Artificial Impressions: Examining AI's Impact on Image Perception

Group 20: Ananya Anand, Sneha Ekka, Zhengyi Hong, Lyushen Song, Leonardo Trucios, Mauro Wang

1. Introduction

In recent years, the capabilities of artificial intelligence (AI) have advanced rapidly. A recent breakthrough is the ability for AI to generate images from text input. These AI-generated images are both realistic and artistic, challenging people's understanding of what creativity means and our concept of aesthetic appreciation. While many appreciate AI's ability to produce images, others have mixed feelings about it. They argue that AI-generated images lack soul and fail to capture the essence of humanity and beauty.

Previous studies have demonstrated a clear bias in the perception of artworks based on their supposed origin, whether created by humans or AI. Moreover, there is emerging discourse around whether Artificial Integene can be considered "artists" of their own right.(Ragot, Martin, and etc, 2020) Research indicates that while the outputs of human and artificial intelligence are judged similarly in terms of being recognized as 'art,' AI is significantly less likely to be acknowledged as 'artists,' primarily due to a perceived lack of being an individual entity capable of intentional creativity. (Mikalonytė, Sigutė, and etc 2022)

Our project aims to explore the effect of knowing that an image is AI-generated on people's appreciation of them. We designed an experiment to investigate how pre-informing participants on whether or not an artwork is AI-generated influences their appreciation of the artwork itself. Hence, we aim to capture the psychological mechanisms that shape human perception of artificial intelligence disruption in the realm of visual art creation.

2. Research Question

How does disclosing the origin of images (AI-generated or not) influence people's appreciation scores of the images?

3. Specific Hypothesis

We have structured our experiment to have our participants divided into treatment and control groups, with the treatment group being told explicitly that the images they are about to see are generated by AI. The control group, however, is only told that they will be reviewing some artwork, with no specification on origin. **Given the above, we hypothesize that the knowledge that an image is AI-generated may negatively influence participants' judgments of the quality and visual attractiveness of the images.** By comparing the ratings between the treatment and control groups, the experiment seeks to determine whether the knowledge that the images are AI-generated results in a significant impact on participants' evaluations.

4. Methodology

4.1 Participants

The participants for this study were selected through two primary channels: scanning a QR code from the flyers we distributed across Boston University and viewing Instagram stories of our team mates that linked to the survey form. Individuals who encountered the QR code or viewed the Instagram story were provided with a brief description of the study before being split by Qualtrics into control and treatment groups. This recruitment method ensured a random selection of participants from a diverse pool of potential candidates.

Fig. 1.1

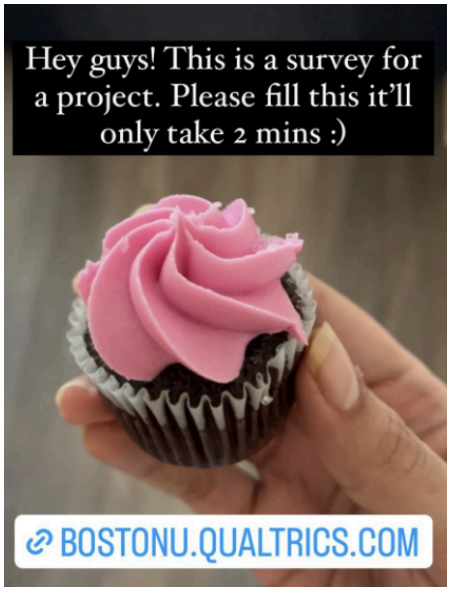

Fig 1.2

4.2 Treatment and Randomization

The treatment in our experiment involved explicitly informing participants that they were to review and rate AI-generated artwork. In the treatment group, participants were told in an introductory screen "Now, we are going to show you some AI generated artwork." Participants in the control group were instead provided with a more generic introductory screen of "Now, we are going to show you some artwork". Considering budget and time limitations, the optimal approach for conducting an experiment with a sufficient sample size was to employ randomization at the user level. Utilizing Qualtrics for our survey allowed us to implement a Randomizer feature, ensuring users were randomly assigned to either the treatment or control group in a 50/50 distribution.

4.3 Procedure Overview

The first step in this experiment involved generating suitable AI images, focusing mainly on artistic themes. We curated a diverse set of 22 images encompassing various visual art genres such as landscape, modern art, realism, and more. See appendix for all images used in this experiment.

