom, confusion, disgust), and sublimation into transformative experiences (i.e., awe) (see Pelowski et al., 2017 for more). The experiences at each of these processing stages have been operationalized in psychometrically validated measures of aesthetic judgment factors, like the ARS (Hager et al., 2012) and the Aesthetic Emotions Scale (AESTHEMOS; Schindler et al., 2017). This has resulted in scales capturing the affective, cognitive, and semantic factors influencing aesthetic judgments. Notably, these aesthetic judgment factors have not yet been examined in relation to machine-generated artwork. The focus of the current study is on these aesthetic judgment factors because they highlight how people may differentially process human- and machine-generated artworks.

Machine-Generated Art

The best-in-class algorithms can generate original or edit existing images, music, poetry, prose, and code at high levels of realism and accuracy based on text prompts from human users. For example, OpenAI's DALL-E 2 is an "AI system that can create realistic images and art from a description in natural language" (https://openai.com/product/dall-e-2) and Google's Imagen is an algorithm with "an unprecedented degree of photorealism and a deep level of language understanding" (https://imagen.research.google/). With these models, users can enter virtually any prompt to receive an assortment of novel images, including prompts like the ones showcased on DALL-E 2's website, "an astronaut playing basketball with cats in space in a watercolor style," or Imagen's website, "a single beam of light enters the room from the ceiling illuminating an easel with a Rembrandt painting of a raccoon."2 At the time of writing, these text-to-image models require some degree of human input to begin the image generation process and some degree of human decision-making to fine-tune and select the final image (see Frolov et al., 2021 for a review of text-to-image models).

On a technical level, the art is generated with neural networks (e.g., Ramesh et al., 2021, 2022; Saharia et al., 2022) which are machine learning algorithms based on models of the human brain (James et al., 2021). Brains are composed of billions of interconnected neurons sending signals to adjacent neurons, forming networks, and giving rise to higher cognition (Fusi et al., 2016; Yuste, 2015). Similarly, artificial neural networks use input (i.e., dendrites), processing (i.e., cell body), and output (i.e., axon terminal) mechanisms to learn and perform (James et al., 2021). The information processed by these networks are vectors of embeddings, which are lists of numerical values representing features of data. Here, the embeddings quantify the text captions or prompts (i.e., text embeddings) and the related images (i.e., image embeddings). To oversimplify generating images from text, encoder models create text embeddings from text prompts and decoder models create images from the text embeddings. A seminal text-to-image algorithm, the Generative Adversarial Network (GAN; Goodfellow et al., 2014; Hudson & Zitnick, 2021), used competing networks, a generator and discriminator, to generate sample images from the latent space and discriminant against low quality images by playing a minmax game until an image of passable quality is generated (Hudson & Zitnick, 2021). Newer algorithms use foundation models, with transformer architecture, to improve flexibility and performance (Bommasani et al., 2021). Transformers use attention mechanisms to tag and track information as it moves across the network nodes to create a deep sense of how features relate to each other (Vaswani et al., 2017).

At the time of writing, some of the best-in-class text-to-image algorithms use similar architectures with encoders, decoders, and transformers—although specific implementations may vary (see Crowson et al., 2022). In practice, these algorithms build on GANs by implementing models that work together to iteratively generate images based on text prompts (e.g., Vector Quantized Generative Adversarial Network [VQGAN]; Esser et al., 2021), evaluate the generated images against the prompt (e.g., Contrastive Language-Image Pretraining [CLIP]; Radford et al., 2021), and refine the image based on the evaluation (e.g., VQGAN-CLIP; Crowson et al., 2022). These models require large amounts of data to train and are computationally intensive. Fortunately, training data have become readily available as museums are continually digitizing their collections to improve the accessibility of their art to the public (Cetinic & She, 2021). While the authors did not train the models, we ran the models and rented virtual graphics processing units to reach the computational power required to generate the study images. Overall, machine-generated art transforms human text inputs into original images through complex transformer-based models.

The major criticisms of this approach to machine-generated art involves the degree of human input from text prompts and the lack of originality from using existing artwork as training data. More specifically, critics claim that these algorithms are lacking authorship and

[2] For more fun examples of "realistic-looking fake versions of almost anything" see https://thisxdoesnotexist.com/.

intention because the algorithms are simply responding to human instruction through the text prompts. They also claim that the algorithms are lacking originality and novelty because the models are trained on datasets of existing human images (Boden, 2007, 2010; Coeckelbergh, 2017; Hertzmann, 2018; McCormack et al., 2019). In other words, art-generating models are "stochastic parrots" (Bender et al., 2021) using pseudorandom walks across hidden layers to create derivative, emotionless "zombie art" (Hassine & Neeman, 2019). On the other hand, creating art by sampling from an existing space with the systematic use of chance has a long history (Dorin, 2013; Todorov, 2019). In 1961, Raymond Queneau published a book *Cent Mille Milliards de Poems* which outlined the content and rules for readers to generate "one hundred thousand million" original poems. More recently, the action art of Jackson Pollock, the conceptual art of Marina Abramovic, and other installation, land, or performance arts emphasize chance and randomness in their artistic process. These examples showcase art as a process, rather than an outcome, taking advantage of stochastic systems to sample the existing space. Modern art-generating algorithms also stochastically sample the existing artworld, using representation learning, to generate new works of art (Cetinic & She, 2021).

It is difficult to determine if these algorithms are *really* creating art, even if the algorithms were not "black box" and uninterpretable (Arrieta et al., 2019), because there are no widely accepted definitions of art (Hertzmann, 2018) or creativity (Amabile & Pillemer, 2012) to use as benchmarks to compare them against. Instead, recent work in computational creativity has been emphasizing the *appearance* of creativity over the existence of *real* creativity (Coeckelbergh, 2017; Colton & Wiggins, 2012; Natale & Henrickson, 2022). The Lovelace Test, popularized as Lady Lovelace's objection by Alan Turing (1950; see Abramson, 2008), is a behavioral test for computational creativity that would require an unbiased audience to believe machine-generated art is human-generated—regardless of how it was made (Bringsjord et al., 2001; Riedl, 2014). The Lovelace Effect describes situations where "the behavior of computing systems is *perceived by users* as original and creative" (Natale & Henrickson, 2022, p. 2). At this point, art-generating algorithms are still demonstrating indeterminacy (i.e., appearing coherent but lacking spatial narrative; Hertzmann, 2020), although transformer-based architectures may solve for this. For example, large scale foundation models like Google's Imagen and OpenAI's DALL-E 2 claim (and appear, to the authors) to be highly realistic (Ramesh et al., 2022; Saharia et al., 2022). The focus on machines *being* creative is an interdisciplinary cognitive and computer science problem where algorithms are design for specific operationalizations of creativity (Larson, 2021). On the other hand, the focus on machines *appearing* creative is a problem of human aesthetic judgment. This is the focus of the current article.

Previous Work on the Human Aesthetic Judgment of Machine-Generated Art

There have been relatively few empirical studies examining the human perceptions of machine-generated artwork. On one hand, it is surprising that there have only been a handful of empirical studies examining the human evaluation of machine-generated art, considering the growing accessibility and excitement around the algorithms. For example, there is work around developing and combining models (Crowson et al., 2022) and work around developing prompts for the models (i.e., prompt engineering; Liu & Chilton, 2022). On the other hand, people are still focused on exploring how to make art with the models and interest may not have trickled down to exploring how we appreciate the art. Across the five empirical studies, there are several broad patterns of findings: people cannot accurately identify classify images as human- or machine-generated, people have more positive aesthetic experiences with human-generated art, and that the factors driving the positive experiences are still relatively unknown (Chamberlain et al., 2018; Elgammal et al., 2017; Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020).

First, these articles have largely found that people have trouble correctly identifying and classifying the images as human- or machine-generated. Across these studies, classification accuracies range from 42% to 64% or about as good as a coin toss (Chamberlain et al., 2018; Gangadharbatla, 2022; Ragot et al., 2020). Ragot and colleagues (2020) found an overall accuracy of 61% (66% and 56% for human- and machine-generated, respectively) with slightly higher classification accuracies for portraits (69%) and lower for landscapes (53%) compared to the overall score. This may be due to a difficulty algorithms have generating faces and realistic figures (Hertzmann, 2020) because the distorted figures may activate certain schemas, biases, or otherwise clue participants in. Along this line, it has also been suggested that the lower classification accuracies for machine-generated art may be driven by the expectation and belief that high-quality art (i.e., the art participants like) could only be made by humans and, as such, simply appreciating the art biased subsequent classifications toward human artists (Chamberlain et al., 2018; Gangadharbatla, 2022). They attributed this to outdated schemas and stereotypes emerging from "the notoriety of abstract and geometric computer art created during the 1960s" (Chamberlain et al., 2018, p. 182).

