# Valuing AI Art: An Experimental Study

## Pre-analysis Plan

Final version: June 7, 2026

**Ethics approval**: Ethics approval for the pre-tests was obtained from Newcastle University (#50559/2023) on September 24, 2024. Ethics approval for the main experiment was obtained from Newcastle University (amendment to #50090/2024) on March 11, 2026. At Hamburg University of Technology, the ethics committee was notified of this experiment on March 13, 2026, in line with the committee's general approval of economic laboratory experiments that comply with the German Association for Experimental Economics Research ethics checklist (https://gfew.de/en-ethik).

**This file includes:**

## I. Outcomes

Primary outcomes: bids (willingness to pay).

## II. Pre-tests: Selection of items

The process for selecting the items to be auctioned in the main experiment follows the pre-testing procedures used for the lab experiment reported in Lane et al. (2025). But it consists of separate but parallel streams for two types of items: visual artworks printed on paper and haikus printed on mugs. For each, we collected source material created by humans and source material generated by AI. We will pretest the detectability of source, and then select the final items to be used in the main experiment.

### a. Visual artworks

The process for selecting visual artworks to be auctioned in the main experiment consists of five stages:

1. *Collection of initial sample of human-created images*. We aimed to collect a sufficiently large initial sample to allow the eventual selection of 10 human-created and 10 AI-generated images for use in the main experiment. To ensure that these artworks span a range of artistic styles as in the lab experiment, we collected artworks from different offline sources.

   First, we collected an initial sample of images created by art students. Students were invited to submit original images in A5-sized portrait format using a range of artistic styles (abstract and representational) and visual characteristics. We collected these images through a call for submissions and required, as conditions of inclusion, that (1) the images are created by the submitting student, who owns the rights to the work and agrees that the research team may use, share, reproduce, and resell the work for research purposes; (2) the images do not contain creators' signatures or any other text; (3) the images are not found using TinEye reverse image search (tineye.com) and (4) the images are not immediately recognizable to the researchers as famous works or close derivatives of famous works based on their personal knowledge.

   Second, we sourced additional artworks from Newcastle University's library archives. Instead of condition (1), we required artworks to hold a CC0 public domain copyright license; we maintained conditions (2), (3) and (4).

   Third, we sourced additional artworks from the collection of Kunsthalle Bremen. Again, we required artworks to hold a CC0 public domain copyright license and maintained conditions (2), (3) and (4).

2. *Generate initial sample of AI-generated images.* The AI-generated images were generated using Midjourney,[1] which enables the generation of images from simple text prompts. Each image was generated using one of two approaches. The first approach uses Anthropic's Large Language Model (LLM) Claude (Sonnet 4.6) 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 images by randomly selecting a human-created image that is first uploaded to Midjourney, which then describes the subject of the image 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 collected an initial sample of human-created and AI-generated images. To identify a candidate sample of images whose source (human or AI) is unlikely to be detected, we will conduct an online experiment (the first pre-test). In this experiment, participants will be shown 50 randomly selected unidentified images from the initial sample (participants are informed that half of the images are human-created and half are AI-generated; the order of image presentation is randomized).

   Participants will be randomly assigned to one of two treatment arms. In the first arm, source evaluations are elicited without incentives (as in the pre-test for the lab experiment). In the second arm, source evaluations are elicited under incentives using a binarized quadratic scoring rule (e.g. Healy and Leo 2025). For each image, participants will be asked to evaluate what is the chance in percentage terms that the image was created by a human and what is the chance in percentage terms that the image was generated by AI. The chance of each source must be a number between 0 and 100 summing up to 100. In a first wave, we will collect the same amount of observations for both treatment arms. After the first wave, we compare outcomes between treatment arms. Specifically, (i) at the image-level we compare whether or not images would be included in the candidate sample under the incentivized and unincentivized treatment arm, respectively, and (ii) at the subject-level we compare the mean error-rates of predictions that the image was human-created. This allows us to assess the extent to which the two treatment arms yield similar candidate-item selections and whether incentivization leads to systematically different discriminating source evaluations. Should the incentivized treatment arm yield significantly lower mean error-rates based on a two-tailed t-test at the 5% significance level, incentives are preferred and we will collect additional observations for a second wave using the incentivized method. Otherwise, data for the second wave will be collected without incentives. Then based on the data from the preferred treatment arm and the second wave, the sample is defined as those images (human-created or AI-generated) for which we fail to reject the null hypothesis at the 5% significance level that the perceived chance of the image being human-created is 50%, based on a two-tailed t-test.