After doing that, in the pre-experimental phase, we conducted a power test to estimate the potential statistical power of the experiment, as further described in 4.5. We found that we would need to collect 973 samples for sufficient experimental power. We tried our best but ended up collecting 183 samples, and after some surveys that glitched out due to a technical error in the logic, we ultimately ended up with 144 samples.

After collecting the data, we processed to estimate the average treatment effect and conditional average treatment effect, considering specific participant characteristics such as gender and if they attend Boston University. We conducted regression analysis and regression analysis with fixed effects to derive these metrics. Finally, to ensure proper randomization, we deployed a proportions_ztest to double check our distributions.

4.4 Survey Design

To collect data, we designed our survey using Qualtrics. The survey began by gathering brief demographic information from participants, including gender, age, and educational background. Participants were then randomly assigned either the treatment or the control group. Subsequently, participants rated 10 randomly selected images from a total of 22 images generated on a scale of 0 to 100. Refer to the appendix for detailed survey design and link.

4.5 Pre-Experiment Power Test Estimation

In a study by Messingschlager and Markus et al. (2023) Cohen's D of -0.09 was calculated as the effect size for their artwork appreciation experiment. Extrapolating their findings to our project, after scaling and calculating, we projected an ATE of -1.76 with a standard deviation of 19.57 for mean rating scores, which derived in a Cohen's d of -0.09, which is very close to the result obtained from their paper. In order to develop an experiment design with at least 80% of statistical power, we need at least a sample of 973 surveys filled, which means that the effect will be detected 80% of the time if we reach that level of surveys under the assumption of a true effect of -1.76.

4.6 Pre-analysis Proper Randomization Check

Once the data was gathered, we used proportions_ztest to validate the randomization of the experiment as done by Qualtrics.

4.6.1 Randomization Check - The Treatment and Control Group

a) Proportions Test

```
Z-test statistic: -0.3334620086260372
P-value: 0.7387855626520152
```

Fig.2

b) Balance Check - Regression

	(1)	(2)	(3)
Any_Treatment	0.048	-0.103	0.004
	(0.147)	(0.082)	(0.084)
Intercept	1.595***	0.446***	0.562***
	(0.086)	(0.058)	(0.058)
Observations	144	144	142
R^2	0.001	0.011	0.000
Adjusted R^2	-0.006	0.004	-0.007
Residual Std. Error	0.872 (df=142)	0.490 (df=142)	0.499 (df=140)
F Statistic	0.109 (df=1; 142)	1.598 (df=1; 142)	0.002 (df=1; 140)
Note:			*p<0.1; **p<0.05; ***p<0.01

Fig. 3

The coefficient for Any_Treatment is 0.004 with a p-value of 0.084. This suggests a very small and statistically insignificant difference in the means of Gender_Bool between the Any_Treatment groups. Based on these balance tests, there are no large or statistically significant imbalances in the observed characteristics between the Any_Treatment groups for Age, Attend_BU, and Gender. The results suggest that the groups are reasonably balanced with respect to these variables.

4.6.2 Balanced Check - AI Images

```
Z-test statistic w.r.t. most shown image: 1.1320898654460456
P-value w.r.t. most shown image         : 0.25759665519286223
```

Fig. 4

As the p-value of 0.25 is greater than 0.05, this suggests that the proportion of the most frequently shown image is not significantly different from the expected proportion.

```
Z-test statistic w.r.t. least shown image: -1.1426756992799223
P-value w.r.t. least shown image:         0.25317326285880204
```

Fig. 5

As the p value is 0.25 greater than 0.05, suggest that the proportion of the least frequently shown image is not significantly different from the expected proportion. As neither the most frequently nor the least frequently shown image were statistically different from the expected proportion, we can conclude that all images were displayed in a balanced and well randomized manner.

5. Outcome and How it was Collected

The outcome collected is the numerical ratings of the AI-generated images from the surveyors. Each image was accompanied by a question header “On a scale of 1-100, how much do you like this photo?”, and a slider scale going from 0 to 100. After collecting the data, we reorganized the data to reflect individual scores of each image a user saw and included a user’s overall average score across all 10 images.

6. Exploratory Data Analysis

After collecting our data, we conducted a brief EDA to explore initial findings and to do a quick gut-feel randomization check. We find that scores trended high, with overall trends remaining consistent among both treatment and control groups.

