Exploring Bias and Identification in Character Art: A Comparison of AI and Human Creations Using Eye-Tracking

Wangfan Li

***Abstract*—With the rapid advancement of artificial intelligence (AI), AI is increasingly capable of generating artwork. AI-generated art, particularly art depicting human characters, has gained significant attention, so it is very important regarding how viewers perceive it compared to human-created works. This study aims to examine the differences in how people evaluate AI-generated vs. human-generated character art, such as eye movements, focusing on two distinct styles: realistic and cartoonish. Using eye-tracking technology, this paper investigates participants' visual attention and decision-making processes during the evaluation of these artworks. Additionally, participants will try to identify which of the two presented artwork was created using AI. By analyzing eye-tracking data and their evaluation, this research seeks to uncover uncommon strategies attempted to identify AI-generated artwork, any potential bias, and human's ability to accurately identify AI art.**

I. Introduction

With the recent development of artificial intelligence (AI), especially generative AI, it is now possible to generate music, visuals and texts that are accessible to most consumers. On the topic of generative AI such as Stable Diffusion and Midjoruney, they can produce artworks unmatched by humans regarding speed, but it is still uncertain how most people will engage with such medium, and many feel such art is ”souless”.

For the potential impact of this technology, many studies have sought to view it from the angle of an artist. [Jiang et al.(2023)], for example, shows how it could potentially hurt an artist's ability to compete and recommends regulation to lessen the impact. However, it is also important to consider how non-artists react to this technology. Many people on online forums are confident in their ability to tell AI vs human-generated art apart from each other and show disgust for AI artwork in general when such examples are identified [Bosonogov and Suvorova(2023)].

This study then aims to investigate how people try to identify the difference between human art and AI art, focusing on eye-tracking measures. Specifically, the study will examine whether viewers display different eye patterns and the corresponding decision-making strategies when evaluating AI vs. human art, how accurately they perform, and if there is any difference when examining human portrait vs landscaping art. The result of this study should shed more light on how people engage with any potential AI art in general.

II. Related Work

With the recent advancement in generative AI models, many studies have examined their effectiveness and accuracy, such as a tool for brainstorming [Mansour(2023)] and character drawing [Jie et al.(2023)], there still exist limitations to it, such as prompt misunderstanding and inaccurate characters.

Studies are also done on trying to examine user's attitudes when it comes to AI, and it is shown that most people still prefer human-made art, even if they are not aware of such bias or accurately identify AI-generated art, using eye-tracking, [Zhou and Kawabata(2023)] found that participants looked longer at artworks they believed to be created by humans, despite being unable to accurately identify the differences. Their findings suggest a subconscious preference for human-made art, mostly because AI arts are perceived as ”lazy”, ”less creative”, possibly because creativity is still an attribute that people associate with only humans [Millet et al.(2023)]. Though there are some that view AI-generated art as art [Hong(2018)]. Importantly, this bias is not based on some inherent attribute that AI art has, but rather a bias that exists disregarding how accurately humans can identify AI art, and indeed there's a significant challenge to accurately identify them [Chamberlain et al.(2018)], showing that while participants found it difficult to distinguish between AI and human works, they liked human-made art more.

Despite these important contributions, it is still to be seen if any differences exist between landscaping art and human portrait art, might affect perception. Additionally, it is important to find out how humans typically go about identifying AI art in terms of eye movement. Lastly, while most existing research has focused on abstract or landscape paintings one at a time, few studies have explored character-based artworks when presented side by side, which provides a different dimension to non-character-based artworks. This study aims to build on the foundation laid by previous work by specifically using cartoon and realistic character-based artworks side by side in identification tasks.

III. Methods

A. Apparatus

Our experiment utilizes a Gazepoint GP3 device with a 60HZ sampling rate on a 1920x1080 screen; the participants will be seated at around 60cm away from the screen. The device will be controlled through the Gazepoint Control software, while Gazepoint Analysis software will be used to operate the experiment and present the task. The resulting data will be analyzed using R software.

B. Stimulus

The material consists of 4 sets of pictures, with each set consisting of 2 pictures side by side, with one AI-generated picture and one human-generated picture. Out of the 4 sets, 2 sets consist of landscaping art, while the other 2 sets consist of human portrait art.

The human-generated arts are obtained through WikiArt, chosen based on how semi-randomly after filtering the subject (landscape vs portrait) and time period (1900). For the AI-generated art, the human picture was first given to GPT 4o to generate a description, and that description was then used as a prompt to give to Midjourney v6.1 to generate a picture both in style and dedication that's similar to the human original. 6 pictures were generated and the best one was chosen based on the researcher's judgement.

Each artwork will be displayed on a neutral light-gray background. The resolution of the images will be normalized to ensure viewing capability.