Second, these articles have found that people still have a preference for human-generated art (Chamberlain et al., 2018; Gangadharbatla, 2022; Hong, 2018; Hong & Curran, 2019; Ragot et al., 2020) and that we have a bias against machine-generated art (Elgammal et al., 2017; Hong & Curran, 2019; Ragot et al., 2020). Notably, several studies have suggested that preference and classification are related because people are more likely to classify images they prefer as human-generated because of schemas, or stereotypes, that typical machine-generated art is still the fractal, pixelated, neon-colored computer art of the 1980's (Chamberlain et al., 2018; Gangadharbatla, 2022). However, there have been some inconsistent findings. For example, Hong and Curran (2019) found that labeling an image as machine-generated can activate existing algorithm averse biases, Ragot et al. (2020) found a biasing effect regardless of existing bias and others reported little to no evidence of bias (Elgammal et al., 2017; Gangadharbatla, 2022). Chamberlain et al. (2018) found that anthropomorphizing the algorithm's artistic process reduced bias (i.e., embodying it with robot arms, paper, and paintbrush) and suggested that our aversion to machine-generated art may involve perceptions of effort (i.e., an effort heuristic; Kruger et al., 2004). Most recently, Gangadharbatla (2022) suggested that people dislike when advanced algorithms generate high-quality, representational, realistic art because it disrupts expectations that machine-generated art is abstract, indeterminate, geometric, and pixelated (Chamberlain et al., 2018; Hertzman, 2020).

Although the previous work has largely been focused on questions of classification and preference, these articles have also found that human-generated artworks typically receive more favorable aesthetic ratings, yet results remain relatively inconclusive. For example, Chamberlain et al. (2018) found that human-generated images had ratings of liking, even after controlling for expertise and statistical image properties. Gangadharbatla (2022) found an interaction effect, based on beliefs around art, such that human-generated art received higher ratings of evaluation, liking, and purchase intentions when the art was representational. They also found that machine-generated art received higher ratings when the images were abstract. Both Hong and Curran (2019) and Ragot et al. (2020) found that human-generated images received higher ratings of composition, emotional expression, novelty, meaning, and overall aesthetic appreciation—even when the labels were switched (i.e., human-generated images were labelled as machine-generated). Much of the empirical work that has been conducted in this space, however, has used psychometrically suspect measures that were not developed with a process model of empirical aesthetics as a theoretical reference point.

The Current Project

Machine-generated art is a novel area based on new, large-scale deep learning models (Bommasani et al., 2021; Cetinic & She, 2021). These foundation models are already causing ethical (Pistilli, 2022; Weidinger et al., 2021), social (Groh et al., 2019; Paullada et al., 2021), and legal controversy (Dee, 2018; Rahwan et al., 2019). They are also challenging our long-held ideas around the importance of art, creativity, and emotional expression in humans (Gunkel, 2021). As such, understanding how people perceive, interpret, and appreciate machine-generated art is an increasingly important area of research (Cetinic & She, 2021; Colton & Wiggins, 2012; Natale & Henrickson, 2022).

Using the process models of aesthetic judgment, this study examines which aesthetic judgment factors in the model differentiate human- and machine-generated artwork. The current study builds on earlier work in several ways. First, we grounded our study in the aesthetic judgment literature, because the perception of art—human- or machine-generated—is a question of aesthetic judgment more so than computational creativity (which may be more closely aligned to artistic creation; Cetinic & She, 2021). In doing so, we used process-based models of aesthetic judgment to inform our design and, particularly, the selection of dependent measures, the aesthetic judgment factors, that were used. More specifically, process models informed the range of cognitive, semantic, and emotional aesthetic judgment factors included, which may be a more comprehensive sampling of the factors that differentiate human- and machine-generated art.

Second, due to the importance of image stimuli (Specker et al., 2020), we used some of the new developments in generative art models (e.g., Crowson et al., 2022) to develop our image stimuli and used a stimulus sampling approach (Highhouse, 2009) to standardize and select our final study stimuli. Third, we designed our study in such a way as to avoid biasing raters against machine-generated art, as found in previous work (e.g., Chamberlain et al., 2018; Hong & Curran, 2019), by only asking participants to consider the possibility of the art being machine-generated after the ratings occurred. We also used an analytical method specifically developed for describing between-group differences for sets of factors (i.e., descriptive discriminant analysis [DDA]; Smith et al., 2020). If people process the human- and machine-generated artwork differently, these processing differences should be reflected in the aesthetic judgment factors. Thus:

> *Research Question 1*: What are the aesthetic judgment factors differentiating between human- and machine-generated art?

Furthermore, this effort also aims to conceptually replicate earlier work, which has found that people have an inconsistent preference for human generated artwork (Chamberlain et al., 2018; Gangadharbatla, 2022) and that they can correctly classify the art source (human or machine) around 42.5%–63.8% of the time (Chamberlain et al., 2018; Elgammal et al., 2017; Gangadharbatla, 2022; Ragot et al., 2020). However, it is worth revisiting these results around preference and classification for several reasons. First, the recent developments in transformer-based architectures (i.e., VQGAN-CLIP, diffusion models) may not have been available during earlier research. The transformer-based models have a distinct style and are thought to be more realistic than pretransformer models. Second, another notable difference in our design is that participants were not informed of the possibility that their image was generated by AI, thereby eliminating any potential bias in ratings of preference (Hong & Curran, 2019). Third, the stimuli and measures used in the current project have been validated. The pilot study used stimulus sampling to identify standardized image stimuli (i.e., balanced on attractiveness, technical skill, and familiarity) and the main study uses psychometrically valid measures of aesthetic judgment. As such, this is a conceptual replication because the phenomena under investigation are similar but the design, sample, stimuli, and measures are different (Derksen & Morawski, 2022). More specifically, with these newer algorithms and more rigorous stimulus sampling and measurement, we are revisiting whether people have a preference for human- or machine-generated art and whether or not they are able to accurately distinguish which image is from which source:

> *Research Question 2*: Do people have a preference for human- or machine-generated art?

> *Research Question 3*: Can people distinguish between human- and machine-generated art?

Each of these research questions can be explored in fairly simple ways, following previous work. For example, preference can be captured simply by asking people how much they liked the image they were given (Chamberlain et al., 2018; Ragot & Martin, 2020). While aesthetic liking is one indicator of preference, another more behaviorally based indicator is a person's willingness to want to hang the image in their home or office. Image classification can be captured by introducing people to the idea that images are being generated by machines and asking whether they think their image was generated by a human or machine artist. Therefore, the current project explores research questions around classification, preference, and the aesthetic judgment factors differentiating human- and machine-generated art. Overall, these efforts taken to ground the current design within the theoretical framework of process models of aesthetic judgment, to use new transformer-based text-to-image algorithms, and the use of well validated image stimuli and measures should improve the validity and generalizability of our results (Fabrigar et al., 2020; Yarkoni, 2022).

Method

Participants

Participants were recruited from MTurk through CloudResearch. Participants were required to be at least 18 years old, based in North America, have intermediate English ability, and have an approval rating of 70%+ with 1,000+ approved HITs. The survey was programmed on Qualtrics. To ensure data quality, there were survey requirements (above), filters (i.e., reCAPTCHA, age, regional emergency phone number identification, and language checks), and instructed attention checks. Participants who did not meet requirements were unable to launch the survey. Participants who failed the filters were sent to the end of the survey. Participant who failed two of three attention checks were filtered out of the dataset during data cleaning (below). The sample size was determined with an *a priori* power analysis, using G*Power (V.3.1.9.2; Faul et al., 2007), which indicated that the minimum sample required to detect a small to medium effect ($f^2 = .10$) with power = .95 for a multivariate analysis of variance(MANOVA) analysis was $n = 132$. A relatively small effect size was used to conservatively establish the minimum sample size to detect a baseline effect (e.g., Hwang et al., 2002) and a MANOVA F test was specified because it is mathematically similar to discriminant analysis (Smith et al., 2020). Out of 200 initial respondents, the final sample was $N = 190$ with an average age of 39 ($SD = 10$ years), 73% white, 58% male, and 55% having a Bachelor's, Master's, or PhD. They also had an average level of self-reported artistic interest ($M = 3.12$, $SD = 1.09$). Participation was voluntary and participants were compensated 1.50 USD (7.50 USD hourly rate).

