Sharing Opinions in the Shadow of AI

Theory and Evidence from Online Interactions

Elliott Ash, Yichuan Lou, and Lena Song*

February 2026

Abstract

We study how the presence of artificial intelligence (AI) agents affects human content production in online interactions. Building on the signaling approach of Bénabou and Tirole (2006), we develop a theoretical framework of social interaction, in which a human sender with both social image concerns and disclosure concerns composes messages for a human receiver. The model characterizes conditions under which the presence of AI co-senders alters human behavior. It predicts that disclosure concerns reduce participation in content production, but that the presence of AI can increase participation by mitigating these concerns. We test the model's predictions using an online survey experiment. The results are consistent with the theory: the effect of AI presence on participation depends critically on whether the sender's identity may be disclosed. When identity disclosure is possible, the presence of AI increases participation, consistent with reduced disclosure concerns. In contrast, when senders remain anonymous, AI presence has no effect or reduces participation, potentially due to diminished signaling value. These effects are more pronounced for women than for men. Our findings contribute to ongoing debates about the societal impact of AI by showing that AI can reshape content production not only through machine-generated content, but also by altering the incentives and disclosure considerations of human contributors.

*Ash: ETH Zurich, ashe@ethz.ch. Lou: UTokyo, yichuanlou@e.u-tokyo.ac.jp. Song: University of Illinois Urbana-Champaign, lenasong@illinois.edu. We thank Sahil Adane and Amin Rahmati for excellent research assistance. The study was approved by Institutional Review Boards at University of Illinois Urbana-Champaign (IRB24-0588). This experiment was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0013548.

1 Introduction

In November 2022, ChatGPT, an online chatbot powered by generative artificial intelligence (Gen-AI), went viral. It demonstrated unprecedented abilities in understanding and generating natural language, engaging users with coherent and contextually relevant conversations. By May 2024, over a hundred million people were using ChatGPT every week.[1] This revolutionized public perception of AI's capabilities, sparked debates about its societal implications, and drew scrutiny from regulators. For example, in October 2023, the US government issued an executive order on the safe, secure, and trustworthy development and use of AI, noting guidance for AI management as "generative AI products become widely available and common in online platforms."[2]



"On the Internet, nobody knows you're a dog."

One type of online platform likely to be significantly impacted is social media (Aridor et al. 2024; Cookson, Mullins and Niessner 2024). There are concerns about AI's role in generating and spreading misinformation and toxic content online, and whether it poses a threat to election security.[3,4,5] At the same time, more optimistic narratives suggest that AI could make online conversations more civil.[6,7] Social media platforms appear to be cautiously embracing this change, incorporating generative AI. Meta, for example, introduced AI

[1] https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

[2] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[3] https://www.cbsnews.com/news/generative-ai-threat-to-election-security-federal-intelligence-agencies-warn/

[4] https://www.theatlantic.com/technology/archive/2023/05/generative-ai-social-media-integration-dangers-disinformation-addiction/673940/

[5] https://www.weforum.org/agenda/2024/01/ai-democracy-election-year-2024-disinformation-misinformation/

[6] https://hbr.org/2024/03/genai-could-make-online-conversations-more-civil

[7] https://www.newscientist.com/article/2398407-hundreds-of-chatbots-could-show-us-how-to-make-social-media-less-toxic

personas on WhatsApp, Messenger, and Instagram.[8]

This paper studies how the presence of generative AI alters human content production in online social interactions. As AI agents become embedded in social media environments and increasingly resemble human users, participation no longer cleanly signals that content is authored by a person. This ambiguity can change the incentives to contribute. On the one hand, AI may dilute the social image value of participation by weakening the inference that a contribution reflects a human's intrinsic motivation. On the other hand, the same dilution may reduce disclosure concerns by loosening the perceived link between content and personal identity. Whether AI discourages or encourages human participation therefore depends on how individuals trade off reputational benefits against privacy or disclosure costs when deciding whether to speak.

We first develop a parsimonious theoretical framework for social interactions, building on the signaling approach of Bénabou and Tirole (2006). In our framework, human agents coexist with two types of non-human agents: traditional rule-based bots and generative AI bots.[9] Human agents have private information about their intrinsic motivation to contribute and choose whether to participate in content production. Participation yields intrinsic utility but also generates belief-based payoffs that depend on how observers interpret the action.

A central feature of the model is that generative AI bots can take the same observable participation action as humans, making participation a noisy signal of human identity. As a result, participation may confer social image benefits when it is attributed to a human, but it may also impose disclosure costs when authorship is linked to personal identity. The presence of AI therefore has ambiguous effects on human participation. By diluting the inference that a participant is human, AI can weaken reputational incentives to contribute, but it can also reduce disclosure concerns by loosening the perceived connection between content and identity. Which force dominates depends on the relative strength of image and disclosure concerns.

In a survey experiment guided by our conceptual framework, we examine how the presence of AI agents affects human participation in opinion content production. Participants complete a writing task in which they are asked to compose a short post that may be shown to future respondents. We experimentally vary two dimensions of the environment. First, we vary whether participants' messages are pooled exclusively with other human-written posts

[8] https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/

[9] The presence of bots on social media is not new. Varol et al. (2017), for example, document the prevalence of automated accounts on Twitter in the 2010s. We distinguish between rule-based bots and technologically sophisticated conversational AI bots to isolate the unique effects of generative AI.

or with a mix that includes generative AI–generated content. Second, we vary whether participants' identity may be disclosed to readers or whether they remain anonymous. This design isolates how AI presence and identity disclosure jointly shape incentives to contribute content, holding fixed the topic, audience size, and absence of direct feedback.

We find that identity disclosure substantially reduces participation in content production, even at a late stage of a paid survey and when non-participation entails fully giving up payment. Strikingly, the presence of AI co-senders mitigates this effect. When participants are identifiable, introducing AI increases completion rates and the likelihood of producing a meaningful post. By contrast, when participants remain anonymous, AI has little or no positive effect on participation. These patterns are strongest among participants with higher disclosure concerns and among women. Together, the results support the model's central mechanism: AI reshapes human engagement not by replacing human expression, but by weakening the perceived link between authored content and personal identity.

These results have implications for how online interactions evolve as public awareness of AI presence on social media continues to grow. By late 2023, 63% of respondents reported having heard of ChatGPT (Loewen et al., 2024), reflecting a rapid shift in beliefs about the prevalence and capabilities of generative AI. Our findings suggest that the mere *perceived* presence of AI in online communication environments can meaningfully affect beliefs and behavior, even when content is produced by humans. When distinguishing between human and machine authorship is difficult, AI alters the social meaning of participation by weakening the link between authored content and personal identity. As a result, AI can either discourage or encourage engagement depending on the extent to which users face disclosure concerns. These effects arise independently of content quality and highlight the importance of perceptions and attribution in shaping online participation.

This paper contributes to the growing body of literature on content production on social media. As reviewed in Aridor et al. (2024), various nonmonetary incentives, including intrinsic or altruistic motives, reputation, attention, feedback, and persuasion, drive content production. For example, Srinivasan (2023) finds that randomly allocating AI-generated replies to Reddit posts increases content production of producers. Moreover, there are diminishing returns to adding AI replies, potentially because they are of lower quality or diversity than human replies. Instead of AI receivers, our paper focuses on the impact of AI presence among co-senders (more precisely, the perceived presence of AI) on human content production.

A related literature in experimental economics has explored human-AI interaction (Chugunova

and Sele, 2022). For example, Maréchal, Cohn and Gesche (2020) show that people are more likely to cheat when they are interacting with machines rather than humans. Bogliacino et al. 2023 find that when participants have the option to use ChatGPT-generated promises, they make promises more often but earn less trust, leading to outcomes that are comparable to those seen in human-only communications.

Building on Bernheim (1994) and Bénabou and Tirole (2006), our theoretical framework adopts a signaling approach to social interactions in which individuals care about how their actions are interpreted by others and therefore behave strategically to influence those beliefs. We extend this framework to an environment in which humans coexist with automated agents whose actions may be observationally indistinguishable from human behavior. A central feature of the model is that generative AI agents can take the same observable participation action as humans, thereby diluting the informational content of participation and weakening the inference that a contributor is human. Participation generates belief-based payoffs only when an action is attributed to a human: a social image benefit from being perceived as intrinsically motivated, and a disclosure cost arising from potential identity exposure. The presence of AI affects behavior by changing the posterior probability that a participant is human, which jointly scales both image benefits and privacy costs. As a result, AI has ambiguous effects on human participation. It can discourage contribution by reducing the signaling value of participation, but it can also encourage contribution by attenuating disclosure concerns when identity exposure is costly.

On the empirical side, reputational or social image concerns on social media have been shown to influence behaviors such as misinformation sharing (Guriev et al. 2023) and protests (Enikolopov, Makarin and Petrova 2020). These concerns have been incorporated into models of online misinformation (e.g., Acemoglu, Ozdaglar and Siderius 2023). Our paper explores how the rapid evolution of technology, specifically advancements that make machines indistinguishable from humans online, mediates these important drivers of human behavior.

The rest of the paper is structured as follows. Section 2 outlines the theoretical framework, while Section 3 describes and reports the results of the empirical design. Section 4 concludes.