   See Section III below for details of the first pre-test protocol and a power calculation. As long as step 3 provides us with a candidate sample of more than

[1] https://midjourney.co/

10 human-created and 10 AI-generated images, we will proceed to step 4. If it provides us with 10 or fewer 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 an online survey (the second pre-test), run on a different sample of participants. In this online survey, we will elicit hypothetical willingness to pay (WTP) for visual artworks based on the candidate-sample images. 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 the image was created by a human or generated by AI. 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 (the artwork). Each participant will rate a random sequence of 50 images selected from the candidate sample – or, if the candidate sample contains 50 or fewer 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-generated images in the candidate sample, we will select one unique image closest in average WTP to each selected human-created image. This will form the final sample of 10 human-created and 10 AI-generated images for use in the main experiment. 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 appeal between the final sample of 10 human-created and 10 AI-generated images, these images will then be taken forward to an online survey (the third pre-test), 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 "very appealing". 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).

**b. Haiku mugs**

The process for selecting haiku mugs to be auctioned in the main experiment follows the same basic procedure as for the visual-artwork stream and consists of the same five stages and three pre-tests.

1. *Collect initial sample of human-created haikus*. Instead of collecting from offline sources, we collected an initial sample of 100 human-created haikus from Project

Gutenberg.[2] We required that the haikus are in the public domain and are not immediately recognizable to the researchers as unusually famous works based on their personal knowledge.

2. *Generate initial sample of AI-generated haikus*. Instead of generating images, we generated AI-written haikus using Claude (Sonnet 4.6). We used two approaches. In the first approach, Claude generated original haikus. In the second approach, Claude was shown the Project Gutenberg haikus and instructed to generate haikus that match their style. Together, these approaches provide the initial sample of AI-generated haikus.

3. *Determine candidate sample of haikus*. As in the visual-artwork stream, we will conduct the first pre-test to identify a candidate sample whose source is unlikely to be detected. Participants will be shown randomly selected unidentified haikus from the initial sample and will evaluate the probability that each haiku was human-written or AI-written. Unlike in the visual-artwork stream, source evaluations will be elicited without incentives only, as the human-written haikus are sourced from the public domain. If the unincentivized treatment arm in the artwork stream turns out to be more discriminating (or not significantly different), we will use the same procedure for source detection of haikus with a 5% significance level as threshold. Should the incentivized treatment arm be more discriminating, we will use the image-level data to calibrate the rejection threshold for the unincentivized treatment arm. More specifically, the candidate sample will then be determined using a calibrated p-value threshold $c^*$ derived from the visual-artwork stream. This threshold will be determined by selecting, among possible p-value thresholds for the unincentivized arm, the threshold that yields candidate-item classifications as similar as possible to those obtained in the incentivized arm using the 5% significance level. More precisely, let $p_j^I$ and $p_j^U$ denote the p-values obtained by a t-test for image $j$ in the incentivized and unincentivized treatment arms, respectively. Then the new threshold $c^*$ is the smallest $c$ that maximizes
$$\frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\left(\mathbf{1}\{p_j^U < c\} = \mathbf{1}\{p_j^I < 0.05\}\right).$$

4. *Select final sample of haikus*. As in the visual-artwork stream, the candidate sample will be taken forward to an online survey (the second pre-test) in which we elicit hypothetical willingness to pay for mugs printed with the candidate haikus.

5. *Elicit aesthetic rating for final sample of haikus*. As in the visual-artwork stream, the final sample will be taken forward to an online survey (the third pre-test).

## III. Pre-tests: Recruitment, protocol and sample sizes

We will conduct the pre-tests on Prolific Academic.[3] To increase comparability with the main experiment (see section V below), participants will be invited from the Prolific

[2] https://projekt-gutenberg.org/
[3] https://www.prolific.com/

Academic database using the regional representative sample feature for the adult population in the United Kingdom (implying they are fluent in English based on Prolific's screening). Prolific samples participants based on age, gender and regional quotas.[4] Furthermore, participants who participated in the pre-tests for the lab experiment are excluded and participants can only participate in one of the pre-tests described above. To prevent retakes, we record the Prolific ID of all participants.

