Pre-analysis Plan - AI vs. Human Writing

I. Background

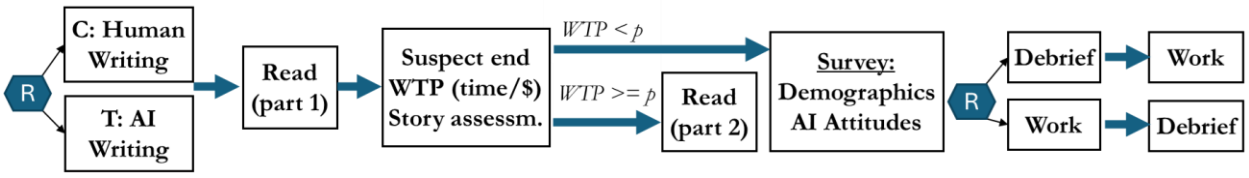
There is a burgeoning literature on how people perceive AI and how people value AI-generated content. Typical studies ask research participants how much they value either human or AI generated content and why. While these are important questions, they cannot isolate the difference in product from the identity of the creator of that product. Put differently, these studies do not answer the question of whether people value human created content differently if AI is able to produce content that is indistinguishable from human creations. As the ability of AI in both cognitive and artistic tasks rapidly increases this question becomes of increasing relevance.

Our study combines elements from economic and literary theory to investigate this and related questions. From economics, we borrow traditional welfare frameworks to test whether the identity of the creator affects the willingness to pay and thus shifts the demand curve of consumers. This is important as it determines how much consumer surplus people derive from consuming a good. Given the sharp decrease in (marginal) costs to produce many goods with AI, a similar willingness to pay for human and AI generated content would suggest a sharp increase in the welfare created on markets, assuming there are no externalities or other adverse effects not priced into the cost of AI. To investigate why people may value whether an identical product is human generated, we employ a framework from literary analyses that distinguishes several channels of how people may derive value from a narrative.

II. Research Questions

R1: Do people value a product differently if it was created by AI?
R2: What is the mechanism why people value it (not) differently?
R3: Do people think that AI manipulates the content of what it generated?

III. Research Design



The figure summarizes the research design which will be implemented using the survey software Qualtrics. We will next discuss the different stages of the design and how they help answer our research questions R1-R3.

i. Sample and Recruitment

We plan to collect a nationally representative sample of 700 participants located in the United States using the online platform Prolific. As part of the recruitment, we inform participants that we are studying what people value about writing. We further inform them that the study takes around 12 minutes and that they will earn $2.50.

ii. Randomization

After giving consent to participate in the study, participants are randomized in equal share into one of two groups: a “Human Writing” group and an “AI Writing” group. Both groups first read about a creative writing professor named Jason Brown.

Next, both groups are told they will next read an unpublished short story. This story was created by a large language model (GPT4) asked to create a story that is representative of the work of Jason Brown. (The author agreed that we can use his name in our research design.) Only the AI group is informed that the story was written by the LLM. This design allows us to test whether differences in the perception of the identity of the writer will lead to differences in valuation of the story.

iii. Willingness to pay and writing assessment

The short story is about a college professor who grapples with the question whether to let students use AI and whether to employ AI for his own writing. About two thirds into the story, we interrupt the reading and ask participants how they suspect the story will continue and whether the professor will decide on the use of AI. Responses to these questions allow us to test whether participants believe that AI generated stories are more likely to have different content, in this case that it is more likely to recommend the use of AI (R3).

Next, we collect participants' willingness to pay to continue reading the story. We employ a Becker-DeGroot-Marschak (BDM) mechanism, a common incentive-compatible method used in experimental economics to elicit participants' valuations for goods. Specifically, we elicit their willingness to pay in two dimensions: i) the share of a $0.50 monetary bonus they are willing to pay and ii) the time (0-6 minutes) they are willing to commit to a transcription task. We will then randomly choose a price category (time vs. money) and the price level in that category (0-50 cents for money, 0-6 minutes for time). If the willingness to pay is below that price, participants will not read the story and do not pay the price. If the willingness to pay is equal to above the price, participants get to read the end of the story. If the price category is money, we will deduct that price from their bonus. If the price category is time, they are asked to transcribe text at the end of the story (more details on this below). Comparing willingness to pay across groups allows us to answer R1.