Procedure

Participants self-selected into the study and were automatically randomized into an "AI" or "Human" condition where they were shown only *one* image and asked to complete all of the measures based on that single image. In the AI condition, participants were randomly shown one of the three machine-generated images. In the Human condition, participants were randomly shown one of the three human-generated images. Participants were randomly assigned to conditions and randomization was balanced to make sure there was an equal number of participants in each condition and for each image. Notably, there was no mention of AI art and participants were not informed that some of the images were machine-generated until the final section of the survey when participants were asked to identify the images as human- or machine-generated.

The procedure was the same for both conditions. First, each participant read and completed a consent form. Next, they were shown their image and instructed to click on the picture to confirm that they had seen it. Afterward, the image was kept at the top of each survey page, for reference, and the participants were asked to rate the image on measures of the aesthetic judgment factors. Specifically, they were asked to, "please take a look at the following painting. This survey will ask you a series of questions on how you think and feel about this particular image. *For your reference, it will stay at the top of each page!*." The aesthetic judgment factor surveys and items were randomized across Participants to avoid order effects (see "Measures" section). After rating the respective images on the aesthetic factors, participants were asked to identify the source of the image (i.e., human- or machine-generated), how confident they were in their estimate, and complete a measure of their artistic interest. At the end of the study, participants were told which condition (AI or Human) they were in and what the source of their image was.

Materials and Measures

Image Materials

The image stimuli used in this study were identified with a stimulus sampling pilot study. Stimulus sampling involves ensuring that our image stimuli are high quality representations of their respective domains with minimal noise (Highhouse, 2009; Judd et al., 2012; Yarkoni, 2022). Image selection is important because the image features determine aesthetic judgments (Dalege et al., 2016; Pelowski et al., 2017)—particularly as our images are presented without context (Krauss et al., 2021; Mullennix et al., 2020). This image sampling process should increase the validity and generalizability of our subsequent results (Fabrigar et al., 2020; Flake et al., 2022; Shadish et al., 2002). The images stimuli were developed in three stages. For a more detailed description of the pilot study, see the OSF.

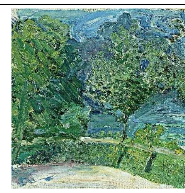
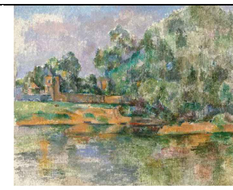
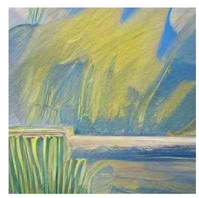
First, the human-generated images were identified through open-source online museum collections (i.e., Art Institute of Chicago, The Metropolitan Museum of Art, the National Gallery of Art, Artvee). The machine-generated images were created using five different algorithms (i.e., Ryan Murdock's Aleph2Image [A2I] and The Big Sleep, Katherine Crowson's Diffusion Model [DD; Crowson et al., 2022], Justin Bennington's S2ML, and DALL-E mini [Dem] hosted on HuggingFace) with text prompts based on the titles of the human images (i.e., "moonlight on the beach" and "lakeshore with reeds"). To reduce the total set of images to 20 (10 human and 10 machine) for the pilot study, the authors used a priori decision rules to standardize the image selection process. Finally, the 20 images were formatted for similar sizing (i.e., 512 px along the smallest dimension).

For the pilot study, 100 Mturk participants were recruited through CloudResearch to rate 20 images (10 human- and 10 machine-generated) on attractiveness, technical skill, familiarity, and willingness to hang the image in their home with single items (i.e., "this painting is attractive [i.e., aesthetically pleasing]," "this painting shows talent [i.e., technical skill]," "I feel like I have seen this painting before," and "I would hang this painting in my home"). Single items have been shown to reduce participant burden and maintain validity (Allen et al., 2022; Matthews et al., 2022). The final sample size was 89 after data cleaning. Overall, three human- and three machine-generated images, that were balanced for attractiveness and technical skill with low ratings of familiarity, were selected for the study (see Figure 1). Three images were used for each condition to avoid idiosyncratic image effects.

Measures

Aesthetic Judgment. The aesthetic judgment process was captured with the Art Reception Survey (ARS; Hager et al., 2012). The ARS is a measure of aesthetic judgment and appreciation that was developed based on process models of aesthetic judgment (e.g., Leder et al., 2004; Leder & Nadal, 2014). It has six dimensions with four to five items per dimension rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). The dimensions are cognitive stimulation ("intellectual engagement of the viewer with an artwork"), negative emotionality ("the arousal of unpleasant affective responses towards the artwork"), expertise ("the extent of

Figure 1
Final Images Used

Image Type	Images and Shorthand Titles		
Human			
	Poplars	Lakeside Road	Banks of Medan
AI			
	A2I lakeshore	DEm lakeshore	DD paris

Note. For the human artwork, Poplars = Paul Klee's Landscape with Poplars (1929; sourced from Artvee, in the public domain). Lakeside Road = Richard Gerstl's Lakeside Road Near Gmunden (1907; sourced from Artvee, in the public domain). Banks of Medan = Paul Cézanne's Banks of the Siene at Médan (c. 1885/1890; sourced from the National Gallery of Art, in the public domain). For the AI artwork, A2I lakeshore = Ryan Murdock's Aleph2Image model (author prompt: "a lakeshore with flower reeds in the style of expressionism"). DEm lakeshore = DALL-E mini model (author prompt: "a lakeshore with reeds in the style of post-impressionism"). DD = Katherine Crawson's Disco Diffusion model (author prompt: "a landscape of Paris in the style of post-impressionism"). AI = artificial intelligence. See the online article for the color version of this figure.

explicit knowledge the person has about artist and painting, as well as a sense of understanding of the artist's intention or idea that was meant to be expressed"), self-reference ("describes whether the recipient is feeling a person connection to the painting, evoking past memories or emotions"), artistic quality ("the level of creativity and artistic skillfulness that is attributed to the painting and painter"), and positive attraction ("subsumes items describing a positive reception of the artwork"). Estimates of internal reliability are shown in Table 1. Example items include, "this painting makes me curious" and "personal memories of mine are linked to this painting" for cognitive stimulation and self-reference.

Aesthetic Emotions. Aesthetic emotions were captured with the AESTHEMOS (Schindler et al., 2017). The AESTHEMOS is a measure with 21 subdimensions that can be grouped into seven broad dimensions representing the broader emotional orientation of the subdimensions. The dimensions are negative emotions (with ugliness, confusion, anger, boredom, and uneasiness subdimensions), prototypical aesthetic emotions (with being moved, surprise, fascination, feeling of beauty/liking, and awe subdimensions), epistemic emotions (with intellectual challenge, interest, and insight subdimensions), animation (with energy, vitality, and enchantment subdimensions), nostalgia (with nostalgia and relaxation subdimensions), sadness (unidimensional), and amusement (with humor and joy subdimensions). Items were rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). Estimates of internal reliability are shown in Table 1. Example items include, "felt deeply

Table 1
Means, SD*s, and Cohen's* d *Effect Size for Aesthetic Judgment Variables by Image Type and Internal Reliability Estimates*

	Image type			
Variables	Human ($n = 91$), M (SD)	AI ($n = 97$), M (SD)	d	a
(ARS) Cog Stim	3.03 (1.18)	3.05 (1.08)	−.02	.91
(ARS) Neg Emo	1.45 (0.74)	1.44 (0.63)	.01	.84
(ARS) Expertise	2.10 (0.78)	2.05 (0.67)	.06	.64
(ARS) Self Reflect	2.20 (1.13)	1.88 (1.02)	.30	.90
(ARS) Art Quality	3.37 (1.06)	3.27 (0.96)	.10	.87
(ARS) Attraction	2.95 (1.02)	2.60 (0.97)	.35	.87
(AES) Neg Emo	1.64 (0.69)	1.70 (0.67)	−.10	.83
(AES) Aesthetic Emo	2.81 (0.96)	2.66 (0.89)	.16	.89
(AES) Epistemic Emo	2.79 (1.02)	2.70 (1.09)	.09	.81
(AES) Animation	2.44 (1.1)	2.27 (1.03)	.16	.92
(AES) Nostalgia	3.09 (1.12)	2.68 (1.06)	.38	.85
(AES) Sadness	1.81 (1.05)	1.66 (0.83)	.16	.73
(AES) Amusement	2.57 (0.99)	2.35 (0.87)	.24	.75

Note. ARS = Aesthetic Response Survey; AES = Aesthetic Emotions Scale; AI = artificial intelligence.

moved," "was unsettling to me," and "energized me" for prototypical aesthetic emotions, negative emotions, and animation, respectively.