For each pre-test, we will set up a Prolific study which directs participants to a common link. Using this link, participants are directed to complete a Captcha 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 or haikus from the initial sample as described above. Participants will complete an attention check question at a random position in the sequence of images or haikus; data from any participant who fails the attention check will not be used. At another point, we will also use a video attention check based on that of Celebi et al. (2026); data from any participant who fails this will also not be used.

In addition, we use Prolific's feature for bot screening available in combination with Qualtrics.[5] Any participant who is flagged as having "Low" or "Mixed" authenticity will be rejected.

From the valid responses, we will exclude 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.

After all items have been presented, participants are asked to elaborate on the reasons for their decisions in an open-text response format. They will be asked to not answer this using generative AI. Answers will be screened with Prolific's LLM checker and any participants flagged with "Low" authenticity will be rejected.

The first pre-test is expected to last approximately 15 minutes, the second pre-test approximately 10 minutes, and the third pre-test approximately 5 minutes. We aim for fixed payments close to an hourly rate of £9, consistent with Prolific's fair payment principles. For the visual artworks, the fixed payment will be £2 for the first pre-test, £1.50 for the second, and £0.75 for the third.

Note that for the incentivized treatment arm for detecting AI-generated images we add an average bonus payment of £0.56 announced only on the consent form. A bonus of £0.75 will be paid for use with a binarized quadratic scoring rule. We ask participants to provide their belief $p$ about an image being AI generated. If that is the fact, they receive a lottery that pays the bonus with probability $1-(1-p)^2$. If the image is instead created by a human they earn the bonus with probability $1-(0-p)^2$. Thus, bonus payments are £0 or £0.75 while a prediction of 50% yields an ex ante expected payment of £0.56.

[4] https://researcher-help.prolific.com/en/articles/445161-what-are-representative-samples-on-prolific
[5] https://www.prolific.com/resources/introducing-authenticity-checks-beta-ensure-genuine-human-responses-in-the-age-of-ai

Since the haiku pre-tests will be done without incentives, the fixed payment for the first pre-test will be raised to £2.25, while for the second and third pre-tests it will be the same as for the visual artworks.

*Sample sizes*:

In the unincentivized pre-test for the lab experiment, the average reported probability for an image being created by a human was 55.85 (standard deviation 28.89). Assuming the same standard deviation for the incentivized elicitation for images and the unincentivized elicitation for haikus, we require a total sample of 184 observations per item when aiming to detect a difference of 6 percentage points from a mean perceived chance of 50%. This is based on a one-sample two-sided t-test yielding 80% power to detect at the 5% significance level (implying Cohen's d = 0.21). Thus, for one type of item, the first pre-test requires a total of 736 participants given an initial sample of 200 items across the human- and AI-generated categories, 50 of which are rated by each participant. Given that 5% of observations will be dropped, we must recruit 775 participants for the preferred elicitation method. For mugs, we will simply recruit 775 participants. For images, because we must first recruit an extra comparison arm in Wave 1, we will recruit a total of 1041 participants, of which a total of 775 will receive the preferred elicitation method..

For images, the sample will be collected in two waves. For comparing error rates under incentivized and unincentivized elicitation for images in wave 1, we require a sample of 253 subjects aiming for Cohen's d = 0.25. This calculation is based on the mean error in participants' predictions in the unincentivized pre-test for the lab experiment, which was 43.54 (standard deviation 5.78), and on a two-sample two-sided t-test with 80% power to detect a difference at the 5% significance level. These participants will report probabilities on a randomly determined subset of 68 images, so that each image in this subset can also be rated 184 times. This subset will then be used to obtain the calibrated p-value threshold $c^*$. Given that 5% of observations will be dropped, we will recruit a sample of 266 in both treatment arms of wave 1. In wave 2, we recruit the remaining 509 participants.
In the pre-tests for the lab experiment, we aimed for 125 observations per item in each of the second and third pre-test. In order to meet Prolific's minimum of 300 participants for using the representative sample feature, we increase this to 150 for each type of item.