2 Theoretical Framework

2.1 Model

The model builds on the large theoretical and empirical literature that explains prosocial behavior as the outcome of intrinsic motivation, participation costs, and social image concerns. We extend this framework to an environment in which human agents coexist with automated agents whose actions may be observationally indistinguishable from human behavior. Two features are central. First, we allow for non-human agents—generative AI and rule-based bots—that can take the same participation action as humans (e.g., posting content and contributing messages), thereby diluting the informational content of participation. Second, beyond standard social image concerns, we incorporate disclosure concerns: conditional on being perceived as human, participation may entail an additional utility cost arising from potential identity leakage. Together, these features capture how the presence of bot agents reshapes the incentives for human participation in prosocial activities.

Agents, types, and actions. There is a continuum of agents. Each agent belongs to one of the three identity groups: humans (h), generative-AI bots (g), and rule-based bots (r). Let (p_h, p_g, p_r) denote the population shares, with $p_h + p_g + p_r = 1$.

Human agents have private type $v \in [\underline{v}, \overline{v}]$, interpreted as intrinsic motivation to participate. The prior distribution of v is F, which we assume to be continuous. Each human agent makes a binary decision: $a \in \{0, 1\}$, where $a = 1$ represents participation and $a = 0$ represents abstention. Participation entails a cost $c > 0$. If a human agent abstains, she receives outside option $u_0 \in \mathbb{R}$. If she participates, she receives *direct utility* $v - c$ plus an indirect payoff that depends on how observers interpret participation.

Bots do not have private types. A generative-AI bot chooses $a = 1$ with probability $\alpha \in (0, 1)$ and takes a detectable action $a_g \notin \{0, 1\}$ with probability $1 - \alpha$. A rule-based bot always takes a detectable action $a_r \notin \{0, 1\}$. Detectable actions are interpreted as non-human and thus never pooled with human participation/abstention.

Information, beliefs, and payoffs. Observers see the realized action. If they observe a detectable action (a_g or a_r), they assign probability one to the agent being a bot. The interesting inference problem arises after observing participation $a = 1$, which can be generated by humans and, with probability α, by generative-AI bots. Let $P(h \mid a = 1)$ denote the perceived probability that a participant is human.

For human agents, participation has reputational consequences beyond its direct payoff. As in Bénabou and Tirole (2006), participation can serve as a signal of prosocial motivation and thus generate social approval: observers who attribute the action to a human may update upward their assessment of the agent's intrinsic motivation. At the same time, participation may expose the agent to unwanted inference about her identity or traits, giving rise to a reputational cost. In our setting, both the approval benefit and the privacy cost are belief-based and thus are relevant only insofar as the participant is believed to be human; if participation is attributed to a non-human sender, neither prosocial credit nor privacy exposure is at stake.

Let $\mu \geq 0$ denote the strength of image concerns and $\gamma \geq 0$ the strength of disclosure concerns. The *indirect/reputational utility* from participation for a human is

$$P(h \mid a = 1)\big[\mu E(v \mid a = 1) - \gamma\big]$$

Intuitively, participation yields an image benefit from being perceived as prosocial, but also entails a privacy cost from being exposed; both are scaled by the likelihood that the participant is human. Thus, a human of type v who participates obtains total utility

$$v - c + P(h \mid a = 1)\big[\mu E(v \mid a = 1) - \gamma\big] \tag{1}$$

If instead the human agent abstains, she receives her outside option with value u_0, which is independent of beliefs and carries no endogenous reputational payoff. Our key modeling assumption is therefore that indirect payoffs arising from social image and disclosure concerns are attached only to participation, not to abstention. In Bénabou and Tirole (2006)'s binary-action framework, each publicly observed action carries a reputational payoff, with "honor" from participating and "stigma" from abstaining. By contrast, in our setting abstention is interpreted as *non-observation*: if an agent does not post, there is no observable object for observers to assess and no individual-level identity inference triggered by the absence of content. Consequently, only participation generates image and privacy payoffs.

Using the payoffs derived above for participation and abstention, a human agent participates whenever $v \geq u_0 + c - P(h|a = 1)\big[\mu E(v|a = 1) - \gamma\big]$.

Timing and equilibrium. Nature first draws the agent's identity. If the agent is human, nature further draws the human agent's type v. The agent then chooses an action. Observers update beliefs using Bayes' rule whenever possible. A (Perfect Bayesian) equilibrium consists

of (i) a strategy for each human type and (ii) beliefs over identity and, conditional on being human, over type v, such that strategies are optimal given beliefs and beliefs are Bayes-consistent on path.[10]

Because only participation can be pooled between humans and generative-AI bots, the equilibrium characterization is conveniently expressed in terms of the participation cutoff v^* and the posterior probability that a participant is human. In any cutoff equilibrium with threshold v^*, human agents participate if and only if $v \geq v^*$, so $\mathbb{E}[v \mid a = 1] = \mathbb{E}[v \mid v \geq v^*]$. Define the *net image-privacy surplus* conditional on being perceived as human by $T(v^*) \triangleq \mu E(v \mid v \geq v^*) - \gamma$. Let $\rho \triangleq \frac{p_g}{p_h + p_g}$ denote the share of generative-AI bots among agents whose actions may be confounded with human participation. Under a cutoff strategy at v^*, Bayes' rule implies that the posterior probability a participant is human is $\pi(v^*) \triangleq P(h \mid a = 1) = \frac{(1-\rho)(1-F(v^*))}{\rho\alpha + (1-\rho)(1-F(v^*))}$. Given a candidate cutoff v^*, a human type v prefers participation if and only if $v - c + \pi(v^*)T(v^*) \geq u_0$.

Define the operator $\Phi(v) \triangleq v - c + \pi(v)T(v)$, where $\pi(v)T(v)$ captures the indirect payoff from participation through prosocial image and privacy channels. We say that participation decisions exhibit *strategic complementarity* if $(\pi(v)T(v))' < 0$ and *strategic substitutes* if $(\pi(v)T(v))' > 0$. To see the intuition, note that a marginal expansion of participation corresponds to a decrease in the cutoff v. When $(\pi(v)T(v))' < 0$, lowering the cutoff improves the indirect value of participation, as the participation pool becomes more favorable in terms of perceived human identity and expected type. As a result, an increase in participation raises the indirect payoff from participation, giving rise to strategic complementarity. When $(\pi(v)T(v))' > 0$, an expansion of participation instead lowers the indirect payoff, yielding strategic substitutability. The sign and magnitude of $(\pi(v)T(v))'$ further determine equilibrium stability. Since $\phi'(v) = 1 + (\pi(v)T(v))'$, the response of marginal types to a small expansion of participation depends on whether the induced change in the indirect payoff is strong enough to offset the decline in intrinsic motivation at the margin. When $(\pi(v)T(v))' < -1$, the the increase in the indirect payoff dominates the loss in intrinsic motivation, so marginal types in $[v^* - \mathrm{d}v, v^*]$, who initially preferred to abstain, now strictly prefer to participate. This self-reinforcing feedback leads to further participation

[10] Unlike in standard signaling games, off-the-equilibrium-path beliefs are immaterial in this environment and therefore need not to be refined à la Cho and Kreps (1987). When full participation ($a = 1$) occurs on path, a human agent's deviation to $a = 0$ does not generate any image or privacy payoff, so beliefs (over both identity and value v) following $a = 0$ are payoff-irrelevant. When no participation ($a = 0$) occurs on path, the action $a = 1$ remains on path because Gen-AI bots take $a = 1$ with positive probability. In this case, following a deviation by a human agent, $P(h|a = 1)$ remains zero, so the indirect payoff is again null.

and continued improvement of the participation pool, implying that only corner equilibria are locally stable. By contrast, when $(\pi(v)T(v))' > -1$, either complementarity is too weak or participation decisions are strategically substitutable, the improvement in the indirect payoff is insufficient to overturn the marginal loss in intrinsic motivation. As a result, types just below the cutoff continue to prefer abstention following a small expansion of participation, and the interior cutoff equilibrium is locally stable.

We henceforth impose $\Phi'(v) > 0$ on $(\underline{v}, \overline{v})$ to ensure a unique stable equilibrium. This condition is always satisfied when participation decisions are strategic substitutes, or when they are strategic complements but the complementarity is sufficiently weak so that the induced increase in the indirect payoff does not outweigh the marginal decline in intrinsic motivation.

Proposition 1. *There is a unique equilibrium:*

(i) If $u_0 \in (\Phi(\underline{v}), \Phi(\overline{v}))$, then $a = 0$ for $v \in [\underline{v}, v^*)$ and $a = 1$ for $(v^*, \overline{v}]$, where v^* uniquely solves $\Phi(v^*) = u_0$.

(ii) If $u_0 \geq \Phi(\overline{v})$, then $a = 0$ for all $v \in [\underline{v}, \overline{v}]$.

(iii) If $u_0 \leq \Phi(\underline{v})$, then $a = 1$ for all $v \in [\underline{v}, \overline{v}]$.

