Pre-analysis Plan: AI vs Human Monitoring

Research Questions

R1: What is the effect of payments on intrinsic motivation and performance?

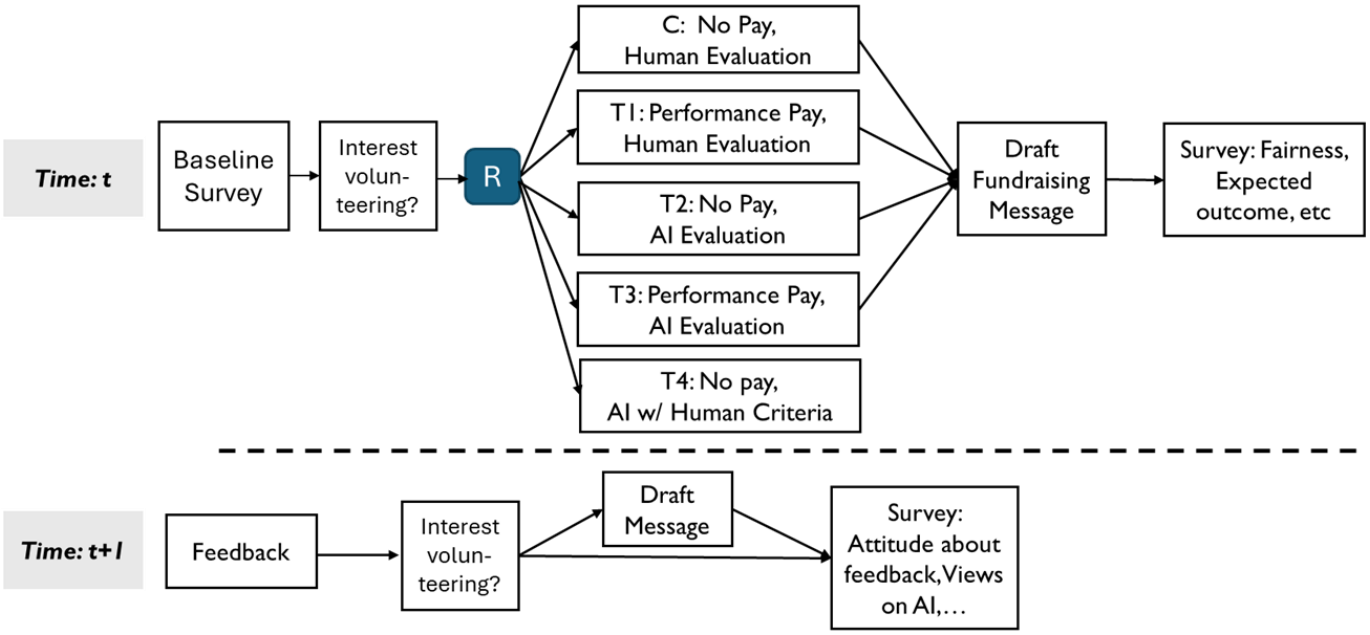
R2: What is the effect of AI vs. human evaluation on intrinsic motivation and performance?

R3: Does the effect of payments on intrinsic motivation / performance depend on whether humans or AI evaluate performance?

R4: What are the mechanisms behind these relationships?

R5: Does informing participants about human-like evaluation criteria mitigate biases against AI?

Study Design



Sampling: We will conduct an experiment with people in the U.S., recruited via Prolific. Participants are informed that the study investigates drivers of prosocial behavior. After participants agree to participate and we administer baseline questions, we ask them if they are interested in volunteering for a prosocial cause. These volunteers then compose the final sample. We aim for a sample size of around 1500 participants that choose to volunteer (the final sample depends on the share willing to volunteer.).

These volunteering participants are introduced to the prosocial cause of food insecurity through an informational video. They asked to create a 3-4 sentence fundraising message.

Randomization: Volunteering participants are randomized into one of five groups that vary in whether they receive payment for their message quality (payment / no payment) and in who will evaluate their message (people / AI / AI with stated human-like criteria).

- Control: participants receive no payment and have human evaluators
- T1 – participants receive no payment and have AI evaluators
- T2 – participants receive payment and have human evaluators
- T3 – participants receive payment and have AI evaluators
- T4 – participants receive no payment and have AI evaluators using human-like criteria.

Feedback: After the AI / humans completed their evaluation, we will invite all participants to a follow-up survey in which they learn about whether their message was evaluated to be: i) below average, ii) above average but not in top 20, iii) above average and in top 20.

After receiving the feedback, participants have a chance to provide campaign slogans as an optional task and they are asked about their perception of the evaluation process.

Key variables

For our analysis, we are next describing the i) measures of effort / intrinsic motivation, ii) mediators, and iii) moderators. We will specify a small number of primary variables that we will focus on in our analysis and additional variables that are more exploratory.