After eliciting participants' willingness to pay, we ask them to assess the quality of the writing across several dimensions. Comparing differences across groups for different categories of evaluations helps shed light on the question of *why* people may value human written stories differently (R2).

iv. *Survey*

After people finish reading the story, we administer a survey in which we ask about participants' demographic characteristics such as age, gender, race, and education levels. We also ask about whether they think they would value a story differently depending on whether it was written by humans or AI and, if so, why. Last, we elicit their familiarity with and attitude towards AI.

IV. Data Analysis

I. Estimation

As our main specification, we estimate the following regression using OLS:

$$y_i = \alpha \ + \ \beta \ \ AI_i \ + \gamma \, X_i + \epsilon_i \qquad (1)$$

The dependent variable y_i measures the outcome for participant i. *AI_i* is an indicator variable measuring whether participant i is assigned to the AI writing group. The coefficient β_1 can thus be interpreted as the average effect of being assigned to the AI writing group. We will estimate specifications with and without controlling for X_i, which presents a vector of participant characteristics

To estimate whether the effect of the random assignment varies across subgroups, we estimate the following specification:

$$y_i = \alpha \ + \ \beta_1 AI_i \ + \ \beta_2 S_i + \ \beta_3 S_i * AI_i \ + \gamma \, X_i + \epsilon_i \qquad (2)$$

Variable S_i is an indicator variable measuring whether participant i is part of this subgroup. The coefficient β_1 can thus be interpreted as the average effect of being assigned a female advisor. The coefficient β_1 can thus be interpreted as the average effect of being assigned to the AI writing group for participants that are not part of subgroup i. And the sum of β_1 and β_3 is the effect of the AI treatment for members of subgroup i. In the same way, we will exploit the second stage random variation, when we interact the indicator of assignment to AI writing with an indicator of whether participants are debriefed before completing the transcription work.

ii. Outcomes

We will next specify the outcomes we analyze to answer our research questions. We follow recommendations by Bannerjee et al. (2020) and distinguish between primary outcomes and secondary outcomes, which are more exploratory in nature.

a) Willingness to pay (R1):
    i) Primary outcomes
        1) Monetary amount (in cents)
        2) Time (in minutes)
    ii) Secondary outcomes: binary measures for WTP > 0.

b) Attention (R2):
    i) Primary outcomes:
        1) Time people spent reading first part of the story (seconds, winsorized at 95th percentile)
    ii) Secondary outcomes:
        1) Binary measure whether people remember details of the story (1/0)

c) Assessment (R2)
    i) Primary outcomes:
        1) Standardized assessment index of categories below (higher values = positive assesment)
    ii) Secondary outcomes
        1) Each of the assessment categories separately
        2) Sympathy for professor (0=none, 1=some, 2=a lot)

d) Recommendation of AI (R3)
    i) Primary outcome
        1) Recommend use for student (-1/0/1) 0=not sure, 1=use AI, -1=not use AI
    ii) Secondary outcome
        1) Professor will use AI (-2/-1/0/1) [-2 = use not happy, -1=not use, 0=use, not sure, 1=use and happy]

e) Effort:
    i) Secondary outcomes
        1) Number of words transcribed in 60 seconds

f) AI preference (R2)
    i) Primary outcome:
        1) Measure whether people think they would value AI differently (-1/0/1)
    ii) Secondary outcome

1) Binary measures of the underlying reason why people claim to (not) value AI differently (1/0)

g) AI attitudes
   i) Secondary:
      1) excitement about AI
      2) curiosity about AI in different applications (books, music, learning, art)

h) Subgroups
   i) Primary subgroups:
      1) Familiarity with AI
      2) Education (completed college vs. no college)
      3) Age
   ii) Secondary subgroups:
      1) Gender
      2) Political leaning
      3) Number of books of fiction read