Proof. Consider a candidate cutoff v^*. The marginal type v^* is indifferent between participation and abstention if and only if $u_0 = \Phi(v^*)$. Under the assumption that $\Phi'(v) > 0$ on $(\underline{v}, \overline{v})$, the equation $\Phi(v) = u_0$ has at most one solution. If $u_0 \in (\Phi(\underline{v}), \Phi(\overline{v}))$, existence of a solution $v^* \in (\underline{v}, \overline{v})$ follows from continuity and the intermediate value theorem, and uniqueness follows from strict monotonicity, establishing (i). If $u_0 \geq \Phi(\overline{v})$, then $\Phi(v) \leq u_0$ for all v, so no type strictly prefers participation; thus $a = 0$ for all v, establishing (ii). The argument for (iii) is symmetric. □

2.2 Key Predictions

In an interior equilibrium, the cutoff v^* satisfies

$$u_0 = v^* - c + \pi(v^*)T(v^*)$$

where $\pi(v^*) = \frac{(1-\rho)(1-F(v^*))}{\rho\alpha + (1-\rho)(1-F(v^*))}$ is the posterior probability that a participant is human, and $T(v^*) = \left[\mu E(v \mid v \geq v^*) - \gamma\right]$ is the net image-privacy surplus conditional on being perceived

as human. The product $\pi(v^*)T(v^*)$ highlights that the indirect payoff from participation depends on two components: how likely participation is to be attributed to a human, and how valuable it is to be perceived as human. Within $T(v^*)$, we refer to $\mu E(v \mid v \geq v^*)$ as the "honor" conferred by participation, capturing the image benefit from being perceived as intrinsically motivated to contribute, and to γ as the "exposure" cost, reflecting the disutility from potential identity exposure induced by participation.

We now describe how the equilibrium cutoff v^* responds to changes in the environment. The effect of Gen-AI bots operates through *dilution*: an increase in either ρ (a higher share of generative-AI bots) or α (a higher probability that bots mimic participation) makes participation a noisier signal of human identity, thereby lowering $\pi(v^*)$. Whether the dilution discourages or encourages human participation depends on whether honor or exposure is the dominant reputational concern. When honor dominates exposure ($T(v^*) > 0$), participation is on net rewarding conditional on being human, so dilution weakens the indirect benefit from participation, making participation less attractive and raising the cutoff v^*. When exposure dominates honor ($T(v^*) < 0$), participation is on net costly conditional on being human, so dilution instead attenuates this indirect loss, making participation more attractive and lowering the cutoff v^*.

By contrast, the parameters μ and γ affect the equilibrium more directly by shifting the value of being perceived as human. An increase in μ raises the honor associated with participation, increasing $T(v^*)$ and thereby lowering the cutoff v^*. An increase in γ raises exposure, reducing $T(v^*)$ and thereby increasing the cutoff v^*.

The following proposition summarizes these comparative statics.

Proposition 2 (Comparative Statics on v^*)**.** *Let $v^* \in (\underline{v}, \overline{v})$ be an interior equilibrium.*

(i) An increase in ρ or α dilutes the likelihood that a participant is human ($\pi(v^*)$ decreases). If the net image-privacy surplus $T(v^*) > 0$, this lowers the indirect payoff from participation and therefore increases the cutoff v^*. If $T(v^*) < 0$, the same dilution attenuates the (negative) indirect payoff and therefore decreases the cutoff v^*.

(ii) An increase in prosocial concern μ raises $T(v^*)$, increasing the indirect payoff from participation and thereby decreasing the cutoff v^*.

(iii) An increase in disclosure concern γ lowers $T(v^*)$, decreasing the indirect payoff from participation and thereby increasing the cutoff v^*.

Proof. For an interior equilibrium, v^* solves the indifference condition

$$u_0 = \Phi(v^*) = v^* - c + \pi(v^*)T(v^*).$$

Since $\Phi'(v^*) > 0$, the implicit function theorem applies and yields, for any parameter θ,

$$\frac{\partial v^*}{\partial \theta} = -\frac{\partial_\theta \Phi(v^*)}{\Phi'(v^*)},$$

so the sign of $\partial v^*/\partial\theta$ is the opposite of the sign of $\partial_\theta \Phi(v^*)$.

(i) $\theta = \rho, \alpha$. The parameters ρ and α enter Φ only through $\pi(\cdot)$. Hence

$$\partial_\rho \Phi(v^*) = T(v^*)\,\partial_\rho \pi(v^*), \qquad \partial_\alpha \Phi(v^*) = T(v^*)\,\partial_\alpha \pi(v^*).$$

From the expression $\pi(v) = \frac{(1-\rho)(1-F(v))}{\rho\alpha + (1-\rho)(1-F(v))}$, we have $\partial_\rho \pi(v^*) < 0$ and $\partial_\alpha \pi(v^*) < 0$ (a higher Gen-AI bot share or a higher mimicry rate lowers the posterior probability that a participant is human). It follows that $\partial_\rho \Phi(v^*)$ and $\partial_\alpha \Phi(v^*)$ have the opposite sign of $T(v^*)$, implying that $\partial v^*/\partial\rho$ and $\partial v^*/\partial\alpha$ have the same sign as $T(v^*)$. This establishes (i).

(ii) $\theta = \mu$. The parameter μ enters Φ only through $T(v^*) = \mu\, E[v \mid v \geq v^*] - \gamma$. Thus

$$\partial_\mu \Phi(v^*) = \pi(v^*)\,\partial_\mu T(v^*) = \pi(v^*)\,E[v \mid v \geq v^*] > 0,$$

so $\partial v^*/\partial\mu < 0$.

(iii) $\theta = \gamma$. Similarly,

$$\partial_\gamma \Phi(v^*) = \pi(v^*)\,\partial_\gamma T(v^*) = -\pi(v^*) < 0,$$

so $\partial v^*/\partial\gamma > 0$. □

3 Experimental Evidence

3.1 Design Overview

Recruitment. We recruited via the University Registration Center for Study (UAST) at ETH Zurich and the University of Zurich . The inclusion criteria are: aged 18+ and a current student at University of Zurich or ETH Zurich. The survey was administered online in batches, with participants required to complete it within a fixed time window.

The resulting sample is 1,075 current students. The sample is close to gender-balanced, with a median age of 23, and around half of participants report Switzerland as their country of origin.

Survey Overview. 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. Donation behavior is elicited through a lottery, following prior work (e.g., Song 2024 and Alesina, Miano and Stantcheva 2023; see Stantcheva 2023 for a review). This approach allows us to measure the extent to which participants are willing to donate, if at all, to organization working on issues of climate change. For a full list of the survey instruments, see Appendix A.1.

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 defined rule-based bots as having "a specific, limited functionality, and it is easy to distinguish them from humans." In contrast, we described generative AI bots as "difficult to distinguish from humans."

Furthermore, in case the participant was unfamiliar with the capabilities of GPT-powered AI bots, we provided two excerpts discussing the importance of donating to a charitable cause – one written by a human and the other by GPT-powered AI bot – which are hard to differentiate in terms of whether they were written by a human or AI. We also provide an example excerpt from a rule-based bot. To further familiarize participants with these concepts and to elicit baseline attitudes and behavior, we collect measures of participants' baseline usage of and familiarity with bots, as well as their preferences for interacting with humans, rule-based bots, or Gen-AI bots. These measures are elicited separately in the contexts of customer service and social media.

Participants then complete a reading task in which they read two posts, presented one at a time, and report their reactions as well as whether they believe 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-written posts or a mix that is mostly AI-generated) varies across the two questions. Both posts are written by humans,[11], and

[11]We use a human-written post to mimic the design in the writing task, and to avoid deception, as a post

the order in which the post texts are shown is randomized. This design allows us to assess whether participants can distinguish between human- and AI-generated posts and how labeling the message pool influences participants' perceptions of the posts.

All participants complete an attention check. Those who indicate that they use Instagram are asked whether they would like to provide their Instagram handle, which is incentivized through entry into a CHF 100 lottery.

After completing these baseline questions, participants are presented with a writing task and then complete a set of questions related to this task for the remainder of the survey.

Writing Task. In the writing task, 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 prompt reads as follows:

> *We have a writing task for you. For this task, we invite you to write a post on the following topic:*
>
> **How can you and other students in Zurich contribute to combating climate change in their daily lives?**

Participants are informed that their response will be added to a database that will be used in a follow-up survey with ETH/UZH students. In the follow-up survey, each post has a 1-in-20 chance of being randomly selected, and recipients will carefully read the selected posts and report their opinions about both the author and the post.[12]

Randomization. In a between-subjects design, participants who arrive at the writing task are cross-randomized along two dimensions: the composition of the co-senders and the identity disclosure condition.

The primary experimental manipulation is the composition of the message pool: participants are randomized with equal probability into either a condition where their message will be pooled exclusively with human-written messages or a condition where it will be pooled with

can be human-written in both conditions.

[12]To isolate the effects of social image and disclosure concerns, participants do not receive any response from the recipient. This design eliminates interaction effects and focuses attention on the sender's incentives and beliefs about how others will perceive them.

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 and express their opinions.

In the No-AI condition, participants see the following:

> *Your response will be added to a database of student-written posts. We will use this database in a follow-up survey with ETH/UZH students, where each post has a 1 in 20 chance of being randomly selected. In this follow-up survey, recipients will carefully read your post and share their opinion about both you and your post. They will be informed that the post is drawn from a database of student-written posts.*

In the AI condition, participants see the same message, with the only difference being that the two instances of "student-written posts" are replaced by "75% AI-generated posts and 25% student-written posts."