Artistic Interest. Artistic interest was self-reported with the Vienna Art Interest & Art Knowledge Questionnaire (VAIAK; Specker et al., 2020). Artistic interest was measured across two scales, with seven items capturing self-reported interest rated on a 5-point Likert scale (1 = *not at all*, 5 = *very much*) and four behavioral items rated on a 5-point frequency scale (1 = *less than once per year*; 5 = *once per week or more often*). The reliability estimates were $\alpha = .92$ and .77, respectively. Example items included, "I am interested in art" and "I enjoy talking about art with others" for the self-report items and "how often do you visit museums or art galleries on average" and "how often do you view images of artworks (picture, internet, books, etc.)?" for the behavioral items. Notably, the VAIAK rating scale was changed from a 7- to 5-point rating scale. While it is recommended to keep a 7-point scale (Specker, 2021), we wanted to use a standardized 5-point scale across measures and reduce participant burden as much as possible and research suggests that there is often minimal impact to validity when moving from a 7- to 5-point scale (Dawes, 2008; Krosnick & Presser, 2010).

Image Classification. The classification of images as human- or machine-generated was captured with a single item: "some artists have started using computer algorithms (like artificial intelligence [AI] and machine learning) to make artwork. Do you think this image was made by a Human or a Machine (AI) artist?." There were two response options ("Human Artist" or "Machine [AI] Artist"). Classification confidence was also assessed with a sliding scale from 0% to 100% asking, "You think this painting/image was made by [insert previous selection]. How confident are you in that choice?." There was also an optional open-ended response box with the caption, "in a few words, how did you decide if this work was from a [insert previous selection]?"

Image Preference. Image preference was captured with two items that were used in the pilot study. Following previous work (Chamberlain et al., 2018; Ragot et al., 2020), perceptions of the images overall attractiveness was assessed with a single item, "this painting is attractive (i.e., aesthetically pleasing)," rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*). Additionally, a behavioral indicator of preference, willingness to hang the image in their home, was captured with, "I would hang this painting in my home," rated on a 5-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*).

Analysis

For data cleaning, three data quality checks were conducted following best practices outline by Curran (2016) and Niessen et al. (2016). First, ocular analysis indicated some cases with straight lining response tendencies so long strings (>2 *SD*) were flagged. Next, negative person-total correlations were flagged (Curran, 2016). Finally, cases with a significant distant Mahalanobi's Distance were flagged at cut-off of $p = .01$, which is stricter than the typical $p = .001$. Respondents with two flags or more were filtered out of the data set. This procedure resulted in dropping 10 cases.

For research question one, DDA (Smith et al., 2020), a variant on linear discriminant analysis (LDA; Fisher, 1936; see Boedeker & Kearns, 2019 for a discussion of LDA variants), was used. DDA is primarily used to describe group differences, based on sets of variables, by estimating the linear combinations maximizing between-group differences (Smith et al., 2020). It allowed us to descriptively explore the aesthetic judgment factors differentiating human- and machine-generated art. The analysis was conducted in RStudio V.4.1.2 with the *candisc* (Friendly & Fox, 2020) and *MASS* (Venables & Ripley, 2002) packages following best practices outlined in Smith et al. (2020). Research question two was investigated with descriptive analyses and research question three was investigated with a group means comparison with base R.

Results

Descriptive Statistics and Assumption Testing

The descriptive statistics, effect sizes, and reliability estimates for the aesthetic judgment factors are shown in Table 1 and more detailed descriptive statistics are available in supplemental tables found in the OSF. As shown there, several of the variables (i.e., negative emotion, self-reference, expertise, and sadness) demonstrated high skew and/or kurtosis which may be indicative of departures from normality. To further investigate multivariate normality, an assumption underlying DDA, we used *Q–Q* plots and a Shapiro–Wilks test. Both tests indicated that the data deviated from multivariate normality. To address this, data transformations were carried out (i.e., log, square root, inverse) based on the 95% confidence intervals around lambda values estimated with a Box–Cox test (Osborne, 2020). However, subsequent analyses were conducted with the raw data because, while the transformations improved the descriptive statistics, *Q–Q* plots and Shapiro–Wilks test indicated that the transformed data still violated the assumption of normality. A Box's *M* test indicated heterogeneity of covariance, but the test is sensitive to multivariate outliers and the ratio of log determinants was in a normal range (Smith et al., 2020). More information on our assumption testing, data transformation, and discussion around using the raw data is included in the OSF.

Aesthetic Judgment Factors

For research question one, exploring the aesthetic judgment factors differentiating human- and machine-generated art, the results are presented in Table 2. There were several variables contributing to the estimation of the linear discriminant function, including attraction ($B = 1.78$), cognitive stimulation ($B = -1.42$), animation ($B = -.55$), expertise ($B = -.44$), and negative emotionality ($B = 0.40$). These standard discriminant function coefficient weights represent how strongly the aesthetic judgment factors contributed to the estimation of the overall discriminant function. For example, attraction, looking at Table 2, had the greatest contribution with a standardized coefficient of 1.78, followed by cognitive stimulation with a coefficient of −.142, whereas aesthetic emotion had a negligible contribution at −.03.3

Before interpreting the discriminant function, it is important to consider the structure coefficients (i.e., r_s column in Table 2). Furthermore,

[3] This pattern of results, shown in Table 2, suggests that Cognitive Stimulation may be a suppressor variable. Further exploration and discussion are available in the OSF.

Table 2
Standardized Discriminant Function Coefficients, Structure Coefficients, and Group Centroids

Variable	Standard coefficients	r_s	r_s^2
(ARS) Cog Stim	−1.42	−0.04	0.00
(ARS) Neg Emo	−0.30	−0.05	0.00
(ARS) Expertise	−0.44	0.08	0.00
(ARS) Self Reflect	0.05	0.38	0.14
(ARS) Art Quality	−0.14	0.11	0.01
(ARS) Attraction	1.78	0.49	0.24
(AES) Neg Emo	0.40	−0.20	0.04
(AES) Aesthetic Emo	−0.03	0.26	0.07
(AES) Epistemic Emo	0.30	0.08	0.00
(AES) Animation	−0.55	0.25	0.06
(AES) Nostalgia	0.30	0.52	0.27
(AES) Sadness	0.21	0.14	0.02
(AES) Amusement	0.39	0.39	0.15
Group	**Centroids**	**Cohen's *d***	
AI	−0.38	0.79 [0.47–1.09]	
Human	0.40		

Note. Statistically meaningful ($r_s > 0.30$) aesthetic judgment factors are in bold. Standard coefficients are contributions to the discriminant functions equation; r_s = structure coefficients, are the bivariate correlations between observed variables and the discriminant function; r_s^2 = squared structure coefficients, are the proportion of explained variance. ARS = Aesthetic Response Survey; AES = Aesthetic Emotions Scale; AI = artificial intelligence.

four of the 13 variables included in the equation explained meaningful variance in the discriminant function (i.e., $r_s > .30$), including self-reflection ($r_s = .40$) and attraction ($r_s = .49$) from the ARS and nostalgia ($r_s = .51$) and amusement ($r_s = .34$) from the AESTHEMOS. The structure coefficients (r_s) represent the bivariate correlations between the aesthetic judgment factors and latent discriminant function, where larger coefficients indicate stronger associations between the aesthetic judgment factors and the discriminant function. Larger structure coefficients suggest that the associated variables contribute to the distinction between groups. Together, these four variables collectively explained 78% of the variance in the discriminant function distinguishing human- and machine-generated art. It is notable that these four significant factors are all emotionally based and characterized by positive valence at varying degrees of arousal (cf., Russell, 1980; van Tilburg et al., 2018).

As shown in Figure 2, the group centroid for the human condition (.40) is substantially higher than the machine condition (−.38). Interpreting this with the structure matrix (Table 2), human-generated art is differentiated from machine-generated art by ratings of self-reference ($r_s = .38$), attraction ($r_s = .49$), nostalgia ($r_s = .52$), and amusement ($r_s = .39$) such that the human art was rated higher on these aesthetic judgment factors compared to machine-generated art. In other words, participants reported experiencing more self-reflection, attraction, nostalgia, and amusement in the human-generated art condition.

Preference

For research question 2, whether people prefer human- or machine-generated art, we found a preference for human-generated art over machine-generated art. First, participants had more positive aesthetic experiences, on almost all of the aesthetic judgment factors, in the human-generated art condition, as shown by the discriminant function (see Figure 2). Second, participants also had significantly higher levels of attraction to human-generated images ($M = 3.74$) compared to machine-generated images ($M = 3.34$); $t(184.75) = -2.26$, $p < .05$; $d = .33$. Participants were also more likely to want to hang human-generated images ($M = 3.19$) in their homes compared to machine-generated images ($M = 2.70$); $t(183.53) = -2.32$, $p < .05$; $d = .34$.