- Effort / intrinsic motivation
  - Primary:
    - **Quality of messages**: rated by humans on a 0 to 100 scale. (All messages will be evaluated by human evaluators and an AI system, but our primary outcome is based on human evaluations.)
    - **Donations**: amount donated by participants (0 to 50 cents)
  - Exploratory
    - **Time**: time participants spent writing the messages (winsorized at the 95th level)
    - **Spelling errors**: share of spelling errors in the message
    - **Attitude towards AI**: Response to the question whether they think AI will be a net positive or net negative to society?
    - **Optional task**: participation in voluntary task (submission of hashtags)

- Mediators:
  - Primary:
    - **Fairness / bias perception**: Response to statement: “I believe the evaluation process is fair and unbiased”
    - **Evaluation criteria perception**: response to question what criteria are being used to evaluate message → code as -1, 0, 1
  - Exploratory:
    - **Effectiveness perception**: Response to statement: “The evaluation process is transparent”
    - **Transparency perception**: Response to statement: “I believe the evaluation process will identify the most effective messages.”

        - **Perceived chance of being in top 20**
        - **Message content**: how does the actual content differ? We measure this through an AI-based evaluation of the content.

- Moderators / Subgroups:
    - Primary:
        - **Attitude towards technology**: Agreement with statement: Overall, do you think recent advances in technology have had a mostly positive or mostly negative impact on society?
        - **Experience with AI**, measured by the question: How regularly do you use AI?
        - **Baseline intrinsic motivation**: support for the cause of the fundraising campaign, measured by the question "In your opinion, how much should the US government prioritize spending resources on addressing food shortage over other important social problems? (0 = lowest priority, 100=highest priority)"
    - Exploratory:
        - **Locus of control**: measured by a standardized index comprising four questions.

Analysis Plan

I. Main Analysis

We will first conduct a simple OLS analysis of how the main outcomes *y* for participant *i* vary by treatment group. We will run the following regression:

$$(1) \qquad y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 T4_i + \delta X_i + \epsilon_i$$

For outcomes y, we will use the primary and exploratory outcomes for intrinsic motivation and mediators listed above. To analyze message quality ratings of participant *i* rated by human-rater *j*, we will run the following regression:

$$(2) \qquad Quality_{ij} = \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 T4_i + \delta X_i + \omega_j + \epsilon_{ij}$$

For both regressions, each beta coefficient measures if a given treatment group's outcome is significantly different from the control group (no payment, human evaluators). We will report results with and without controlling for a vector of participant control variables X_i . For the quality regression, we report results with and without rater fixed effects ω_j . We will use two-way clustering of standard errors $\epsilon_{ij}$by participant *i* and rater *j*.

II. Pooled Analysis

We plan to conduct pooled analyses to focus on the difference between payment and no-payment groups and human evaluator vs. AI evaluator groups. Pooling will maximize precision. We will exclude T4 from this analysis (no payment, AI evaluator w/ human criteria) given that we do not have a payment variation for this treatment. We will run:

$$(3) \qquad y_i = \alpha + \beta_1 AI_i + \delta X_i + \epsilon_i$$

Where AI_i is a dummy variable for whether participant *i* is in treatment group T1 or T3, who were told to have AI evaluators. The control group and T2 were told to have human evaluators. Furthermore, we will run:

$$(4) \qquad y_i = \alpha + \beta_1 Payment_i + \delta X_i + \epsilon_i$$

Where $Payment_i$ is a dummy variable for whether participant *i* is in treatment group T2 or T3.

We will then run a regression including an interaction term of payment and AI

$$(5) \qquad y_i = \alpha + \beta_1 Payment_i + \beta_2 AI_i + \beta_3 Payment_i * AI_i + \delta X_i + \epsilon_i$$

The analysis of letter quality will follow the specification described above.

III. Subgroup Analysis (Moderators)

To estimate heterogeneity of treatment effects, we estimate the following regression:

$$(6) \qquad y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \beta_3 T3_i + \beta_4 T4_i + \beta_5 T1_i * S_i + \beta_6 T2_i * S_i + \beta_7 T3_i * S_i + \beta_8 T4_i * S_i + \beta_9 * S_i + \delta X_i + \epsilon_i$$

For a given moderator variable listed above we will create an indicator variable S_i measuring the level of that moderator. We will also explore more parsimonious specifications in which we interact subgroup indicators with the pooled treatment arms discussed in part II.

IV. Analysis of Feedback

For the feedback analysis, we divide our sample into three groups based on how the messages are rated: below average, above average, above average and selected as one of the top 20. Given the small sample size of the latter group, we will exclude these 20 observations from the analysis. (As a robustness test, we will pool them with the above average group.)

Our main analysis follows the subgroup analysis discussed in part III with the indicator variable S_i measuring if the message was rated as above average (i.e. the participant received positive feedback).

Our main outcomes for the analysis of feedback are whether people complete the voluntary task (writing slogans), their perceptions about the fairness of the process, and whether they think humans or AI are better as evaluators. For exploratory outcomes, we use perceived effectiveness / transparency and whether they are interested in learning what an effective message is.