Participants are independently randomized into one of two identity conditions. In the identified condition, participants are informed that, upon completing the writing task, their first name (and Instagram handle, if provided), university, major, and year of study will be displayed alongside their response in the follow-up survey. In the anonymous condition, participants are informed that their response will be shared with information about their university, major, and year of study, but without any identifying information. To ensure that AI-generated posts cannot be easily distinguished from human-generated posts, participants are informed that, if an AI-generated post is selected, it will be presented using the credentials of a member of the research team, without disclosing that it was generated by AI.

The main difference between the identified and anonymous conditions is the disclosure of personal identifiers. For participants who did not provide an Instagram handle, only their first name is disclosed in the identified condition; for those who did, both their first name and handle are disclosed. These additional details are intended to heighten participants' concerns about others' beliefs, consistent with the methodology in Henry and Sonntag (2017). To maximize statistical power for testing predictions in the identified condition, participants are randomized such that 80% are assigned to the identified condition and 20% to the anonymous condition.

Quality control. Beyond including an attention check, we implement several measures to encourage effort in the writing task. Participants are instructed to complete the task in one sitting:

Participants were instructed to complete the writing task in one sitting:

> *Please complete this task in one session and avoid switching to other browser tabs, windows, or engaging in other activities until you finish.*

If participants switch to another browser tab during the writing task, they receive a pop-up reminder instructing them not to do so. In addition, we disable copy-and-paste functionality in the writing task to reduce the likelihood that participants rely on external tools, such as ChatGPT, to complete the survey. We also monitor tab switching during the survey, which allows us to detect whether participants navigate to other browser tabs while completing the task.

3.2 Outcomes and Estimation

Our primary outcome is whether a participant completes the survey. Given the very low attrition elsewhere in the survey, this behavioral outcome captures a binary measure of participation in content production and opinion sharing. Because participants complete an extensive baseline questionnaire before reaching the writing task, and attrition is typically low in laboratory-style samples such as ours, non-completion reflects a meaningful decision to forgo a CHF 10 participation payment as well as potential lottery payments.

In addition to survey completion, we present results on two alternative measures of participation – the length of the written response and whether the response constitutes meaningful engagement with the prompt, as classified using a GPT-based label. These secondary measures capture variation in effort and intensity in content production and opinion sharing.

To estimate treatment effects, define Y_i as an outcome for participant i (e.g., survey completion or length of message). 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 $\boldsymbol{X}_i$ 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:

$$Y_i = \alpha + \beta T_i + \delta D_i + \theta \left(T_i \times D_i\right) + \boldsymbol{X}_i' \boldsymbol{\gamma} + \varepsilon_i \tag{2}$$

We will potentially control for age, gender, education, university, and country of origin in Appendix Figures if any of the covariates are unbalanced.

To study heterogeneous treatment effects, we estimate treatment effects separately across prespecified moderators, including whether the participant shared their Instagram handle (as a proxy for disclosure concerns) and demographic characteristics such as age, gender, and country of origin.

3.3 Pre-registration and Deviations from the Pre-analysis Plan

The pre-analysis plan specifies the experimental design, outcome variables, estimating equations, and moderators.

The primary pre-specified outcomes are the effort participants put in writing the post and their willingness to publicly disclose the content. In particular, after submitting their message participants are given the opportunity to pay to have their response removed from the database. We use a Becker–DeGroot–Marschak (BDM) mechanism (Becker, DeGroot and Marschak, 1964) to elicit an incentive-compatible valuation of how much participants are willing to pay to keep their post from being shared. This design choice reflects our ex ante intent to measure disclosure behavior directly, rather than relying solely on stated preferences or hypothetical responses.

Our main deviation from the pre-analysis plan is that we use survey completion as the primary outcome. The PAP specified outcomes measured after the writing task, including effort and willingness to pay to remove the post. In the experiment, however, we observed substantial attrition precisely at the writing task, despite very low attrition earlier in the survey. A nontrivial share of participants chose to exit at this stage, thereby forgoing the CHF 10 participation payment, rather than proceed with the writing task. As a result, post-level outcomes are observed only conditional on completion. We therefore focus on survey completion as the primary outcome, interpreting it as a revealed-preference measure of participation in opinion sharing and content production. We additionally analyze writing-task outcomes, coding them as zero for participants who exited prior to completing the task, and show that the qualitative patterns are consistent across measures.

3.4 Results

Reading Task and Perceptions of Authorship. As part of the survey before the writing task, participants completed a reading task designed to assess their ability to distinguish between human- and AI-generated content. Participants were shown a single potential social media post on a climate change, and was informed that the post was randomly drawn from a pool of four posts, consisting of one post written by a climate-conscious human and three posts generated by GPT-powered Gen-AI bots. After reading the post, participants were asked to guess whether the text had been written by a human, a rule-based bot, or a Gen-AI bot. The distribution of responses is shown in Figure 1.

The results indicate that human and AI authorship is difficult for readers to distinguish in this setting. A substantial fraction of participants classified human-written posts as AI-generated. In other words, even when a post was produced by a human author, readers frequently inferred that it was written by a generative AI bot.

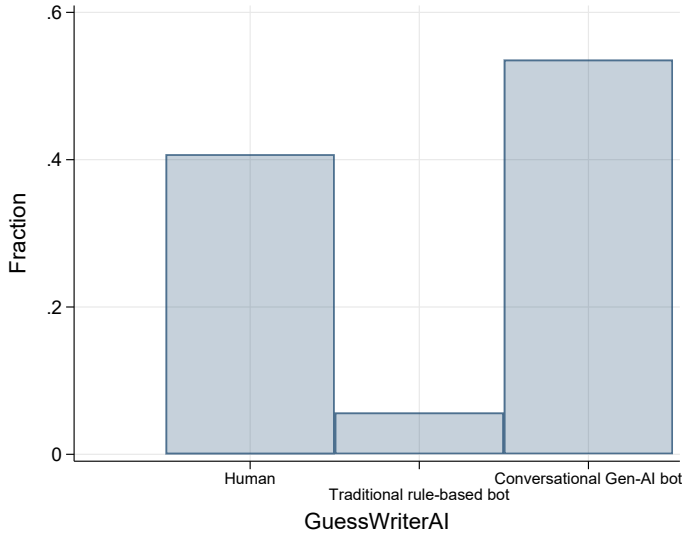
This finding is important for interpreting subsequent behavior in the experiment. Human senders who participate in the reading task are made aware that their writing may plausibly be mistaken for AI-generated content by others. As a result, the presence of AI in the environment does not merely add machine-generated messages. It also changes the informational and reputational context in which humans communicate. In particular, the possibility that human content may be perceived as AI-generated weakens the link between authored content and personal identity, which is a key channel through which disclosure concerns operate in the theoretical framework.

Writing Task and Main Outcomes. After completing a lengthy baseline survey, participants were given a writing task. They were informed that their response would be added to a database consisting of approximately 75 percent AI-generated posts and 25 percent student-written posts. Depending on treatment assignment, participants were further informed whether their identity might be disclosed to readers and whether AI agents were present as co-senders in the environment.

We study two main outcomes, shown in Figure 2. Panel A shows completion of the study, while Panel B shows whether participants wrote a meaningful post. Both outcomes display a consistent interaction between identity disclosure and the presence of AI co-senders.

When participants are not identified, completion rates are high and similar across AI conditions. In Panel A, completion is approximately 97 percent in the no-AI condition and 93 percent in the AI condition, and the difference is not statistically significant ($p =$