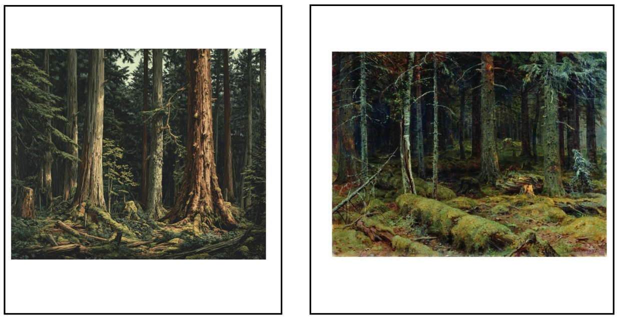

Fig. 1. *One set of images in the landscape group, the one on the left is AI generated, with the two AOI outlined.*

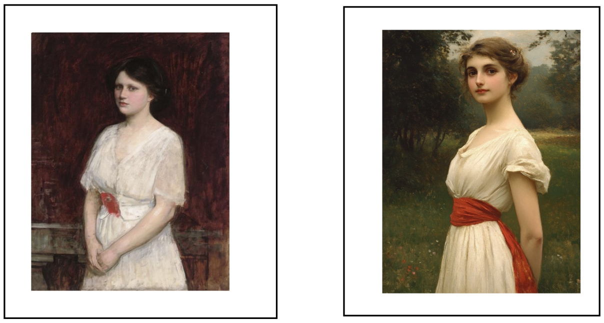

Fig. 2. *One set of images in the portrait group, the one on the right, is AI generated, with the two AOI outlined.*

C. Subjects

For this study, 9 participants were recruited from the University, which included 7 males and 2 females,. All of them are students and are aware of generative AI and use it to some degree in work or outside of it.

D. Experimental Design

Using a within-subject design, the participants were exposed to all 4 sets of pictures; the only variable differentiating the two different conditions was the subject depicted (landscape vs. portrait). The dependent variables are eye-tracking metrics (fixation duration, count, saccades) and the decision-making metrics in which the participants think which of the two artworks is AI generated and the corresponding confidence score and rate which artwork they prefer.

We hypothesize that: H1: Participants will more easily identify the AI-generated image in the portrait condition since we as humans are very sensitive to the human form, and any minor mistake can tip the participants off that something is wrong [Mori et al.(2012)]. H2: Participants will have an accuracy higher than a random chance for identifying the AI picture. H3: Participants will have the highest fixation count for the AI image as they find out any minor mistakes that AI makes and focus on them.

E. Procedures

The participants will be seated first in front of the computer screen, which is equipped with the Gazepoint eye-tracking device; they will then be provided with a brief overview of the experiment, explaining that they will view a series of artworks, and be asked to make evaluations as to which of the two artworks is AI-generated. The participants will then undergo the standard 5-point calibration procedure for the device.

The participants will view 4 sets of artworks; the artworks will be presented randomly with one AI-generated and one human-created artwork, two at a time, side by side so that the participants will be viewing two at a time. After 15 seconds, the participants will be directed to a survey screen to answer which one they think is the AI-generated artwork, how confident they are using a Likert scale, and which artwork they prefer. There will be no time limit on rating, and the participants can proceed to the next set when they are ready.

When all sets are viewed, the participants will answer a short post-experiment survey, including how difficult they find this activity and how they would rate their ability to identify AI artworks.

IV. Result

For accuracy, the total accuracy is 0.53, while it's also divided into portrait accuracy, which is 0.694 and landscape, which is 0.416. A one-tailed paired t-test was conducted to check for significance. The results show that the difference between landscape accuracy and portrait accuracy was statistically significant at the 0.05 level, t(8) = -2.294, p = 0.02593. The mean difference was -0.278.

A one-sample t-test was conducted to compare the mean total accuracy to the chance level of 0.5 since if the participants were just guessing, it should be around 0.5 or 50/50. The results indicated that the mean total accuracy (M=0.53, t(8)=0.32, P =0.76) is not significant enough to show that the participants did better than random on average.

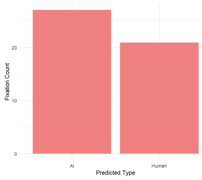

Fig. 3. *Fixation count by type.*

Another two-tailed paired t-test was done to check for significant differences between user's choice of what they think is AI at the end and the fixation count. The results show that the mean fixation count for pictures judged human was lower than that for AI-generated fixations, with a mean difference of -7. However, this difference was significant, t(7) = -2.36, p = 0.050. We then check the multiple linear regression to examine the relationship between the fixation number of what the participant predicts to the AI and the actual accuracy; the resulting model is not significant at F(2, 5) = 0.714, p = 0.534, R squared = 0.222, the two predictors used, the fixation number on what the participant ended predicting to be AI(p=0.286), and the fixation number what the participant ended predicting to be human created (p=0.86), are both nonsignificant.

