Valuing AI Art: An Experimental Study

Pre-analysis Plan

Final version: September 24, 2024

Abstract: In this study, we will investigate differences in individuals' willingness to pay for artwork which is either human- or AI- created. First, using an online survey experiment, we will identify a candidate sample of images, spanning a range of artistic styles, which are statistically indistinguishable in terms of humans' judgement about whether they are created by a human artist or AI program. Then, we will use laboratory experiment auctions to compare incentivized valuations for human- versus AI- created artwork.

Ethics approval: Ethics approval for the online survey experiment was obtained from Newcastle University (#50559/2023) on September 24, 2024. We will obtain ethics approval for the lab experiment before data collection commences.

This file includes:

I. Outcomes

Primary outcomes: bids (willingness to pay).

II. Selection of artwork

The process for selecting artwork to be auctioned in the laboratory consists of four stages:

1. *Select initial sample of human-created images*. We will first identify an initial sample of 100 images created by identified human artists, spanning a range of artistic styles. We will randomly select images for this initial sample using the open access art aggregator rawpixel,[1] conditional upon (1) the images having a CC0 license[2] (public domain), which waives any claims to copyright by the creator and permits unrestricted use of the images, including resale; (2) the images not containing artists' signatures; and (3) the images not being immediately recognizable to the researchers as famous works based on their personal knowledge.

2. *Generate initial sample of AI-created images.* The AI art will be created using Midjourney,[3] which enables the generation of original images from simple text prompts. Each image is created using one of two approaches. The first approach uses Anthropic's Large Language Model (LLM) Claude[4] to generate a description of an imaginary artwork, which is then used as a prompt in Midjourney's web application. The second approach "re-imagines" real artwork by randomly selecting a real work of art that is first uploaded to Midjourney, which then describes the subject of the art in text. The text is then used as a prompt along with the "Style reference" option.

3. *Determine candidate sample of images.* Based on steps 1 and 2, we will have an initial sample of 100 human-created art images and 100 AI-created images spanning a range of artistic styles. To identify a candidate sample of images to be auctioned in the laboratory whose source (human or AI) is undetectable, we will conduct an online survey. In this survey, participants will be shown 50 randomly selected unidentified images from the initial sample created (participants are informed all images are drawn from a database where half the images are human-created and half are AI-created; the order of image presentation is randomized). For each image, participants will be asked to evaluate what is the chance in percentage terms that the image was created by a human artist and what is the chance in percentage terms that the image was created by an AI-art generator. The chance of each source should be a number between 0 and 100 and the chances given to the two possible sources should add up to 100. The candidate sample is defined as those images (human or AI) for which we fail to reject the null hypothesis that the perceived chance of the image being human

[1] https://www.rawpixel.com/public-domain
[2] https://creativecommons.org/public-domain/cc0/
[3] https://midjourney.co/
[4] https://claude.ai/

created is 50%, based on a two-tailed t-test.[5] See Section III below for details of the online survey protocol and a power calculation.

As long as step 3 provides us with a candidate sample of at least 10 human-created and 10 AI-created images, we will proceed to step 4. If it provides us with fewer than 10 images of either type, we will add more images to the initial sample and repeat step 3 until we have a sufficient candidate sample.

4. *Select final sample of images*. So that the final sample spans a range of potential values, the candidate sample of images will then be taken forward to a second survey, run on a different sample of participants. In this survey, we will elicit hypothetical willingness to pay (WTP) for small prints of the candidate sample images (i.e., the physical format of the images that will be auctioned in the laboratory). Specifically, for each image, participants will be presented with a multiple price list of amounts in increments of £1 ranging from £0 to £15 but will not be informed about whether they are created by a human artist or AI program. Participants must select the largest amount that they would be willing to pay for each image from the list if a physical copy of it were to be made available for them to purchase right now as an A5 print. Each participant will rate a random sequence of 50 images selected from the candidate sample – or, if the candidate sample contains 50 or less images, they will rate all images. For each image, we will compute the average WTP across participants. For the set of human-created images in the candidate sample, we will select the image with the highest and lowest average WTP, find the 8 evenly spaced points on the monetary interval between them, and select the final 8 images as those images with average WTP closest to each point. Then, among the set of AI-created images in the candidate sample, we will select the 10 images closest in average WTP to each selected human-created image. This will form the final sample of 10 human-created and 10 AI-created images to be auctioned in the laboratory. We will deviate from this strategy only if the obtained final sample has such divergence in the WTP distributions for the human and AI images that the highest ranked human image is ranked below 6 or more of the AI images, or vice versa. In that instance, we will add more images to the initial sample and repeat steps 3 and 4 until we achieve a final sample with sufficiently balanced distributions.