Figure 1: Reading Task - Guessing Writer in Presence of AI



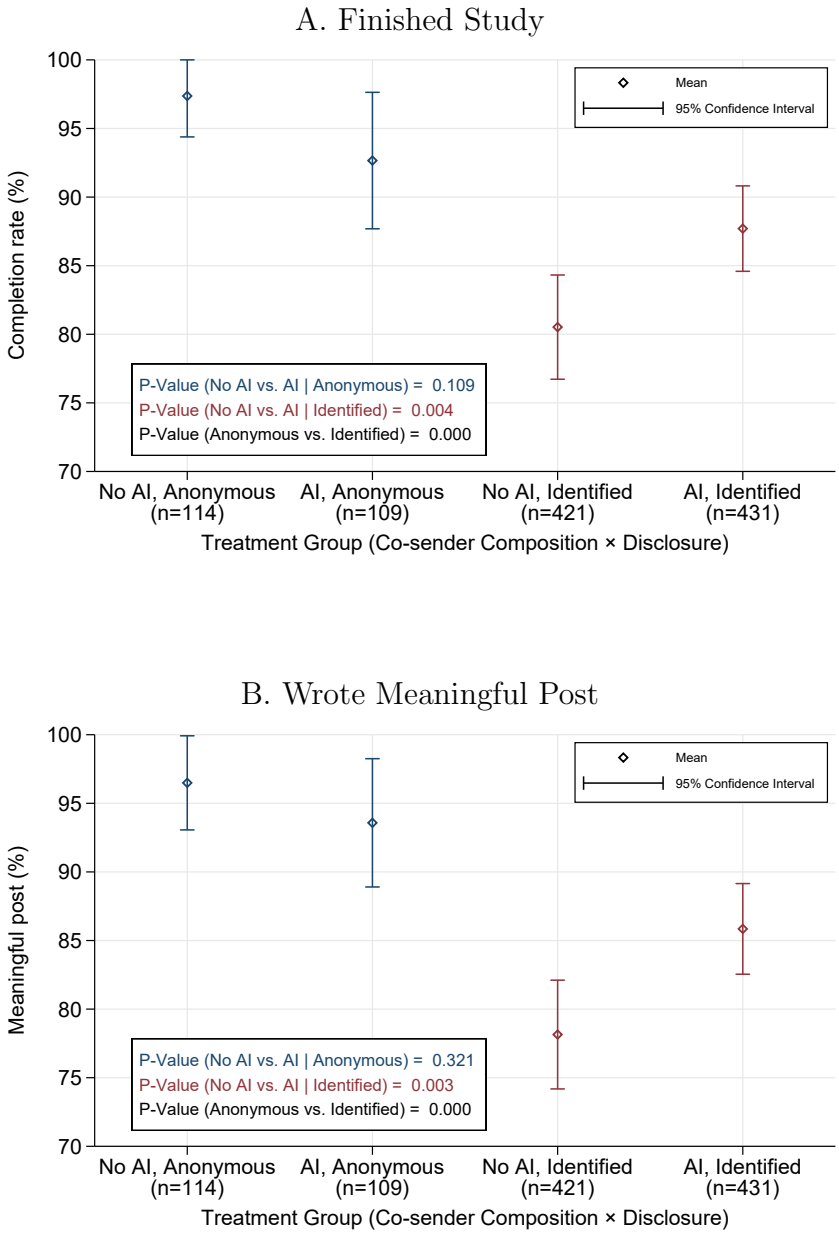
Notes: This figure shows the distribution of participants' guesses about the authorship of a human-written post shown in a reading task prior to the writing task.

0.109). A similar pattern holds for writing a meaningful post in Panel B, where participation remains high and the difference between AI and no-AI conditions is small and statistically insignificant ($p = 0.321$). This is consistent with the model's prediction that, in the absence of disclosure concerns, AI does not increase participation and may slightly reduce it by weakening signaling incentives.

By contrast, when participants are identified, outcomes change sharply. In the no-AI condition, identification substantially reduces participation. Completion falls to roughly 80 percent, and the share writing a meaningful post drops to about 78 percent. These declines are large relative to the anonymous conditions and are highly statistically significant (p ¡ 0.001). This pattern aligns with the model's prediction that identity disclosure increases disclosure concerns and discourages participation.

Next, we see that the presence of AI mitigates these identification effects. When AI co-senders are present, completion rises from about 80 percent to nearly 88 percent ($p = 0.004$), and the probability of writing a meaningful post increases from about 78 percent to 86 percent ($p = 0.003$). These gains are sizable and statistically significant, and they reverse a substantial portion of the participation loss induced by identification. This is exactly the pattern predicted by the model, in which AI reduces disclosure concerns by weakening the perceived link between authored content and personal identity. Appendix Figure A1 shows

Figure 2: Main Outcomes by Treatment Group



Notes: This figure shows our main outcomes, finishing the study (Panel A) and writing a meaningful post (Panel B), by the treatment groups. Estimates are based on the full sample of respondents who passed the attention check. Error bars indicate 95% confidence intervals, which are truncated at 100%.

similar results for the number of characters written.

It is important to emphasize that completion of the study is a high-stakes outcome. Participants reached the writing task only after completing a lengthy baseline survey, so attrition at this stage reflects a costly decision to disengage (and leave a significant payment on the table) rather than inattention or low effort. Attrition is typically low in surveys of this type, a fact reflected in the very high completion rates in the anonymous conditions. The sharp drop in completion under identification, and its partial reversal with AI presence, therefore indicates meaningful changes in participation incentives rather than noise or fatigue.

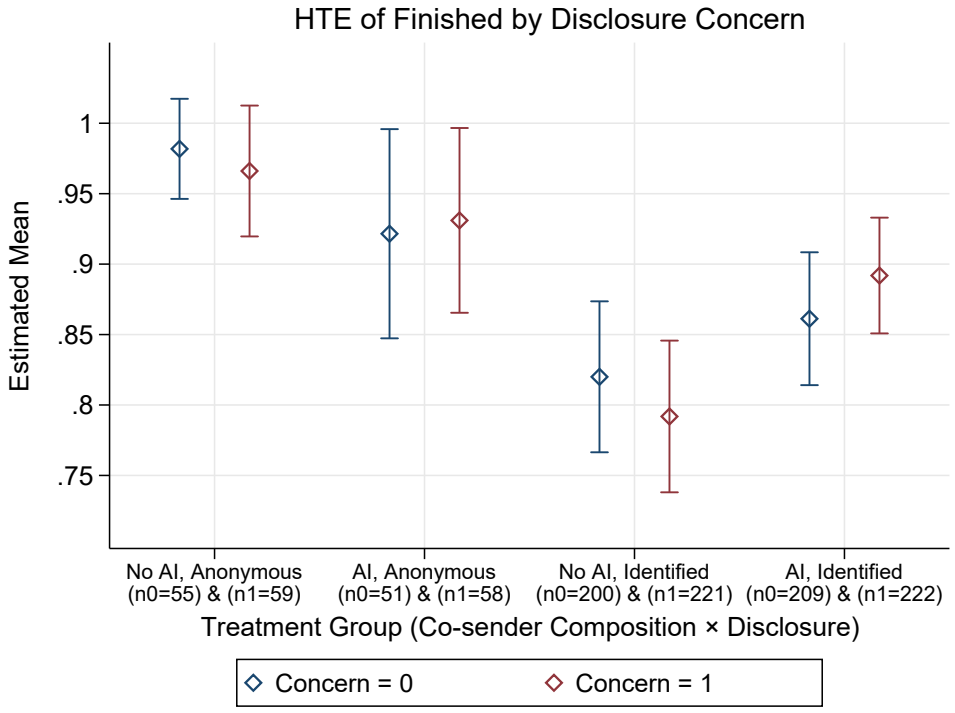
Taken together, Figure 2 provides strong support for the core mechanism of the model. Identification suppresses participation, AI increases participation when identification is present, and AI has little or no positive effect when identification is absent. These results show that AI reshapes human content production not only through machine-generated text, but by altering the disclosure and signaling environment faced by human contributors.

Heterogeneity by Baseline Disclosure Concerns. Figure 3 examines treatment effect heterogeneity by participants' baseline disclosure concerns, measured by whether respondents report reluctance to share their personal Instagram handle ('not shared handle'). The figure plots estimated completion rates separately for participants with lower and higher disclosure concern across the four treatment groups.

The results show clear and systematic heterogeneity. The effect of identification on participation is substantially larger for participants with higher disclosure concerns. Under identification and no AI, completion rates are markedly lower for those who report not sharing a handle compared to those with lower disclosure concern. This pattern is consistent with the model's emphasis on disclosure costs. Participants who are more sensitive to disclosure face higher perceived costs from having their content linked to their identity, and therefore are more likely to disengage when identification is salient. Notably, this holds even though participants who shared an Instagram handle have more information disclosed (their handle in addition to their first name) than those who did not.

Further, the mitigating effect of AI presence under identification is also stronger among participants with higher disclosure concern. When AI co-senders are present, completion rates increase more for individuals with higher disclosure concern than for those with lower concern. This suggests that AI is especially effective at reducing disclosure concerns for individuals who are most sensitive to them. In the language of the model, AI more strongly weakens the perceived mapping between authored content and personal identity for these

Figure 3: Heterogeneous Effects by Disclosure Concern



Notes: This figure reports treatment effects on study completion by disclosure concern. Disclosure concern is proxied by willingness to share Instagram handle where those who did not share are classified as high disclosure concern and those who shared as low disclosure concern. Estimates are based on respondents who passed the attention check. Error bars denote 95% confidence intervals.

participants, thereby restoring participation incentives.

By contrast, in the anonymous conditions, completion rates are high for both privacy groups, and differences between AI and no-AI conditions are small. This again matches the theoretical prediction that when disclosure concerns are already muted by anonymity, the presence of AI has little scope to increase participation. Because sharing an Instagram handle is only possible for Instagram users, our high–disclosure-concern group includes both non-users and users who choose not to share their handle. To isolate disclosure concerns from platform usage, Appendix Figure A2 shows this result holds when limiting to the sample of participants that have an Instagram account.

These heterogeneity patterns reinforce the proposed mechanism. The main effects are not driven by generic reactions to AI or survey fatigue, but by differential disclosure concerns across participants. The fact that the interaction between identification and AI presence is

largest among those with higher baseline disclosure concern provides direct evidence that AI operates through the disclosure channel highlighted in the theoretical framework. The same qualitative patterns appear both in the full sample and in the Instagram-user subsample, indicating that the results are not driven by platform usage or selective participation.