Classification Accuracy

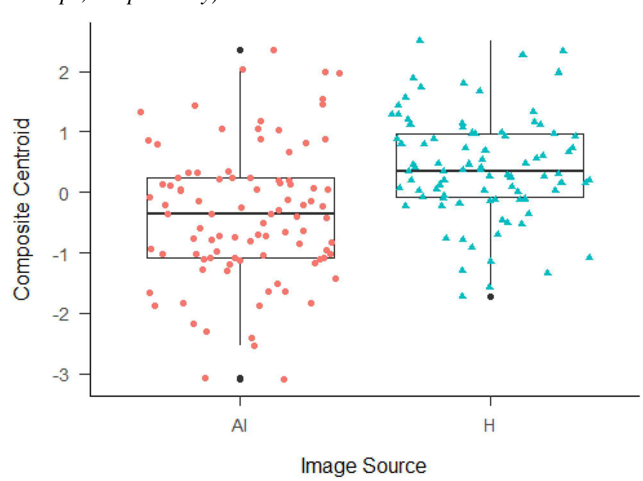
For research question 3, exploring whether people can distinguish between human- and machine-generated art, we found that the overall classification accuracy for the images was 60% (see Table 3). This means that participants were able to successfully classify the images as human- or machine-generated art 60% of the time. The classification accuracy was higher for human images at 85% and lower for machine images at 35%, which may reflect a schema or assumption that art is typically human made (Chamberlain et al., 2018).

Discussion

General Discussion

The purpose of this project was to explore the aesthetic judgment factors differentiating human- and machine-generated art. With recent advancements in AI and generative art, there has been as surge of ground-breaking research investigating how people experience machine-generated artwork. Most of this work has focused on preferences (Chamberlain et al., 2018; Elgammal et al., 2017), biases (Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020), and ability to identify and classify the image source (i.e., Turing or Lovelace Test; Chamberlain et al., 2018; Elgammal et al., 2017; Gangadharbatla, 2022; Ragot et al., 2020) with a smaller

Figure 2
*Image Source (AI or Human) Plot of Composite Centroids With 95% Confidence Intervals (*n = *100 and 94 for AI and Human Groups, Respectively)*



Note. AI = artificial intelligence. See the online article for the color version of this figure.

Table 3
Overall Classification Accuracy

Participant classification	True source: Machine generated	True source: Human generated	Classification accuracy (%)
Machine	34	13	35
Human	63	78	85
Total	97	91	60

Note. Participant classification reflects participant rating (i.e., "Do you think this image was made by a human or machine [AI] artist?); True Source reflects the experimental condition participants were randomly assigned to (i.e., the real image source). Classification accuracy was found by dividing the condition's participant classification by the true source total (e.g., 34/97 = 0.35). AI = artificial intelligence.

focus on the audience's subjective, aesthetic judgments of the artwork itself.

There are three primary contributions with this work. First, as a theoretical and methodological contribution, our framing of the study as *human perception of machine-generated art* allowed us to ground it in current theoretical models of aesthetic judgment and measure the aesthetic judgment factors with psychometrically valid measures. Second, our pilot study used a stimulus sampling procedure to validate our image stimuli and, hopefully, improve the ecological validity and generalizability of our results (Highhouse, 2009; Wells & Windschitl, 1999). The pilot study procedure and resulting images are available on the OSF and can be freely used in future research (https://osf.io/z4cnj/). Finally, this effort answers calls for more research around the human appreciation of machine-generated art (Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020), calls for continued research grounded in empirical aesthetics (Nadal & Vartanian, 2021), and calls to study AI from the human judgment perspective (i.e., The Lovelace Test; Gunkel, 2021; Natale & Henrickson, 2022; Ren & Bao, 2020). To our reading, this is one of the first studies to investigate human perceptions of machine-generated art with established theories and measures of aesthetic judgment. Overall, we found that aesthetic judgment factors characterizing positive emotions were higher for human-generated artwork—even when people had trouble identifying the art source.

Our findings indicated that positively valenced emotions collectively explained 78% of the variance in the function discriminating between human- and machine-generated art. In other words, it was the emotional, instead of cognitive or semantic, factors that differentiated. The specific factors were self-reflection (i.e., feeling a personal connection to the image), attraction (i.e., feeling a general positive attitude toward the image), nostalgia (i.e., feelings of relaxation, melancholy, or peace of mind), and amusement (i.e., feelings of joy and humor). Overall, these results suggest that people prefer and have more positive emotional experiences with human art, over machine art, even though they may not always identify which is which. Our findings are aligned with previous work suggesting that emotions are central to aesthetic processing (Leder et al., 2012; Menninghaus et al., 2019; Skov & Nadal, 2020). This preference for human-generated images is aligned with previous research but is particularly notable because participants in our study were not aware of which condition (human or AI) they were in before viewing and rating the images. They likely had the assumption that the images came from the same stereotypical source of art—humans—and were unbiased by schemas or stereotypes about artists (Hong & Curran, 2019; Ragot et al., 2020) when rating the images.

We also found that people have difficulty accurately identifying and classifying images as human- or machine-generated. More specifically, people were able to classify the images correctly 60% of the time—or slightly better than a coin toss. These results are in line with previous work that found overall accuracies ranging from 44% to 61%, human-generated image accuracies ranging from 60% to 84%, and machine-generated image accuracies ranging from 25% to 56%. The accuracy we found for machine-generated images (35%) was at the low end of the range established by previous research—only higher than the finding of 25% by Elgammal et al. (2017) when they introduced the Creative Adversarial Network model. Interestingly, successfully identifying machine-generated images as machine-generated 35% of the time means that the generative art algorithms were "tricking" participants into believing their work was made by humans 65% of the time. Based on these results, these algorithms would pass a Turing or Lady Lovelace Test as colloquially formed (i.e., greater than chance; French, 2000; Riedl, 2014).4

A Tentative Interpretation of the Results

Although this research was exploratory, we offer a tentative explanation for our findings based on aesthetic judgment fluency. Again, we found that, while people seem to prefer and experience more positive emotions with human-generated art, they cannot easily identify which image is human- and which image is machine-generated. In other words, people may *feel* better without *knowing* better. This pattern of results is consistent with a dual-processing account of judgment and decision making (Kahneman & Frederick, 2002; Stanovich, 1999), where people automatically use fast, intuitive, emotionally based System 1 or engage deliberate, controlled, cognitively based System 2 processes to make judgments (see Evans & Stanovich, 2013). In the artworld, these processes—aesthetic judgments—are determined by the fluency, or discrepancy, of an aesthetic experience. Fluency refers to how easily a work of art can be perceived, processed, and understood and has been related to positive aesthetic experiences (Graf & Landwehr, 2015, 2017; Reber et al., 2004; Winkielman et al., 2003).

There are two major models of aesthetic processing that consider fluency. The Pleasure-Interest model of Aesthetic liking (PIA; Graf & Landwehr, 2015, 2017) proposes that positive aesthetic experiences emerge from System 1 as pleasure (i.e., positive emotion) and from System 2 as interest (i.e., cognitive engagement). An audience will automatically engage their System 1 processes to determine their affective reactions based on perceptual fluency (i.e., how easily one can identify image characteristics; Reber et al., 2004). In contrast, System 2 processing is cognitively based, leading to impressions of interest, confusion, or boredom, and is deliberately engaged when a viewer is motived to take a closer look at an artwork (i.e., on their own volition, at the request of a date, or even by experimental demand). System 2 processing may be related to conceptual fluency (i.e., how easily one understands and relates to the art; Reber

[4] In fact, images generated by the more advanced algorithms (i.e., transformer based VQGAN and CLIP models) were classified as human-generated, or "tricking" participants, around 87% of the time.

et al., 2004). Similarly, the VIMAP (Pelowski et al., 2016, 2017) outlines a process where people can have positive, negative, or neutral aesthetic experiences emerging from different levels of stimuli discrepancy and self-relevance. More specifically, positive aesthetic experiences emerge from low discrepancy or from discrepancy reducing mechanisms (i.e., self-reflection; Pelowski et al., 2017).5 Discrepancy occurs when there is a belief that an artwork has a subjective meaning that is not immediately accessible (Parsons, 1987). Both of these models, the PIA and VIMAP, suggest that positive aesthetic experiences may emerge from fluent, low discrepancy, automatic processing.