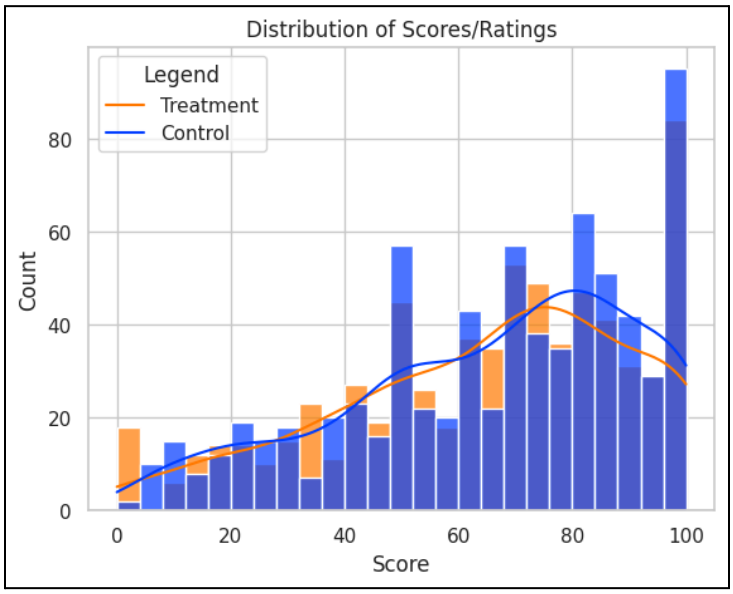

Fig.6

Looking into the age distribution, we see that most of our respondents were younger, falling into the 18-24 and 25-34 range. This is likely as most of the respondents to our survey were our friends or people on the street of the BU campus, which will trend young. Distribution between control and treatment were relatively even as well.

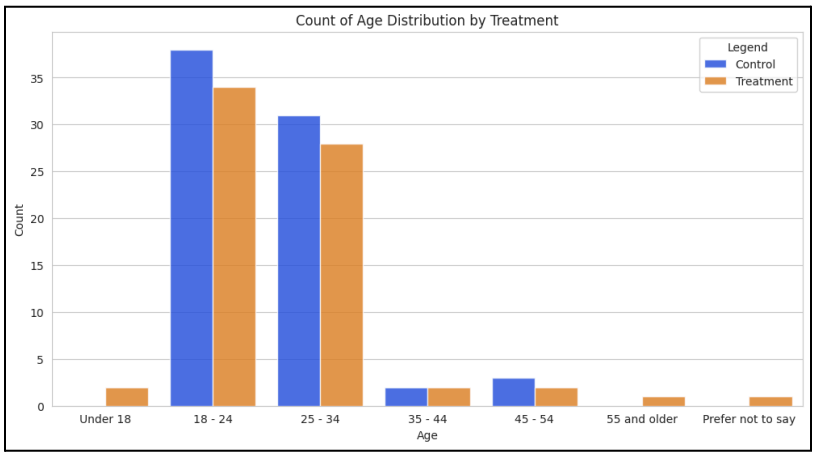

Fig.7

Gender split looked good as well, evenly distributed between treatment groups among all options available. Slightly more women than men responded.

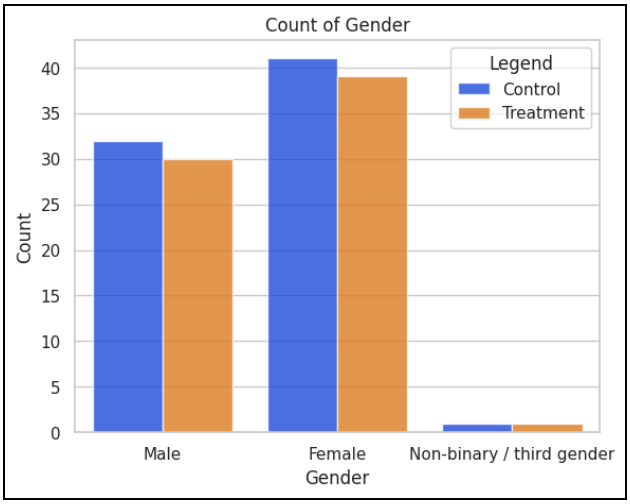

Fig.8

Interestingly, we had more non-BU attendees than BU attendees. Out of the BU attendees, unsurprisingly, most of the respondents hailed from Questrom.