Heterogeneity by Demographics. In the appendix, we examine heterogeneity in participation outcomes across demographic groups. While baseline participation levels differ across demographics, the qualitative pattern of treatment effects remains largely stable, with one important exception related to gender.

Younger participants are more likely to drop out than older participants across all treatment conditions (Appendix Figure A3). However, the interaction between identification and AI presence is qualitatively similar for both age groups. Identification reduces participation, and the presence of AI under identification increases participation relative to the no-AI condition. This suggests that age-related differences affect baseline engagement rather than the underlying disclosure mechanism.

Swiss participants exhibit higher dropout rates than non-Swiss participants (Appendix Figure A4), while dropout rates are similar across groups under the anonymous condition. This pattern indicates that identification has a larger effect on participation for Swiss than for non-Swiss participants. By contrast, under the identified condition, the direction and relative magnitude of the effects of AI presence are qualitatively similar across nationality groups, indicating that the main results on AI presence are not driven by country-specific norms or survey familiarity.

Gender shows a more distinct pattern (Appendix Figure A5). Women report higher disclosure concerns than men, and the treatment effects differ accordingly. Under identification, women are more likely to disengage in the absence of AI, consistent with higher perceived disclosure costs. At the same time, the presence of AI under identification leads to a larger increase in participation for women than for men. This differential response aligns closely with the model's predictions and with the heterogeneity by disclosure concern documented above. For participants with higher disclosure sensitivity, AI more strongly reduces perceived identity exposure and restores participation incentives.

Overall, these demographic heterogeneity results reinforce the interpretation that the observed effects operate through disclosure concerns rather than through general attitudes toward AI or survey participation. Differences across age and nationality affect baseline engagement, while differences across gender interact meaningfully with AI presence in a way

that mirrors differences in disclosure sensitivity.

4 Conclusion

Many online platforms are built around social interaction and identity, yet these environments increasingly include artificial agents that are difficult to distinguish from humans. This paper develops a signaling framework in which participation in content production yields social image benefits but may also impose privacy and disclosure costs when authorship can be linked to identity. We show that the presence of generative AI alters human participation by weakening the perceived connection between content and personal identity. In an online writing experiment, making participants identifiable sharply reduces participation, even late in a paid survey. However, introducing AI co-senders substantially mitigates this effect. When participants remain anonymous, AI has little or no positive effect on participation. These patterns are strongest among participants with higher disclosure concerns and among women. Together, the results show that AI reshapes human content production not only by generating text, but by changing the disclosure incentives faced by human contributors.

Our findings clarify why the effects of AI on human engagement can appear contradictory across settings. When participation primarily serves as a signal of prosocial motivation, the presence of AI dilutes signaling value and can discourage human contributions. By contrast, when participation carries meaningful disclosure risk, AI can increase engagement by attenuating exposure. This distinction helps reconcile concerns that AI may crowd out human expression with evidence that AI can sometimes expand participation. Whether AI reduces or increases human engagement depends on how strongly content is tied to personal identity and how costly that linkage is perceived to be.

The results also speak directly to platform design. Decisions about attribution, identity disclosure, and the visibility of AI-generated content are not neutral. By shaping the extent to which authored content is perceived as uniquely human, platforms influence who chooses to speak and who remains silent. In environments where disclosure concerns are salient, introducing AI or reducing certainty about authorship may broaden participation, particularly among users who are more privacy sensitive. Conversely, in settings where reputation and signaling are central, the same design choices may weaken incentives to contribute.

Several avenues for future research follow naturally. First, repeated interaction and feedback may amplify or dampen the disclosure mechanism we identify, especially as users

update beliefs about AI prevalence and attribution over time. Second, richer attribution regimes, such as probabilistic labeling or optional identity disclosure, could allow users to sort into environments that better match their preferences. Finally, understanding how AI affects participation in higher-stakes domains such as political speech, workplace communication, or mental health forums remains an important open question. More broadly, our results suggest that as AI becomes embedded in social environments, its most consequential effects may arise not from what machines say, but from how their presence reshapes the social meaning of human speech.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2023. "A Model of Online Misinformation." *The Review of Economic Studies*, rdad111.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2023. "Immigration and redistribution." *The Review of Economic Studies*, 90(1): 1–39.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2024. "The economics of social media." *Journal of Economic Literature*, 62(4): 1422–1474.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral science*, 9(3): 226–232.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652–1678.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841–877.

Bogliacino, Francesco, Paolo Buonanno, Francesco Fallucchi, and Marcello Puca. 2023. "Trust in times of AI." Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy CSEF Working Papers 689.

Cho, In-Koo, and David M. Kreps. 1987. "Signal Games and Stable Equilibria." *The Quarterly Journal of Economics*, 102(2): 179–222.

Chugunova, Marina, and Daniela Sele. 2022. "We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines." *Journal of Behavioral and Experimental Economics*, 99: 101897.

Cookson, J. Anthony, William Mullins, and Marina Niessner. 2024. "Social Media and Finance."

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social media and protest participation: Evidence from Russia." *Econometrica*, 88(4): 1479–1514.

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. 2023. "Curtailing False News, Amplifying Truth."

Henry, Emeric, and Jan Sonntag. 2017. "Measuring Image Concern." *AARN: The Evolution of Social Behavior (Topic).*

Loewen, Peter John, Blake Lee-Whiting, Maggie Arai, Thomas Bergeron, Thomas Galipeau, Isaac Gazendam, Hugh Needham, Lee Slinger, and Sofiya Yusypovych. 2024. "The Global Public Opinion on Artificial Intelligence survey, Schwartz Reisman Institute for Technology and Society."

Maréchal, Michel André, Alain Cohn, and Tobias Gesche. 2020. "Honesty in the Digital Age (December 1, 2020). University of Zurich, Department of Economics, Working Paper No. 280."

Song, Lena. 2024. "Closing the distance: The effects of social media content on support for racial justice."

Srinivasan, Karthik. 2023. "Paying Attention."

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics*, 15(Volume 15, 2023): 205–234.

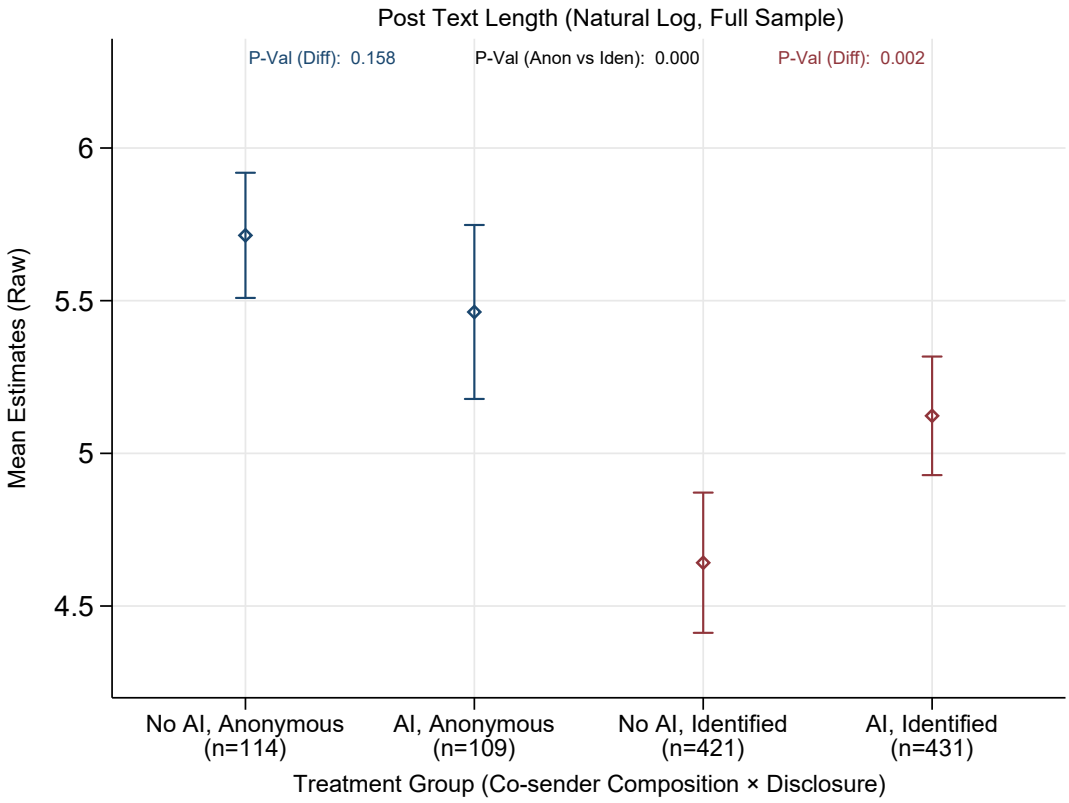
Varol, Onur, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. "Online Human-Bot Interactions: Detection, Estimation, and Characterization." *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 280–289.