5. *Elicit aesthetic rating for final sample of images*. To further control for differences in "quality" between the two final samples, the final samples of images will then be taken forward to a third and final survey, run on a different sample of participants. In this survey, each image in the final sample will be rated according to an aesthetic rating scale from 1 to 5, where 1 is "very unappealing" and 5 is

[5] We will deviate from this strategy only if, unexpectedly, the evaluations are skewed such that the entire sample of images (or the vast majority of it) is perceived as all being probably human or all being probably AI created, such that we have no, or very few, images which receive an average evaluation in the vicinity of 50%. In that case, we will select those images for which we cannot reject the null hypothesis that the perceived chance of the image being human created is equal to the mean perceived chance across all images in the combined sample. This would still achieve our main objective of selecting images where the perceived likelihood of the source being human/AI is similar between those which are actually human and those which are actually AI created.

"very appealing".[6] This measure will be used as a control variable in the data analysis (see section VII below). Each participant will evaluate all 20 images (in random order).

III. Online survey recruitment, protocol and sample sizes

We will conduct the online surveys on Prolific Academic.[7] To increase comparability with the lab sample (see section V below), participants for the online surveys will be invited from a selection of the Prolific Academic database which is defined as follows: (a) students, (b) located in the United Kingdom, (c) fluent in English, and (d) approval rate above 90%. Participants from the first survey are excluded from the second, and participants from the first and second survey are excluded from the third. To prevent retakes, we record the Prolific ID of all participants. The share of male and female participants will additionally be balanced.

For each survey, we will set up a Prolific study which directs participants to a common link. Using this link, participants are directed to complete a Captcha (to rule out bots) and to provide informed consent. Those participants who consent will then be directed to read the study instructions. Participants are then presented with the unidentified images from the initial sample as described above. Participants will complete an attention check question at a random position in the sequence of images; data from any participant who fails the attention check will not be used.[8] After all images have been presented, participants are asked to elaborate on the reasons for their decisions in an open-text response format.

The first two surveys are expected to last approximately 10 minutes and the third survey approximately 5 minutes. The fixed payment will be £1.50 for the first two surveys and £0.75 for the third, which equates to an hourly rate of £9, consistent with Prolific's fair payment principles.

Sample sizes: We require a total sample of 125 per image for the first survey, which based on a one-sample two-sided t-test yields 80% power to detect at the 5% significance threshold a difference of 6 percentage points from a mean perceived chance of 50%, with Cohen's d = 0.25. Thus, this survey's total required sample given an initial sample of 200 images across the human- and AI-created categories, 50 of which are rated by each participant, is 526 (given that 5% of observations will be dropped). We aim for approximately 125 observations per image in each of the second and third surveys, but the final sample sizes will be determined by eventual budget, depending on whether resampling turns out to be necessary in the first survey.

We will initially soft launch the first survey with 30 subjects to check the technical setup and understanding of the task. If this shows completion times that imply our payments

[6] A similar aesthetic rating scale is used by Kirk et al. (2009).
[7] https://www.prolific.com/.
[8] We will also exclude from the analysis the fastest 5% of responses, on the grounds that such respondents are probably not paying sufficient attention for their provided data to be reliable. Furthermore, we will exclude from the analysis any respondent who gives the same response to all questions.

are not complying with Prolific's fair payment principles, we will adjust the fees. If and only if the soft launch requires us to make any changes to the survey, we will not use the data obtained from it.

IV. Lab experiment treatments

We implement 4 treatments in the lab using a between-subjects design:

1. The artworks being sold are created by an AI program, and subjects are not informed of this.

2. The artworks being sold are created by an AI program, and subjects are informed of this.

3. The artworks being sold are created by human artists, and subjects are not informed of this.

4. The artworks being sold are created by human artists, and subjects are informed of this.

V. Lab experiment recruitment, protocol and sample size

We will recruit for the lab experiment from a broad participant pool of university students at the Experimental and Behavioural Economics Lab at Newcastle University. No subject can participate in more than one session. Students voluntarily register to receive emails to participate in studies through the recruitment platform ORSEE (Greiner, 2015) maintained by the laboratory.