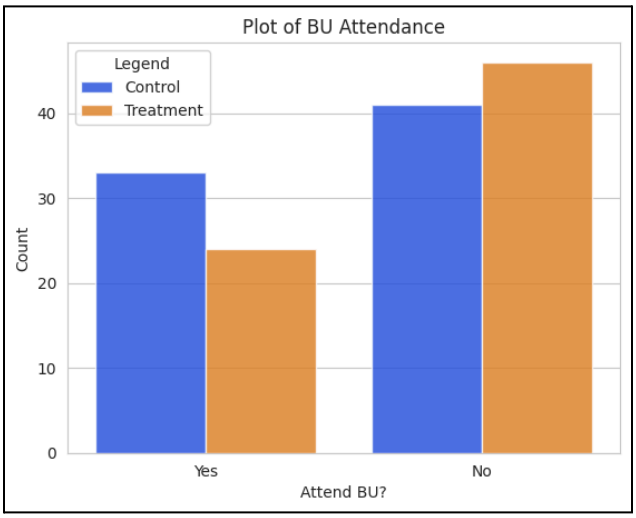

Fig. 9

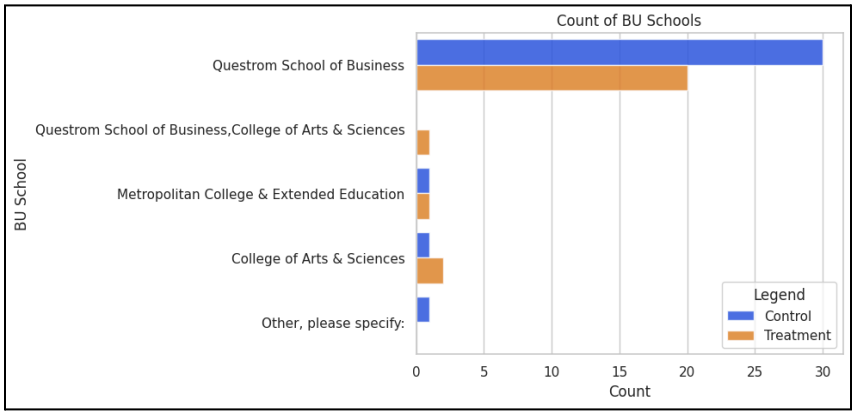

Fig. 10

Looking at scores distributed by image ID, we see that generally, the image scores did not stray very far from each other. Certain images, like image 11, 12 and 13 (see appendix) have scores that are more tightly distributed higher than other images. Still, there were no stand out first or last place performers, which is a good sign to show that the images were generally comparable in terms of appeal across the survey.