Online Appendix: Not for Publication

Sharing Opinions in the Shadow of AI: Theory and Evidence from Online Interactions

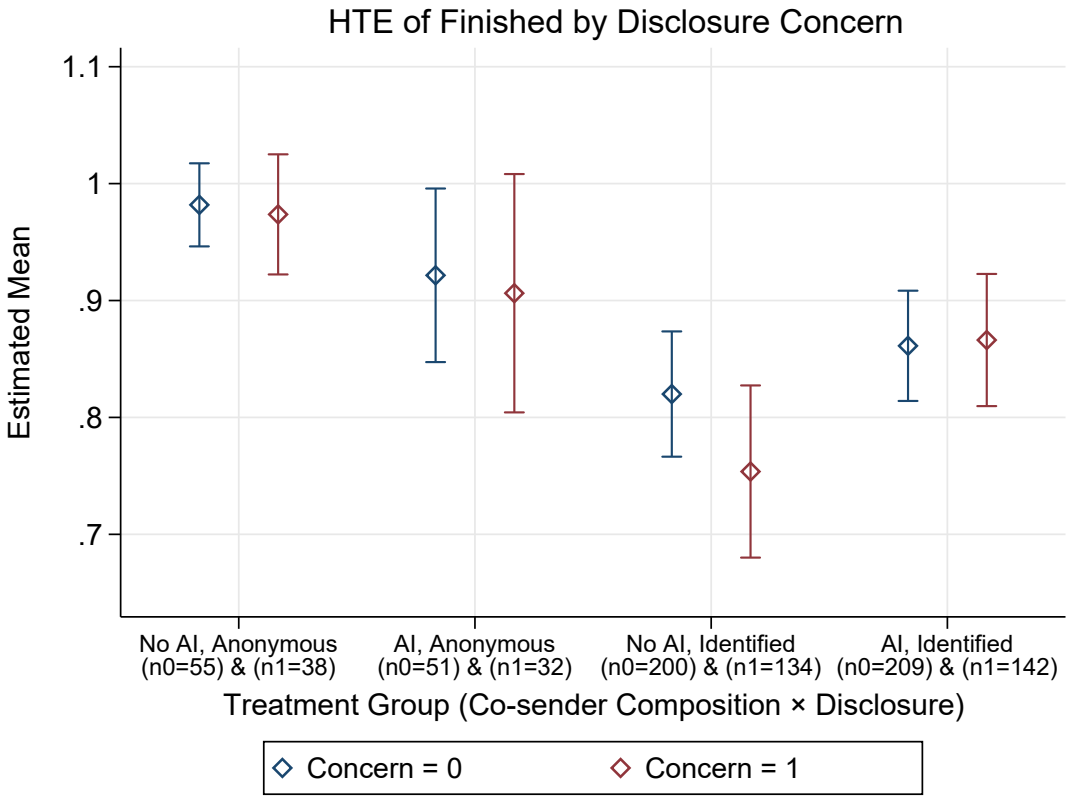
Elliott Ash, Yichuan Lou, Lena Song

Figure A1: Post Length by Treatment Group



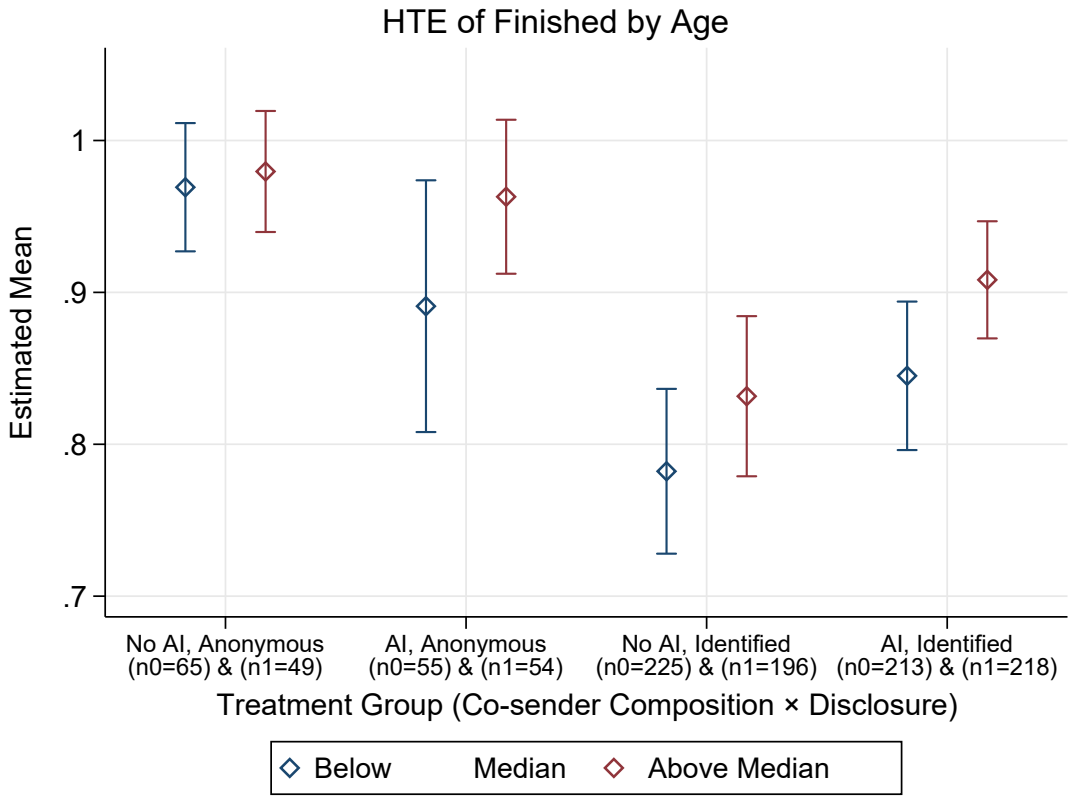
Notes: This figure shows post length by the treatment groups. Estimates are based on the full sample of respondents who passed the attention check. Error bars indicate 95% confidence intervals.

Figure A2: Heterogeneity by disclosure concern - Instagram User Sample



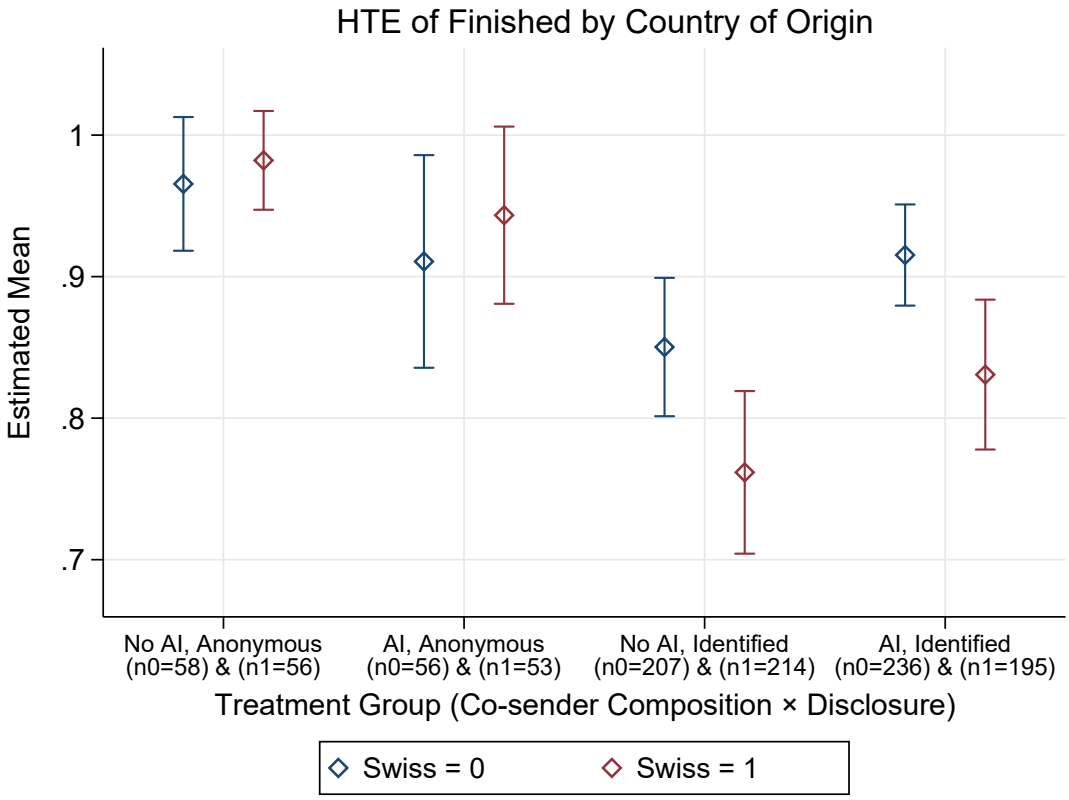
Notes: This figure reports treatment effects on study completion by disclosure concern, for Instagram users only. Disclosure concern is proxied by willingness to share Instagram handle where those who did not share are classified as high disclosure concern and those who shared as low disclosure concern. Estimates are based on respondents who passed the attention check. Error bars denote 95% confidence intervals.