In each session, subjects participate in a series of 10 2nd price Vickrey auctions (Vickrey, 1961). At the beginning of a session, subjects are matched into groups of 4 bidders. These groups remain fixed across rounds. Each subject is endowed with an income that they can bid (likely to be around £10-15, to be finalized based on the elicited hypothetical WTP values from the online study – we need to ensure that subjects' budgets typically exceed plausible WTP values for each item). In each round, a small physical print of one of the artworks (lots) from the relevant final sample (human- or AI-created depending on the treatment) is auctioned within each group. Each bidder is invited to observe the physical copy of the artwork before the auction and an image is also displayed on the computer screen. The order of lots is randomized across sessions. Subjects submit their bids for each lot using a sealed-bid format and no feedback is provided on auction outcomes within the group until the end of the session. Thus, the subject is the independent level of observation for bids/willingness to pay (which are theoretically equivalent in the Vickrey auction). After all of the auctions rounds are completed, one round is selected at random separately for each group. The subject who submitted the highest bid in the group in the selected round's auction wins the artwork and second-highest bid in the group is subtracted from the highest bidder's income.

The rules of the auction are explained to subjects at the beginning of each session, together with the strategic reasoning for why they should bid the highest amount of money that each artwork is worth to them. Providing this information is appropriate for

experiments (such as this one) in which the goal is to elicit homegrown values under the assumption that subjects know the dominant strategy property of the auction is to bid their value (see Harrison et al., 2004, for a discussion). To further facilitate learning of the dominant strategy, each subject will individually undertake a hypothetical training exercise in which they will receive feedback if they fail to select the dominant strategy.

At the end of each session, we will run a post-experimental survey, including questions about demographics, and subjects' views towards and experiences with AI and art (more information in Section VIII). We will also elicit open-ended responses about reasons underlying bidding decisions in the different treatments.

The target sample size is 240 participants split equally across the four treatments (15 groups of four bidders per treatment). This is determined by budget constraints at the time of writing.

VI. Hypotheses

The hypotheses to be tested in the lab experiment are as follows:

Hypothesis 1: *Bids in Treatment 2 will be lower than in Treatment 1*

Hypothesis 2: *Bids in Treatment 4 will be higher than in Treatment 3*

Hypothesis 3: *Bids in Treatment 4 will be higher than in Treatment 2*

We also expect that bids in Treatment 1 and Treatment 3 will not significantly differ. This is expected to occur by design (the artworks being of indistinguishable source and similar quality), and we do not make it a formal hypothesis.

VII. Statistical methods

Each hypothesis will be tested using both two-sample t-tests and two-sample Kolmogorov-Smirnov tests. For all treatment comparisons, we will run these tests using the within-subject average bid across all 10 lots as the outcome variable. For comparisons of treatments using the same images (Treatment 1 vs Treatment 2; Treatment 3 vs Treatment 4), we will also run separate tests for each lot, using the amount bid on that lot as the outcome variable. All tests are two-sided and will use the 5% significance threshold.

We will also conduct regression analyses at the individual level controlling for the correlates of outcomes below, accounting for the panel structure of the data and censoring of the outcome variables. The dependent variable will be the amount a subject bids for a given artwork, with treatment dummies as independent variables, and additional controls for the artwork's aesthetic quality and hypothetical WTP value. An additional model will control for the individual-level variables referred to in Section VIII below, in case any of these variables are by chance unbalanced across treatments.

VIII. Correlates of outcomes

We will conduct an exploratory analysis to check for heterogeneity in the treatment effects based on standard demographic variables (e.g., age, gender, socio-economic background, field of studies) and the following AI and art-related variables which will be elicited in the end-of-experiment questionnaire:

- Extent to which subject has used generative AI
- Favourability of attitude towards AI
- Knowledge of art

This heterogeneity will be explored using interaction terms in the regressions, as well as machine learning techniques, such as the Causal Forest.

IX. References

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114-125.

Harrison, G. W., Harstad, R. M., & Rutström, E. E. (2004). Experimental methods and elicitation of values. *Experimental Economics, 7*, 123-140.

Kirk, U., Skov, M., Hulme, O., Christensen, M. S., & Zeki, S. (2009). Modulation of aesthetic value by semantic context: An fMRI study. *Neuroimage, 44*(3), 1125-1132.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance, 16*(1), 8-37.