Interpreting our findings through this lens, it may be the case that human-generated images typically have greater levels of perceptual fluency, or fewer discrepancies, and that machine-generated images are harder to process because they are perceptually disfluent, or have more discrepancies. Perceptual fluency (i.e., how easily a viewer can make sense of an image) is generally a function of the statistical image characteristics (i.e., color, contrast, symmetry) while conceptual fluency is processed by more complex cognitive and semantic processes (Graf & Landwehr, 2017; Reber et al., 2004). Machine-generated images are often characterized by visual indeterminacy (i.e., appearing coherent and meaningful, but missing definition or spatial interpretation on closer inspection; Hertzmann, 2020) and are often criticized for lacking intentionality, composition, and narrative (Cetinic & She, 2021; Coeckelbergh, 2017; McCormack et al., 2019). It may be the case that the disfluent characteristics of machine-generated images (i.e., visual indeterminacy) are adversely influencing participant's System 1 processing and resulting in less positive aesthetic experiences. When participants were instructed to identify the image source, however, they may have engaged in System 2 processing and were unable to identify the source, through controlled reflection, because the discrepancy was at the System 1 level. Overall, there appears to be *some* psychological differences between human- and machine-generated art, but this is only a tentative interpretation and there are several limitations and future directions worth discussing.

Limitations and Future Directions

There are several limitations with the current study and directions for future research. First, one limitation of the current study was sampling. The sample of participants, image stimuli, and context was limited. In terms of participants, a sample of MTurk participants was used. Future samples could include students, art students, or even members of an art community to capture varying levels of exposure, expertise, and interest in art. In terms of image stimuli, the stimuli used were sampled to make sure they were good representations of their content domains. However, the style (i.e., impressionist, postimpressionist, and expressionist) and content (i.e., landscapes) was limited to control for extraneous effects. Domain representativeness means that the images are good examples of high-quality human- and machine-generated art. Minimal noise means that the images include as little irrelevant or idiosyncratic information as possible to help identify a "true" effect (Podsakoff et al., 2003). Despite stimulus sampling efforts, another limitation is that the human-generated images were chosen from open-source museum collections, meaning that they had been already identified as high quality and curated by expert museum curators. In contrast, the machine-generated art was selected from a variety of algorithm image outputs, by the project authors based on our *a priori* criteria. Although psychological differences may exist between museum curated human art and researcher chosen AI art, practically, we wanted to identify images balanced on attractiveness, technical skill, and fame otherwise an unbalanced image may not be representative of a typical painting for that group and distract from the true effect. For example, comparing van Goh's The Starry Night against a pixelated, fractal, indeterminate computer piece would not be a fair comparison because the image source would be obvious.

Future research should continue to use broader sets of image stimuli, including different styles (i.e., abstract, classical, modern) and content (i.e., people, animals, objects) as algorithms continue advance and avoid the uncanny valley (Gangadharbatla, 2022). In fact, at the time of writing the results, OpenAI has launched DALL-E 2 (Ramesh et al., 2022), Google has launched Imagen (Saharia et al., 2022), Midjourney has entered open-beta, and Stability AI has launched Stable Diffusion. These projects all seem to generate landscapes, objects, and animals at high degrees of realism. With these increasingly realistic machine-generated artworks, future work should consider how people curate the AI art and identify which aesthetic judgment factors or image characteristics influence the decision to select the final AI image. Finally, in terms of context, contextual features play a large role in aesthetic judgment (Pelowski et al., 2017). Laboratory and survey research remove contextual variability which, as discussed by Carbon (2020), "[a]ccording to Brunswik (1956), we will not get 'fully representative' (p. 67) research with laboratory-oriented research that ignores such typical viewing and inspection behavior" (p. 3). Future work should continue to explore the influence of context on evaluations of machine-generated artwork.

Second, the current work was both a conceptual replication and exploratory. It was a conceptual replication (Derksen & Morawski, 2022) of earlier findings around preference for human art and classification of image source (i.e., research questions two and three). However, it took more of an inductive approach (Woo et al., 2017) to answering the primary research question around identifying the aesthetic judgment factors differentiating human- and machine-generated art. Future research could take a confirmatory approach to replicate and expand on the current findings. Future research should also continue to consider moderators of the aesthetic judgment process. Previous work has considered the effect of bias (Hong & Curran, 2019; Ragot et al., 2020) and effort heuristics (Chamberlain et al., 2018), however other individual differences could include artistic expertise (e.g., Hawley-Dolan & Winner, 2011, Moffat & Kelly, 2006), personality (e.g., Cleridou & Furnham, 2014; Swami & Furnham, 2012), decision-making style (Scott & Bruce, 1995), and expand on the effects of bias by considering algorithm aversion (Dietvorst et al., 2015; Logg et al., 2019). Finally, future work should continue examining the aesthetic judgment factors—particularly the aesthetic emotions. Aesthetic emotions appear to be involved in differences between human- and machine-generated art. Confirmatory research should aim to establish directionality (e.g., are emotions driving classifications or are classifications driving emotions) and

[5] It is worth noting that participants in the machine-generated image condition did not necessarily have negative aesthetic experiences, but rather they had less positive experiences compared to the human-generated image condition.

exploratory research should continue exploring the dimensionality of aesthetic emotions. As it stands, the AESTHEMOS captures the content of aesthetic emotions but does not have a simple factor structure (Schindler et al., 2017).

A final direction for future research is to consider the machine side of the human judgment of machine-generated art. More specifically, future work should continue exploring how the algorithms work, how text prompts influences output, and how changing the model parameters changes the output. In doing so, future research can continue to refine the algorithms by identifying the mechanisms underlying "discrepancies" and "disfluencies," like visual indeterminacy (Hertzmann, 2020), and other characteristics that may be driving differences in the perception of human- and machine-generated art. Future work could manipulate image characteristics with prompt engineering (i.e., Liu & Chilton, 2022; Zhou et al., 2022) to manipulate the experimental stimuli and create experimental conditions. For example, as nostalgia was found to be an important aesthetic judgment factor, future work could manipulate the degree of positive emotionality or nostalgia in the image outputs by including simple prompt additions like, "a happy…," "a sad…," "a nostalgic painting of…," or "a carefree painting of…." Similarly, perhaps different degrees of positivity or nostalgia can be induced by prompting the algorithms to manipulate image characteristics, like the vividness of colors, by including words like "bright" or "dull" (see Graham & Redies, 2010 for more on image characteristics). In addition to prompt engineering, parameter tuning gives researchers even more control over the output of the algorithms. The combination of prompt engineering and parameter tuning gives has implications for experimental design, more generally, as algorithms are able to create novel stimuli with high degrees of experimenter control (see Utz & DiPaola, 2021).

As the fields of aesthetic philosophy and computation creativity converge around these novel text-to-image algorithms, developments in algorithmic image generation can inform our understanding of art, aesthetics, and creativity. Similarly, our understanding of aesthetic perception and judgment can continue to inform the tuning of existing transformer models and the development of new models. There are interesting parallels between AI art and the development of earlier art technologies, such as cameras and digital editing. In the 1840s, after seeing a collection of photographs, the French painter Paul Delaroche is alleged to have exclaimed, "from today painting is dead!." Like a paintbrush, camera, or editing software, these new foundational text-to-image models are still tools that human artists can use to enhance their own artistic creations. As it stands, humans still have control over the input, the algorithm parameters, and the final decision around what image is most impactful and meaningful. And as the current findings show, people still feel better about human-generated art. However, after asking ChatGPT,6 a cousin of the algorithms used in this project, what it thought of the results, it wrote: "as an AI, I would say that the study's findings are interesting, but I would also caution against drawing broad conclusions about the relative merits of human versus AI-generated artwork based on a single study. It's important to remember that AI is a tool that can be used in many different ways, and its capabilities can vary greatly depending on how it is designed and trained. Therefore, it's possible that in some cases, AI-generated artwork may be just as attractive, exciting, and nostalgic as human-generated artwork. Ultimately, the value of any artwork should be judged on its own merits, rather than on the basis of who or what created it" (see the OSF for the full conversation transcript: https://osf.io/5kqdh).

Conclusion

Creative, artistic pursuits have long been thought to be safe from machines (Bakhshi et al., 2015; Boden, 2010) in a world where upward of 50% of employment is at risk of automation (Manyika et al., 2017). Algorithms are now able to perform these deeply human pursuits (Rahwan et al., 2019) and this has prompted researchers to study how people perceive machine-generated art (e.g., Chamberlain et al., 2018; Gangadharbatla, 2022; Hong & Curran, 2019; Ragot et al., 2020). In an emerging area of research, our study also built on the earlier work by integrating theoretical frameworks of aesthetic processing and using psychometric measures of aesthetic judgment to explore which aesthetic judgment factors differentiate human and machine art. Overall, we found that people cannot distinguish between human- and machine-generated art but that they prefer the human artwork and experience more positive emotions when shown human-generated art. In other words, while people may *perceive* human- and machine-generated art in the same way, the way they *feel* about the art is different. People have more positive emotions with human-generated art.