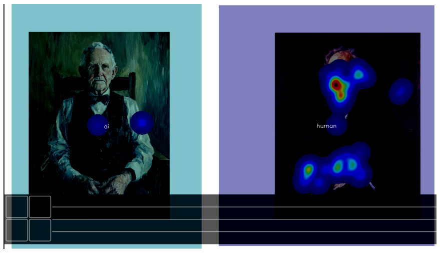

Fig. 4. *An example of the heatmap generated.*

V. Discussion

With a mean difference of around -0.278 and a significant difference in accuracy, it shows that there is a significant accuracy difference between portrait and landscape groups, and it makes sense regarding the literature, as humans are very good at identifying human features. The researcher also observed that one of the sets in the portrait group depicts a woman. The AI tends to increase the woman's attractiveness to near model level while the human-drawn one is very realistic, showing that this might be a bad pairing that presents an easy clue to the participants, skewing the result. This also shows the tendency for AI to generate more beautiful women while not being the case with the man that's paired in the other set, but this finding supports our H1 hypothesis.

With the result showing that the participants did not significantly perform better than random chance, it lends to reject our H2; I personally think this shows the pace of improvement regarding generated AI, where even I, as the researcher, have a hard time identifying, and that it's pretty much impossible when it comes to landscape art. However, further studies are warranted to prove this.

The follow up of the t-test showed a significant result for the fixation number on the image that's predicted to be AI, while it's not significant as a predictor for accuracy, meaning the image that participants paid more attention to is likely to be identified as AI, but extra attention is not shown to actually improvement accuracy. This might mean that as participants paid more attention, they identified more perceived flaws and thus were more likely to select that image while not equally weighing the other image that was paid less attention; this could be a form of bias and needs more data from further study.

A big limitation of this study is the limited number of participants, where there's not enough power to find small or medium effects, and because of the recruitment strategy, all participants are familiar with how generative AI works, so the findings might not hold true for the general population. For future studies, it might be interesting to test out different art styles instead of just the classical paintings used in this study. It might also be of interest to examine more subjects that are outside of landscaping and portraits, but with how fast this area is being developed, a worry might be that the findings shed more light on the particular new models of AI instead of something fundamental to human psychology.

VI. Conclusion

In conclusion, the study explored how participants differentiate between AI-generated and human-created artworks. The findings show a significant difference in accuracy between portrait and landscape art, with participants performing better at identifying AI-generated portraits. However, participants did not perform significantly better than random chance overall, suggesting the increasing capability of AI-generated art generators.

Interestingly, the fixation count was significantly higher for images perceived as AI-generated, indicating that participants tended to scrutinize these images more. However, this increased attention did not translate into greater accuracy, showing that this might be a form of bias instead of a valid strategy.

The study is limited by its small sample size and the demographics of the participants, as the researcher had some trouble accessing more recruitment strategies. Future research should involve a larger, more diverse population and explore other art styles and genres to generalize findings.

References

[Bosonogov and Suvorova(2023)] Semen Dmitrievich Bosonogov and Alena Vladimirovna Suvorova. 2023. Perception of AI-generated art: Text analysis of online discussions. 529, 0 (2023), 6–23.

[Chamberlain et al.(2018)] Rebecca Chamberlain, Caitlin Mullin, Bram Scheerlinck, and Johan Wagemans. 2018. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts* 12, 2 (2018), 177.

[Hong(2018)] Joo-Wha Hong. 2018. Bias in perception of art produced by artificial intelligence. In *Human-Computer Interaction. Interaction in Context: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II 20*. Springer, 290–303.

[Jiang et al.(2023)] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 363–374.

[Jie et al.(2023)] Pingjian Jie, Xinyi Shan, and Jeanhun Chung. 2023. Comparative Analysis of AI Painting Using [Midjourney] and [Stable Diffusion]-A Case Study on Character Drawing. *International Journal of Advanced Culture Technology* 11, 2 (2023), 403–408.

[Mansour(2023)] Soha Mansour. 2023. Intelligent graphic design: The effectiveness of midjourney as a participant in a creative brainstorming session. *International Design Journal* 13, 5 (2023), 501–512.

[Millet et al.(2023)] Kobe Millet, Florian Buehler, Guanzhong Du, and Michail D Kokkoris. 2023. Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior* 143 (2023), 107707.

[Mori et al.(2012)] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[Zhou and Kawabata(2023)] Yizhen Zhou and Hideaki Kawabata. 2023. Eyes can tell: Assessment of implicit attitudes toward AI art. *i-Perception* 14, 5 (2023), 20416695231209846.