7. Results and its Interpretation

7.1 Regression Analysis on Score

	est1	est2	est3
depvar	Score	Score	Score
Intercept	69.806*** (5.725)	69.595*** (5.724)	
Any_Treatment	-2.468 (1.311)	-2.032 (1.298)	-2.468 (1.301)
Age[T.18 - 24]	-2.318 (4.402)	-4.428 (4.367)	-2.318 (4.368)
Age[T.25 - 34]	-1.953 (4.412)	-3.042 (4.391)	-1.953 (4.378)
Age[T.35 - 44]	3.297 (5.406)	2.419 (5.432)	3.297 (5.365)
Age[T.45 - 54]	-16.647** (5.149)	-17.247*** (5.128)	-16.647** (5.109)
Age[T.55 and older]	7.273 (5.138)	7.349 (5.110)	7.273 (5.098)
Age[T.Prefer not to say]	-4.206 (7.926)	-8.740 (7.798)	-4.206 (7.865)
Gender[T.Male]	4.216** (1.295)	4.269** (1.300)	4.216** (1.285)
Gender[T.Non-binary / third gender]	-36.302*** (5.297)	-34.425*** (5.373)	-36.302*** (5.256)
Attend_BU[T.Yes]	-4.515** (1.378)		-4.515*** (1.367)
Image_ID[T.10]	-9.225 (4.810)	-9.629* (4.822)	
Image_ID[T.11]	6.204 (4.554)	5.657 (4.570)	
Image_ID[T.12]	13.259** (4.657)	12.722** (4.678)	
Image_ID[T.13]	9.862* (4.472)	9.391* (4.521)	
Image_ID[T.14]	-11.485* (4.996)	-11.564* (5.010)	
Image_ID[T.15]	5.016 (4.868)	4.697 (4.865)	
Image_ID[T.16]	-0.854 (4.590)	-0.900 (4.623)	
Image_ID[T.17]	1.853 (4.787)	1.745 (4.833)	
Image_ID[T.18]	-24.181*** (4.908)	-24.381*** (4.955)	
Image_ID[T.19]	8.707 (4.468)	8.374 (4.478)	
Image_ID[T.2]	2.444 (4.947)	2.264 (4.966)	
Image_ID[T.20]	2.185 (4.942)	1.781 (4.928)	
Image_ID[T.21]	-10.455* (4.643)	-10.537* (4.675)	
Image_ID[T.22]	3.017 (4.954)	2.843 (4.963)	
Image_ID[T.3]	6.561 (4.993)	6.296 (5.034)	
Image_ID[T.4]	2.483 (4.759)	2.341 (4.766)	
Image_ID[T.5]	-5.295 (4.840)	-5.584 (4.875)	
Image_ID[T.6]	-9.810* (4.780)	-10.222* (4.791)	
Image_ID[T.7]	3.882 (4.931)	3.513 (4.996)	
Image_ID[T.8]	-8.981 (4.742)	-9.107 (4.775)	
Image_ID[T.9]	3.264 (4.937)	3.006 (4.964)	
Image_ID	-	-	x
R2	0.158	0.151	0.158
S.E. type	hetero	hetero	hetero
Observations	1440	1440	1440

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 11

We utilized "Score" as our target variable, representing the subjective rating by each participant to the displayed image during the survey. This variable reflects the individual perception and appreciation of the image's quality among participants.

Our regression analyses employed various control variables, including age, gender, attendance at BU, and image ID (fixed effects), to account for potential confounding factors. This approach allowed us to isolate the specific effects of our variables of interest.

The findings across all three regressions consistently indicated that the treatment variable, which denotes whether participants were informed that the images were AI-generated, did not yield statistically significant differences in the final ratings. This suggests that participants' awareness of the images being AI-generated were not effective in their evaluation.

7.2 Conditional Average Treatment Effect

	est1	est2
depvar	Score	Score
Any_Treatment	-3.984* (1.823)	-2.095 (1.634)
Gender[T.Male]	2.174 (1.786)	4.251** (1.304)
Gender[T.Non-binary / third gender]	-25.089** (8.121)	-36.297*** (5.274)
Age[T.18 - 24]	-2.107 (4.220)	-2.086 (4.411)
Age[T.25 - 34]	-1.523 (4.231)	-1.787 (4.392)
Age[T.35 - 44]	4.655 (5.300)	3.388 (5.370)
Age[T.45 - 54]	-16.505*** (4.987)	-16.500** (5.118)
Age[T.55 and older]	8.401 (5.022)	7.290 (5.100)
Age[T.Prefer not to say]	-5.270 (7.794)	-3.676 (8.030)
Attend_BU[T.Yes]	-4.582*** (1.387)	-4.102* (1.758)
Any_Treatment:Gender[T.Male]	4.266 (2.658)	
Any_Treatment:Gender[T.Non-binary / third gender]	-22.372* (9.262)	
Any_Treatment:Attend_BU[T.Yes]		-0.959 (2.830)
Image_ID	x	x
R2	0.162	0.158
S.E. type	hetero	hetero
Observations	1440	1440

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 12

For the conditional average treatment effects, we used attended BU and gender to calculate conditional average treatment effects with fixed effects of Image id. There are no significant conditional treatment effects detected by the regression. Notably, for the non-binary group, we didn't have enough participants to indicate a significance.

Nevertheless, we can see that for CATE for gender, there is a statistical difference between users in the treatment and control as we can see from the star over Any_Treatment. However, this might be due to chance, since our confidence interval is very close to incorporating our result score.