References

Abramson, D. (2008). Turing's responses to two objections. *Minds and Machines*, *18*(2), 147–167. https://doi.org/10.1007/s11023-008-9094-6

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*(1), 1–5. https://doi.org/https://doi.org/10.1027/1015-5759/a000699

Amabile, T. M., & Pillemer, J. (2012). Perspectives on the social psychology of creativity. *The Journal of Creative Behavior*, *46*(1), 3–15. https://doi.org/10.1002/jocb.001

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bakhshi, H., Frey, C. B., & Osborne, M. (2015). *Creativity vs robots: The creative economy and the future of employment*. Nesta.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Boden, M. A. (2007). Authenticity and computer art. *Digital Creativity*, *18*(1), 3–10. https://doi.org/10.1080/14626260701252285

Boden, M. A. (2010). The Turing test and artistic creativity. *Kybernetes*, *39*(3), 409–413. https://doi.org/10.1108/03684921011036132

Boedeker, P., & Kearns, N. T. (2019). Linear discriminant analysis for prediction of group membership: A user-friendly primer. *Advances in Methods and Practices in Psychological Science*, *2*(3), 250–263. https://doi.org/10.1177/2515245919849378

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., ... Liang, P. (2021). *On the opportunities and risks of foundation models*. ArXiv: arXiv:2108.07258v3. https://doi.org/10.48550/arXiv.2108.07258

[6] https://openai.com/blog/chatgpt/.

Bringsjord, S., Bello, P., & Ferrucci, D. (2001). Creativity, the Turing test, and the (better) Lovelace test. *Minds and Machines*, *11*(1), 3–27. https://doi.org/10.1023/A:1011206622741

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Carbon, C.-C. (2020). Ecological art experience: How we can gain experimental control while preserving ecologically valid settings and contexts. *Frontiers in Psychology*, *11*, Article 800. https://doi.org/10.3389/fpsyg.2020.00800

Carnovalini, F., & Rodà, A. (2020). Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, *3*, Article 14. https://doi.org/10.3389/frai.2020.00014

Cetinic, E., & She, J. (2021). *Understanding and creating art with AI: Review and outlook*. arXiv http://arxiv.org/abs/2102.09109

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 177–192. https://doi.org/10.1037/aca0000136

Cleridou, K., & Furnham, A. (2014). Personality correlates of aesthetic preferences for art, architecture, and music. *Empirical Studies of the Arts*, *32*(2), 231–255. https://doi.org/10.2190/EM.32.2.f

Coeckelbergh, M. (2017). Can machines create art? *Philosophy & Technology*, *30*(3), 285–303. https://doi.org/10.1007/s13347-016-0231-5

Collingwood, R. G. (1938). *The principles of art*. Oxford University Press.

Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? *European Conference on Artificial Intelligence*, *242*, 21–26. https://doi.org/10.3233/978-1-61499-098-7-21

Crowson, K., Biderman, S. R., Kornis, D., Stander, D., Hallahan, E., Castricato, L., & Raff, E. (2022). VQGAN-CLIP: Open domain image generation and editing with natural language guidance. *arXiv: 2204.08583v2*. https://doi.org/10.48550/arXiv.2204.08583

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, *123*(1), 2–22. https://doi.org/10.1037/a0039802

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–104. https://doi.org/10.1177/147078530805000106

Dee, C. M. A. (2018). Examining copyright protection of AI-generated art. *Delphi—Interdisciplinary Review of Emerging Technologies*, *1*(1), 31–37. https://doi.org/10.21552/delphi/2018/1/11

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication". *Perspectives on Psychological Science*, *17*(5), 1490–1505. https://doi.org/10.1177/17456916211041116

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dorin, A. (2013). Chance and complexity: Stochastic and generative processes in art and creativity. *Proceedings of the virtual reality international conference: Laval virtual* (pp. 1–8). https://doi.org/10.1145/2466816.2466837

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). *CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms*. arXiv http://arxiv.org/abs/1706.07068

Esser, p., Rombach, R., & Ommer, B. (2021, June 19–25). Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual conference* (pp. 12868–12878). https://ieeexplore.ieee.org/document/9578911

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576–588. https://doi.org/10.1037/amp0001006

French, R. M. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences*, *4*(3), 115–122. https://doi.org/10.1016/S1364-6613(00)01453-4

Friendly, M., & Fox, J. (2020). *candisc: Visualizing generalized canonical discriminant and canonical correlation analysis*. https://CRAN.R-project.org/package=candisc

Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *Neural Networks*, *144*, 187–209. https://doi.org/10.1016/j.neunet.2021.07.019

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gangadharbatla, H. (2022). The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts*, *40*(2), 125–142. https://doi.org/10.1177/0276237421994697

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27). https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Graf, L. K. M., & Landwehr, J. R. (2015). A dual-process perspective on fluency-based aesthetics: The pleasure-interest model of aesthetic liking. *Personality and Social Psychology Review*, *19*(4), 395–410. https://doi.org/10.1177/1088868315574978

Graf, L. K. M., & Landwehr, J. R. (2017). Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Frontiers in Psychology*, *8*, Article 15. https://doi.org/10.3389/fpsyg.2017.00015

Graham, D. J., & Redies, C. (2010). Statistical regularities in art: Relations with visual coding and perception. *Vision Research*, *50*(16), 1503–1509. https://doi.org/10.1016/j.visres.2010.05.002

Groh, M., Epstein, Z., Obradovich, N., Cebrian, M., & Rahwan, I. (2019). *Human detection of machine manipulated media*. arXiv http://arxiv.org/abs/1907.05276

Gunkel, D. J. (2017). Special section: Rethinking art and aesthetics in the age of creative machines: Editor's introduction. *Philosophy & Technology*, *30*(3), 263–265. https://doi.org/10.1007/s13347-017-0281-3

Gunkel, D. J. (2021). Computational creativity: Algorithms, art, and artistry. In E. Navas, O. Gallagher, X. Burrough (Eds.), *The Routledge handbook of remix studies and digital humanities* (pp. 385-395). Routledge.

Hager, M., Hagemann, D., Danner, D., & Schankin, A. (2012). Assessing aesthetic appreciation of visual artworks—The construction of the Art Reception Survey (ARS). *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 320–333. https://doi.org/10.1037/a0028776

Hassine, T., & Neeman, Z. (2019). The zombification of art history: How AI resurrects dead masters, and perpetuates historical biases. *Journal of*

Science and Technology of the Arts, *11*(2), 28–35. https://doi.org/10.7559/citarj.v11i2.663

Hawley-Dolan, A., & Winner, E.. (2011). Seeing the mind behind the art: People can distinguish abstract Expressionist paintings from highly similar paintings by children, chimps, monkeys, and elephants. *Psychological Science*, *22*(4), 435–441. https://doi.org/10.1177/0956797611400915

Hertzmann, A. (2018). Can computers create art? *Arts*, *7*(2), Article 18. https://doi.org/10.3390/arts7020018

Hertzmann, A. (2020). Visual indeterminacy in GAN art. *Leonardo*, *53*(4), 424–428. https://doi.org/10.1162/leon_a_01930

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods*, *12*(3), 554–566. https://doi.org/10.1177/1094428107300396

Hong, J. W. (2018). Bias in perception of art produced by artificial intelligence. In M. Kurosu (Ed.), *Human-computer interaction. Interaction in context. HCI 2018*. Lecture notes in computer science (Vol. 10902). Springer. https://doi.org/10.1007/978-3-319-91244-8_24

Hong, J.-W., & Curran, N. M. (2019). Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *15*(2s), 1–16. https://doi.org/10.1145/3326337

Hudson, D. A., & Zitnick, L. (2021). Generative adversarial transformers. *Proceedings of the 38th international conference on machine learning* (pp. 4487–4499). https://proceedings.mlr.press/v139/hudson21a.html

Hwang, D., Schmitt, W. A., Stephanopoulos, G., & Stephanopoulos, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, *18*(9), 1184–1193. https://doi.org/10.1093/bioinformatics/18.9.1184

Jacobsen, T. (2010). Beauty and the brain: culture, history and individual differences in aesthetic appreciation. *Journal of Anatomy*, *216*(2), 184–191. https://doi.org/10.1111/j.1469-7580.2009.01164.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, Article 106553. https://doi.org/10.1016/j.chb.2020.106553

Krauss, L., Ott, C., Opwis, K., Meyer, A., & Gaab, J. (2021). Impact of contextualizing information on aesthetic experience and psychophysiological responses to art in a museum: A naturalistic randomized controlled trial. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 505–516. https://doi.org/10.1037/aca0000280

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Emerald Group.