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

Table 1: Summary of pre-tests

| Pre-test | Stage | Visual artworks | Haiku mugs |
|---|---|---|---|
| 1 | Determine candidate sample | Wave 1:<br>N = 266 (unincentivized)<br>N = 266 (incentivized) | N = 775 (unincentivized) |

| | | | |
|---|---|---|---|
| | | Wave 2:<br>N = 509 (incentivized or unincentivized) | |
| 2 | Select final sample | N = 150 | N = 150 |
| 3 | Elicit aesthetic rating | N = 150 | N = 150 |

## IV. Main experiment: Treatments

We implement four treatments using the representative online sample provided by Pureprofile in a between-subjects design:

1. The visual artworks being sold are based on AI-generated images, and subjects are informed of this.

2. The visual artworks being sold are based on human-created images, and subjects are informed of this.

3. The haiku mugs being sold are based on AI-generated poetry, and subjects are informed of this.

4. The haiku mugs being sold are based on human-created poetry, and subjects are informed of this.

## V. Main experiment: Recruitment, protocol and sample size

The sample for the main experiment will be provided by Pureprofile.[6] The sample will be stratified to match the composition of the sample collected for detection of AI-generated images and haikus based on age, gender and region. Data will be collected using Qualtrics.

Subjects participate in a series of 10 2nd price Vickrey auctions (Vickrey, 1961). Subjects are informed that they will be matched (ex post) into groups of 4 bidders for determining auction outcomes. These groups remain fixed across rounds. Each subject is endowed with an income of £12 that they can bid. In each round, one item – either a visual artwork or a haiku mug – from the relevant final sample (human-created or AI-generated depending on the treatment) is auctioned within each group. We will vary between subjects whether they participate in auctions of a visual artwork or a haiku mug. An image of the artwork or mug is displayed on the respective Qualtrics screen. The order of lots is randomized across participants. Subjects submit their bids for each lot using a sealed-bid format and no feedback is provided on auction outcomes within the group until after 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 auction 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

[6] https://business.pureprofile.com/

the item 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, together with the strategic reasoning for why they should bid the highest amount of money that each item 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, participants participate in a post-experimental survey, including questions about demographics, and subjects' views towards and experiences with AI and art (more information in Section VIII).

Before the main experiment, participants are directed to complete a Captcha and to provide informed consent. Those participants who consent will then be directed to read the study instructions. Participants will complete an attention check question at a random position in the sequence of lots; data from any participant who fails the attention check will not be used. At another point, we will also use a video attention check based on that of Celebi et al. (2026); data from any participant who fails this will also not be used.

In addition, we use Qualtric's feature for bot detection.[7] Based on consultation with the survey provider, data from any participant with a "Q_RecaptchaScore" below 0.5 will not be used. Pureprofile excludes participants they identify as duplicate respondents, and those showing suspicious device behaviour, VPN/proxy usage or automated traffic patterns.

From the valid responses, we will exclude from the analysis the fastest 5% of responses within each treatment, on the grounds that such respondents are probably not paying sufficient attention for their provided data to be reliable.

The target sample size is 252 participants per treatment. This is based on an effect size of Cohen's d = 0.25 and the pooled standard deviation of 2.33 observed across the Human Info and AI Info treatments in the lab experiment and rounding to the closest multiple of four. Given that 5% of observations will be dropped, we recruit a total 1068 participants with valid responses. Again, we assume a two-sample two-sided t-test with 80% power to detect a difference at the 5% significance level.

## VI. Main experiment: Hypotheses

We hypothesize that informing participants that an item is based on AI generation lowers bids relative to informing them that it is based on human creation. We expect this effect to arise both for visual artworks and for haiku mugs.

[7] https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

Thus, the hypotheses to be tested in the main experiment are as follows:

**Hypothesis 1**: *For visual artworks, average bids are lower in the AI Info treatment than in the Human Info treatment.*

**Hypothesis 2**: *For haiku mugs, average bids are lower in the AI Info treatment than in the Human Info treatment.*

## VII. Main experiment: 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. 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 item, with treatment dummies as independent variables, and additional controls for the item's appeal 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.

These regressions will be conducted separately for the artworks and mugs to test our hypotheses. We will also run pooled models combining the artwork and mug data to check for an overall effect of AI production on bids across the two item types.

## VIII. Main experiment: 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/poetry

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

## IX. References

Celebi, C., Exley, C., Harrs, S., Kivimaki, H., Serra-Garcia, M., & Yusof, J. (2026). Mission possible: The collection of high-quality data. *Working paper,* https://www.ifo.de/cesifo/9qP.

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

Healy, P. J., & Leo, G. (2025). Belief elicitation: A user's guide. In: *Handbook of Experimental Methodology* (Vol. 1, No. 1, pp. 81-162). North-Holland.

Lane, T., Pickard, H., & Walker, M. J. (2025). No silver lining: Consumer indifference between human and AI production. *Working paper*, https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5184807.

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