7.3 Regression Analysis on Mean Score

	est1	est2	est3
depvar	Mean_Score	Mean_Score	Mean_Score
Intercept	70.860*** (3.655)	70.673*** (3.651)	
Any_Treatment	-2.163** (0.820)	-1.775* (0.821)	-2.163** (0.813)
Age[T.18 - 24]	-4.698 (3.162)	-6.574* (3.127)	-4.698 (3.137)
Age[T.25 - 34]	-3.543 (3.189)	-4.512 (3.177)	-3.543 (3.164)
Age[T.35 - 44]	2.251 (3.955)	1.470 (3.987)	2.251 (3.924)
Age[T.45 - 54]	-17.640*** (3.172)	-18.174*** (3.157)	-17.640*** (3.148)
Age[T.55 and older]	5.535 (3.119)	5.603 (3.113)	5.535 (3.095)
Age[T.Prefer not to say]	-8.072* (3.216)	-12.105*** (3.111)	-8.072* (3.191)
Gender[T.Male]	3.636*** (0.816)	3.684*** (0.822)	3.636*** (0.809)
Gender[T.Non-binary / third gender]	-35.577*** (2.586)	-33.909*** (2.707)	-35.577*** (2.566)
Attend_BU[T.Yes]	-4.015*** (0.838)		-4.015*** (0.832)
Image_ID[T.10]	0.024 (2.486)	-0.336 (2.512)	
Image_ID[T.11]	0.945 (2.593)	0.459 (2.621)	
Image_ID[T.12]	1.234 (2.673)	0.756 (2.690)	
Image_ID[T.13]	1.313 (2.513)	0.894 (2.557)	
Image_ID[T.14]	0.001 (2.702)	-0.069 (2.703)	
Image_ID[T.15]	2.018 (2.856)	1.734 (2.863)	
Image_ID[T.16]	0.568 (2.584)	0.527 (2.571)	
Image_ID[T.17]	-0.003 (2.804)	-0.099 (2.833)	
Image_ID[T.18]	-2.838 (2.590)	-3.016 (2.598)	
Image_ID[T.19]	0.949 (2.532)	0.653 (2.538)	
Image_ID[T.2]	-1.690 (2.680)	-1.850 (2.686)	
Image_ID[T.20]	0.890 (2.513)	0.530 (2.520)	
Image_ID[T.21]	0.525 (2.513)	0.452 (2.518)	
Image_ID[T.22]	-1.659 (2.715)	-1.814 (2.710)	
Image_ID[T.3]	0.517 (2.715)	0.281 (2.736)	
Image_ID[T.4]	0.330 (2.824)	0.204 (2.819)	
Image_ID[T.5]	1.438 (2.613)	1.180 (2.618)	
Image_ID[T.6]	-0.242 (2.642)	-0.608 (2.642)	
Image_ID[T.7]	0.890 (2.607)	0.562 (2.634)	
Image_ID[T.8]	-0.257 (2.609)	-0.370 (2.631)	
Image_ID[T.9]	0.806 (2.618)	0.577 (2.661)	
Image_ID	-	-	x
R2	0.136	0.123	0.136
S.E. type	hetero	hetero	hetero
Observations	1440	1440	1440

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 13

For regression regarding the mean-score, it is very similar to the regression model on score. Both regression models were designed and constructed exact the same. Instead of the score as the dependent variable, we here have the mean_score of the users as the dependent variable.

The regression analyses employed various control variables, including age, gender, attendance at BU, and image ID (fixed effects). Here, we can see that any_treamtent was statistically significant. These three coefficients mean the treatment (informed that all the arts are AI- generated) had a negative effect on the mean_score.

These results helped to support our hypothesis that knowing art that was generated by artificial intelligence would have negative effects on their perception of the appreciation of the art. Adding attended BU as a fixed effect had a negative effect on mean_score and both coefficients are statistically significant. This might be explained by the fact that BU students (our data mostly came from MSBA students) might be more aware of the controversies of AI-generated images and this fact might have a negative effect on their perception.

