

Pre-Analysis Plan: How AI Agents Shift Online Interactions

Elliott Ash, Yichuan Lou, and Lena Song*

April 30, 2025

*Ash: ETH Zurich, ashe@ethz.ch. Lou: UTokyo, yichuanlou@e.u-tokyo.ac.jp. Song: University of Illinois Urbana-Champaign, lenasong@illinois.edu.

1 Introduction

In this project, we empirically assess the impact of artificial intelligence (AI) agents on the dynamics of information production and processing on social media. The substantive goal is to investigate how the presence of AI as senders alters the behavior and engagement of human users. Using survey experiments, our treatments randomly expose subjects to varying levels of the presence of AI agents. We will measure the effects of AI presence on human messaging investment and engagement behaviors. This approach allows us to isolate the effect of AI agent presence on online interactions and to evaluate potential disruptions caused by AI in digital communication environments.

2 Experimental Design

2.1 Recruitment

We will recruit around 1,000 ETH Zurich and University of Zurich participants via University Registration Center for Study (UAST) in 2025. The inclusion criteria are: aged 18+ and a current student at University of Zurich or ETH Zurich.

2.2 Experimental Design

The survey begins with a screener, informed consent, demographic questions, and baseline measures of attitudes toward climate change. Participants also respond to an open-ended question about climate change, which provides a baseline measure of key outcome variables (e.g., message length). In addition, participants report their social media usage and are given an opportunity to donate to a charity, providing a behavioral measure of prosocial tendencies related to climate change. To ensure shared understanding of key concepts, we include a brief explanation of the differences between traditional rule-based bots and conversational generative AI (Gen-AI) bots. We also elicit participants' use, familiarity with, and attitudes towards bots.

Participants then complete a receiver task, where they read two simulated social media posts and are asked to report their reactions and guess whether the author is a human, a rule-based bot, or a Gen-AI bot. The description of the message pool from which the posts are drawn (either consisting entirely of human-generated posts or a mix that is mostly AI-generated) varies across the two questions. In addition, the order in which the post texts are shown is randomized. This design allows us to measure how the labeling of the message pool influences participants' perceptions of the posts.

All participants complete an attention check, and those who indicate they use Instagram are asked to provide their handle.

The main task involves a writing prompt. Participants are instructed to submit a short text response, with the understanding that their response may be shared in a future online survey. They

are told that a future participant may read and react to their response.

The primary experimental manipulation is the composition of the message pool: participants are randomized into either a condition where their message will be pooled exclusively with human-written messages or a condition where it will be pooled with a mix of human and AI-generated content. This design allows us to examine whether the presence of AI alters how people choose to communicate.

To manipulate the strength of the signaling motive, we independently randomize participants into one of two identity conditions. In the identified condition, participants are told that—if they complete the writing task and provide their Instagram handle—their first name and handle may be shown alongside their response in the follow-up survey. In the anonymous condition, participants are informed that their response will be shared without any identifying information.

In both identity conditions, after submitting their message, participants are given the opportunity to pay to have their response removed from the message pool. We use a Becker-DeGroot-Marschak (BDM) mechanism to elicit an incentive-compatible valuation of how much participants are willing to pay to have their post removed. This enables us to assess how the presence of AI in the message pool affects participants’ willingness to keep their own message private.

At the end of the survey, we include questions designed to elicit potential mechanisms, such as the strength of participants’ signaling motives. We also implement several checks to ensure data quality, including measures to verify that participants remained engaged throughout the survey and did not rely on external AI tools to complete their responses.

2.3 Randomization

We randomize exposure to treatment for participants who satisfy the inclusion criteria. The main randomization is in the composition of senders. We randomize across the following compositions with equal probability:

- AI condition: 75% AI-generated posts and 25% student-written posts
- Human condition: 100% student-written posts

To manipulate the strength of the signaling motive, we independently randomize participants into one of two identity conditions, with 1/5 in the anonymous condition and 4/5 in the identified condition.

2.4 Outcome Variables

The primary outcome variables are:

- Time spent on the post (in seconds)

- Length of the message, measured by the number of characters
- Willingness to pay (WTP) to remove the post

In addition to these primary outcomes, we will explore several secondary measures related to participants’ perceptions, motivations, and writing characteristics:

- Perceived effectiveness of AI
- Perceived “AI-ness” of their own writing
- Importance placed on being perceived positively by others
- Beliefs about how others perceive their engagement with climate change issues
- Degree of “AI-ness” in the submitted text (using the state-of-art AI detection methods)
- Use of personal anecdotes in the written response
- Textual features, including overall length, average sentence length, and number of typos
- Emotional tone and argumentative content

Given the high dimensionality of text data, we will also use inductive, data-driven approaches to identify additional dimensions of language and style that may systematically vary across treatment conditions.

3 Empirical Strategy

3.1 Empirical Strategy

To estimate treatment effects, define Y_i as an outcome for participant i (e.g., time spent on the post, message length, or willingness to pay to remove the post). Define T_i as an indicator for whether participant i was assigned to the AI treatment condition, in which their message is pooled with AI-generated content. Define D_i as an indicator for assignment to the identified condition (Instagram handle and name may be disclosed to others). Define $\mathbf{X}_i$ as a vector of pre-treatment covariates that may improve precision and account for residual variation, including baseline attitudes toward climate change, charitable donation decision, and a baseline measure of the outcome of interest (derived from the participant’s earlier open-ended response about climate change). To estimate the causal effect of AI presence on participant outcomes, we will estimate the following linear regression model:

$$Y_i = \alpha + \beta T_i + \delta D_i + \mathbf{X}_i' \boldsymbol{\gamma} + \varepsilon_i \quad (1)$$

We will potentially control for age, gender, education, university, and country of origin in Appendix Figures if any of the covariates are unbalanced. We will calculate Lee bounds for treatment effects if there is differential attrition (Lee, 2009).

To study heterogeneous treatment effect, we will estimate the effects separately for different moderators, or include an interaction term. Specifically, we will examine heterogeneity based on:

- Baseline attitudes toward climate change
- Willingness to donate to a climate-related cause (as a measure of prosociality)
- Baseline perceived difference between one’s own writing and AI-generated content
- Image concerns
- Whether the participant shared their Instagram handle

We also plan to explore heterogeneity along additional dimensions, including:

- Prior familiarity with AI tools
- Demographic characteristics (e.g., age, education, political party)
- Baseline writing style or tone (e.g., emotionality, narrative vs. analytical style in the baseline response)

These exploratory analyses will help us understand for whom and under what conditions the presence of AI affects self-expression or willingness to share content.