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9

Larson, E. J. (2021). The myth of artificial intelligence: Why computers can't think the way we do. In *The myth of artificial intelligence*. Harvard University Press. https://doi.org/10.4159/9780674259935

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, *95*(4), 489–508. https://doi.org/10.1348/0007126042369811

Leder, H., Gerger, G., Dressler, S. G., & Schabmann, A. (2012). How art is appreciated. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 2–10. https://doi.org/10.1037/a0026396

Leder, H., & Nadal, M. (2014). Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode—Developments and challenges in empirical aesthetics. *British Journal of Psychology*, *105*(4), 443–464. https://doi.org/10.1111/bjop.12084

Lindell, A. K., & Mueller, J. (2011). Can science account for taste? Psychological insights into art appreciation. *Journal of Cognitive Psychology*, *23*(4), 453–475. https://doi.org/10.1080/20445911.2011.539556

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. *CHI conference on human factors in computing systems* (pp. 1–23). https://doi.org/10.1145/3491102.3501825

Llano, M. T., d'Inverno, M., Yee-King, M., McCormack, J., Ilsar, A., Pease, A., & Colton, S. (2022). *Explainable computational creativity*. arXiv http://arxiv.org/abs/2205.05682

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S.. (2017). *Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages*. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

Matthews, R. A., Pineault, L., & Hong, Y. H. (2022). Normalizing the use of single-item measures: Validation of the single-item compendium for organizational psychology. *Journal of Business and Psychology*, *37*(4), 639–673. https://doi.org/10.1007/s10869-022-09813-3

Mazzone, M., & Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. *Arts*, *8*(1), Article 26. https://doi.org/10.3390/arts8010026

McCormack, J., Gifford, T., & Hutchings, P. (2019). Autonomy, authenticity, authorship and intention in computer generated art. In A. Ekárt, A. Liapis, & M. L. Castro Pena (Eds.), *Computational intelligence in music, sound, art and design* (pp. 35–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-16667-0_3

Menninghaus, W., Wagner, V., Wassiliwizky, E., Schindler, I., Hanich, J., Jacobsen, T., & Koelsch, S. (2019). What are aesthetic emotions? *Psychological Review*, *126*(2), 171–195. https://doi.org/10.1037/rev0000135

Moffat, D., & Kelly, M. (2006, August 28–29). *An investigation into people's bias against computational creativity in music composition*. Paper presented at The Third Joint Workshop on Computational Creativity (ECAI'06). Riva del Garda, Italy. https://computationalcreativity.net/ijwcc06/

Mullennix, J. W., Kristo, G. M., & Robinet, J. (2020). Effects of preceding context on aesthetic preference. *Empirical Studies of the Arts*, *38*(2), 149–171. https://doi.org/10.1177/0276237418805687

Nadal, M., & Vartanian, O. (2021). Empirical aesthetics: An overview. In M. Nadal & O. Vartanian (Eds.), *The Oxford handbook of empirical aesthetics* (pp. 3 - 38). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198824350.013.1

Natale, S., & Henrickson, L. (2022). The Lovelace effect: Perceptions of creativity in machines. *New Media & Society*. https://doi.org/10.1177/14614448221077278

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Osborne, J. (2020). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, *15*(12). https://doi.org/10.7275/qbpc-gk17

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology*, *64*(1), 77–107. https://doi.org/10.1146/annurev-psych-120710-100504

Parsons, M. J. (1987). *How we understand art: A cognitive developmental account of aesthetic experience*. Cambridge University Press.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), Article 100336. https://doi.org/10.1016/j.patter.2021.100336

Pearce, M. T., Zaidel, D. W., Vartanian, O., Skov, M., Leder, H., Chatterjee, A., & Nadal, M. (2016). Neuroaesthetics: The cognitive neuroscience of aesthetic experience. *Perspectives on Psychological Science*, *11*(2), 265–279. https://doi.org/10.1177/1745691615621274

Pelowski, M., Markey, P. S., Forster, M., Gerger, G., & Leder, H. (2017). Move me, astonish me… delight my eyes and brain: The Vienna Integrated Model of top-down and bottom-up processes in Art Perception (VIMAP) and corresponding affective, evaluative, and neurophysiological correlates. *Physics of Life Reviews*, *21*, 80–125. https://doi.org/10.1016/j.plrev.2017.02.003

Pelowski, M., Markey, P. S., Lauring, J. O., & Leder, H. (2016). Visualizing the impact of art: An update and comparison of current psychological models of art experience. *Frontiers in Human Neuroscience*, *10*, Article 160. https://doi.org/10.3389/fnhum.2016.00160

Pistilli, G. (2022). What lies behind AGI: Ethical concerns related to LLMs. *Revue Ethique et Numérique*. hal-03607808f

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021, July 18–24). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). https://proceedings.mlr.press/v139/

Ragot, M., Martin, N., & Cojean, S. (2020). AI-generated vs. human artworks. A perception bias towards artificial intelligence? *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–10). https://doi.org/10.1145/3334480.3382892

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). *AI and the everything in the whole wide world benchmark*. arXiv http://arxiv.org/abs/2111.15366

Ramachandran, V. S., & Hirstein, W. (1999). The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies*, *6*(6–7), 15–51.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. arXiv:2204.06125v1. https://doi.org/10.48550/arXiv.2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-shot text-to-image generation*. arXiv http://arxiv.org/abs/2102.12092

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's Processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Ren, F., & Bao, Y. (2020). A review on human–computer interaction and intelligent robots. *International Journal of Information Technology & Decision Making*, *19*(1), 5–47. https://doi.org/10.1142/S0219622019300052

Riedl, M. O. (2014). *The Lovelace 2.0 test of artificial creativity and intelligence*. arXiv http://arxiv.org/abs/1410.6142

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). *Photorealistic text-to-image diffusion models with deep language understanding*. arXiv https://doi.org/10.48550/arXiv.2205.11487

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., & Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLOS ONE*, *12*(6), Article e0178899. https://doi.org/10.1371/journal.pone.0178899

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*, *55*(5), 818–831. https://doi.org/10.1177/0013164495055005017

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Skov, M., & Nadal, M. (2020). A farewell to art: Aesthetics as a topic in psychology and neuroscience. *Perspectives on Psychological Science*, *15*(3), 630–642. https://doi.org/10.1177/1745691619897963

Smith, K. N., Lamb, K. N., & Henson, R. K. (2020). Making meaning out of MANOVA: the need for multivariate post hoc testing in gifted education research. *Gifted Child Quarterly*, *64*(1), 41–55. https://doi.org/10.1177/00169862198903

Specker, E. (2021). Further validating the VAIAK: Defining a psychometric model, configural measurement invariance, reliability, and practical guidelines. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000427

Specker, E., Forster, M., Brinkmann, H., Boddy, J., Pelowski, M., Rosenberg, R., & Leder, H. (2020). The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A unified and validated measure of art interest and art knowledge. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(2), 172–185. https://doi.org/10.1037/aca0000205

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Erlbaum.

Swami, V., & Furnham, A. (2012). The effects of symmetry and personality on aesthetic preferences. *Imagination, Cognition and Personality*, *32*(1), 41–57. https://doi.org/10.2190/IC.32.1.d

Todorov, P. (2019). *A game of dice: Machine learning and the question concerning art*. arXiv http://arxiv.org/abs/1904.01957

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Utz, V., & DiPaola, S. (2021). Exploring the application of AI-generated artworks for the study of aesthetic processing. *2021 IEEE 4th international conference on multimedia information processing and retrieval (MIPR)* (pp. 393–398). https://doi.org/10.1109/MIPR51284.2021.00073

van Tilburg, W. A. P., Wildschut, T., & Sedikides, C. (2018). Nostalgia's place among self-relevant emotions. *Cognition and Emotion*, *32*(4), 742–759. https://doi.org/10.1080/02699931.2017.1351331

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, December 4–7). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA* (pp. 5998–6008). https://nips.cc/Conferences/2017

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language models*. arXiv http://arxiv.org/abs/2112.04359

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. https://doi.org/10.1177/01461672992512005

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The Hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Lawrence Erlbaum Associates Publishers.

Woo, S. E., O'Boyle, E. H., & Spector, P. E. (2017). Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review*, *27*(2), 255–264. https://doi.org/10.1016/j.hrmr.2016.08.004

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), 487–497. https://doi.org/10.1038/nrn3962

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). *Learning to prompt for vision-language models*. arXiv http://arxiv.org/abs/2109.01134

Received September 2, 2022
Revision received December 2, 2022
Accepted January 12, 2023 ■