7.4 Conditional Average Treatment Effect

	est1	est2
depvar	Mean_Score	Mean_Score
Any_Treatment	-3.989*** (1.121)	-1.603 (1.084)
Gender[T.Male]	1.282 (1.021)	3.688*** (0.828)
Gender[T.Non-binary / third gender]	-25.286*** (1.085)	-35.570*** (2.615)
Age[T.18 - 24]	-4.434 (2.872)	-4.350 (3.178)
Age[T.25 - 34]	-3.116 (2.906)	-3.294 (3.176)
Age[T.35 - 44]	3.796 (3.750)	2.387 (3.928)
Age[T.45 - 54]	-17.505*** (2.898)	-17.420*** (3.164)
Age[T.55 and older]	6.839* (2.870)	5.562 (3.090)
Age[T.Prefer not to say]	-9.260** (2.925)	-7.276* (3.347)
Attend_BU[T.Yes]	-4.120*** (0.848)	-3.396*** (0.997)
Any_Treatment:Gender[T.Male]	4.918** (1.667)	
Any_Treatment:Gender[T.Non-binary / third gender]	-20.532*** (1.687)	
Any_Treatment:Attend_BU[T.Yes]		-1.439 (1.726)
Image_ID	x	x
R2	0.149	0.137
S.E. type	hetero	hetero
Observations	1440	1440

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001

Fig. 14

For the conditional average treatment effects on 'score', we used attended BU and gender to calculate conditional average treatment effects with fixed effects of Image id. There are no significant conditional treatment effects detected by the regression. Notably, for the non-binary group, even though the coefficient is statistically significant, we didn't have enough participants(only 3 participants) to indicate a significance.

8. Limitations

The duration of our experiment was restricted to 8 days, with data collection primarily concentrated within the initial 48-hour window following survey release. This may not be enough time to collect the magnitude of data we needed to.

Our sample size remained constrained, yielding 183 samples for analysis post data cleaning. This was not enough to reach the 973 samples needed for sufficient statistical power. This limitation necessitates careful consideration when assessing the robustness of our findings, particularly in calculating metrics such as Additional Treatment Effect, Conditional Additional Treatment Effect, and statistical power.

We focused on BU Master of Business Analytics students, distributing flyers at the Hariri Building and utilizing social media. However, our outreach mainly reached acquaintances and family, limiting diversity. A more representative sample may yield more useful results.

Lastly, due to the nature of the way we solicited responses, which included flagging down a group of friends in public for them to scan the QR code to the survey at once, it resulted in some spillover effects as some of them would make comments to each other about the survey, about how much they liked or disliked certain photos, etc.

9. Conclusion

In conclusion, using our final regression model where we control for age, gender, BU attendance, also adding in image ID as fixed effects, we do see that awareness of an image being AI generated decreases the rating scores when compared to the control group that didn't get the AI prompt. This shows that AI is still associated with negative attributes in the creative space, and affects people's judgment on their appreciation of art. Moving forward, as AI continues to improve, this trend is something we should further monitor to see if it will continue to intensify. However, it is key to remember that this study likely did not have sufficient statistical power due to the small sample size despite having statistically significant results so we must take the findings with a grain of salt.

Appendix

The 22 AI generated images used for the survey
https://drive.google.com/drive/folders/1aU_sEyIEjI9jzTUs90UrQU_0NCAzQnSy?usp=sharing

Survey Link
https://bostonu.qualtrics.com/jfe/form/SV_aXBbgDhPUMe09MO

Bibliography

1. Mikalonytė, Elzė Sigutė, 0000-0002-3047-7729View Profile, Markus Kneer Department of Philosophy, Markus Kneer, Department of Philosophy, Abbott Ryan, Ambrosio Chiara, et al. "Can Artificial Intelligence Make Art?: Folk Intuitions as to Whether AI-Driven Robots Can Be Viewed as Artists and Produce Art." ACM Transactions on Human-Robot Interaction, December 1, 2022. https://dl.acm.org/doi/10.1145/3530875.

2. Martin Ragot, Nicolas Martin, Nicolas Martin, Salomé Cojean Université d'Angers, Salomé Cojean, et al. "Ai-Generated vs. Human Artworks. A Perception Bias towards Artificial Intelligence?: Extended Abstracts of the 2020 Chi Conference on Human Factors in Computing Systems." ACM Conferences, April 1, 2020. https://dl.acm.org/doi/abs/10.1145/3334480.3382892?casa_token=96cFvMGMXpkAAAAA%3AehU4V4u9ytzrRi8UfNZAwEm6eFMzqLx8049_m4t4QOjlYeFQJcb0LWvFae0TQyzWQBHFzreYsIJJ-TU

3. Messingschlager, Veronika and Appel, Markus, "Mind ascribed to AI and the appreciation of AI-generated art", Sage Journals, September 26, 2023. Mind ascribed to AI and the appreciation of AI-generated art - Tanja Veronika Messingschlager, Markus Appel, 2023 (sagepub.com)