Figure A3: Heterogeneity by Age



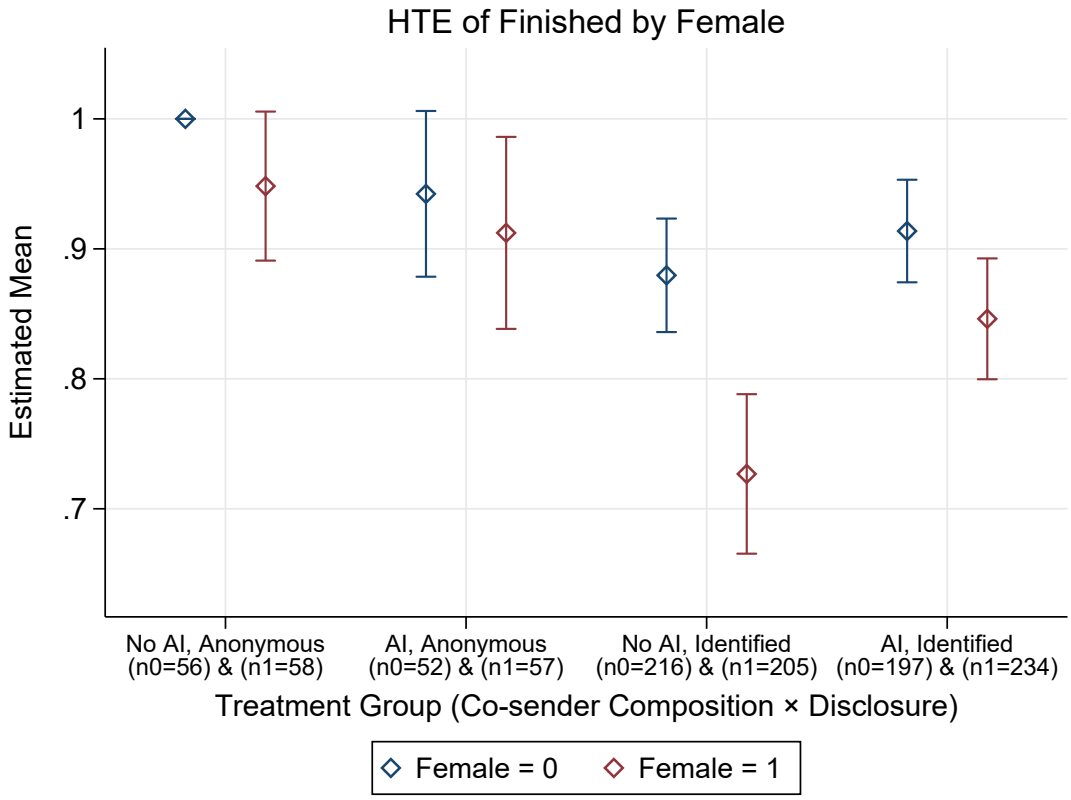
Notes: This figure reports treatment effects on study completion by age group. Age is defined relative to the sample median, with respondents classified as below-median or above-median age. Estimates are based on respondents who passed the attention check. Error bars denote 95% confidence intervals.

Figure A4: Heterogeneity by Nationality



Notes: This figure reports treatment effects on study completion by nationality. Respondents are classified as Swiss or non-Swiss. Estimates are based on respondents who passed the attention check. Error bars denote 95% confidence intervals.

Figure A5: Heterogeneity by Gender



Notes: This figure reports treatment effects on study completion by gender. Respondents are classified as female or non-female. Estimates are based on respondents who passed the attention check. Error bars denote 95% confidence intervals.

A Experimental Design Appendix

A.1 Survey Instruments

Demographic Questions

1. Age: What is your age? [number entry box]
2. Gender: What gender do you most strongly identify with? [Male / Female / Other or Non-binary]
3. University: Which university are you at? [ETH / UZH / Other]
4. GradYear: What is your anticipated year of graduation? [2024 or earlier / 2025 / 2026 / 2027 / 2028 or after]
5. FirstName: What is your first name? [text entry box]
6. OriginCountry: What is your country of origin? [dropdown menu]
7. Education: What level of education are you currently pursuing? [Bachelor's degree / Master's degree / PhD / Other]

Field of Study

1. ETHFaculty: [If ETH] What faculty do you study in? [Architecture and Civil Engineering / Engineering Sciences / Natural Sciences and Mathematics / System-oriented Natural Sciences / Management and Social Sciences]
2. UZHFaculty: [If UZH] What faculty do you study in? [Faculty of Theology / Faculty of Law / Faculty of Business, Economics, and Informatics / Faculty of Medicine / Vetsuisse Faculty / Faculty of Arts and Social Sciences / Faculty of Science]

Political and Climate Opinions

1. SwissPoliticalParty: Which party did you vote for in the last election? [Did not vote / Swiss People's Party / Social Democratic Party / FDP.The Liberals / Green Party / The Centre / Green Liberal Party / Evangelical People's Party / Federal Democratic Union / Ticino League / Swiss Labour Party / Small leftist parties / Center Left-CSP / Liberal-Democratic Party / Movement of the Citizens of French-speaking Switzerland / Pirate Party Switzerland / Mixed vote / Other]

2. ClimateWorry: How worried are you about climate change? [Not at all worried / Not very worried / Moderately worried / Very worried / Extremely worried]

3. ClimatePersonal: To what extent do you feel a personal responsibility to try to reduce climate change? [Not at all responsible / Slightly responsible / Moderately responsible / Very responsible / Extremely responsible]

4. ClimateLong: What do you think are the most serious consequences of climate change for Europe and the world? [open text]

Social Media Use

1. SocialMediaPlatform: Which of the following social media platforms do you use? [Facebook / Instagram / Snapchat / TikTok / X (Twitter) / BlueSky / LinkedIn / Reddit / YouTube / Other / I don't use social media] [multiple answer]

2. UseImageConcern: [If uses social media] How important is it to you that other people perceive you positively on social media? [Not important at all / Slightly important / Moderately important / Very important / Extremely important]

Prosocial Behavior

1. Donation: As part of completing this survey, you are automatically enrolled in a lottery to win CHF X. Would you be willing to donate part of your winnings to an organization working on issues of climate change? Enter the amount you would like to donate, if any. [number entry box]

2. DonateOrg: [If Donation > 0] Which organization would you like your donation to support? [text entry box]

Bot and AI Opinions

Participants were first shown an explanation distinguishing *rule-based bots* from *conversational generative AI (Gen-AI) bots*:

BotExplain: In this study, we are interested in your views about two types of "bots" – that is, automated programs designed to perform tasks or engage in conversations with humans.

The first type, known as 'rule-based bots,' operate on a set of predefined rules and responses. They are programmed to recognize specific commands and provide corresponding

answers. These bots are often found in customer service chat functions, where they can efficiently handle routine inquiries without human intervention. Rule-based bots have a specific, limited functionality, and it is easy to distinguish them from humans.

The second type represents a more advanced form of technology: 'Conversational Generative AI (Gen-AI) bots.' These bots leverage artificial intelligence to perform complex problem-solving tasks and to generate responses that mimic human conversation. Unlike their rule-based counterparts, Gen-AI bots can understand and respond to a wider range of queries with greater nuance, variability, and creativity. They learn from interactions to improve their responses, making them more adaptable and capable of handling deep conversations. As a result, Gen-AI bots are difficult to distinguish from humans.

1. BotKnow: How familiar are you with ChatGPT and other Gen-AI bots? [Very familiar and use frequently / Somewhat familiar and use occasionally / Somewhat familiar but rarely use / Aware but do not use / Not familiar at all]

2. BotSupport: When seeking customer service or support, what are your preferences for interacting with humans, traditional rule-based bots, or conversational Gen-AI bots? [5-point Likert scale from Strongly Prefer Not to Strongly Prefer]

3. BotSocialMedia: On social media, what are your preferences for interacting with humans, traditional rule-based bots, or conversational Gen-AI bots? [5-point Likert scale from Strongly Prefer Not to Strongly Prefer]

Participants are then presented with examples:
Here is a potential social media post:

> Some of us start closer to the finish line than others. Some individuals need a boost to get their lives back on track. It feels good to give when you are fortunate to have enough. Let's all do our part and donate to things that matter to us.

1. PostReact: How would you react if you saw this post on your social media timeline? [Like / Dislike / I would not react to this post]

On the next page, participants were informed that the post they just reacted to was written by a human.

For your information, here is a potential social media post, written by a GPT-powered Gen-AI bot:

> Many people face obstacles that are too big to tackle alone. Donating provides essential support to those who need a little extra help to overcome challenges. When we contribute to causes close to our hearts, we create a ripple of positive change in the world. Let's all pitch in and make a difference where we can.

Here is another potential social media post, written by a rule-based bot:

> I am a bot programmed to remind people that donating is important because it helps those in need.

On the next page, participants answered the following question:

1. InitialAIEffective: How effective do you think Gen-AI bots are at producing persuasive messages? [Not at all effective / Slightly effective / Moderately effective / Very effective / Extremely effective]

Reading Tasks

Participants were informed that they would read and respond to social media posts that may have been written by humans, or Gen-AI bots. Participants were told that all posts written by humans were created without any AI assistance.

Here is a potential social media post, drawn from a pool of four posts from climate-conscious humans:

> Addressing climate change is an important global issue, and it is great to see countries starting to collaborate on it. We are doing our part to help with it, too. However, while addressing climate change is important, we have other important local issues to focus on. Improving healthcare, reducing inflation and financing retirement provisions deserve our attention, too.
>
> If we are in a spot to contribute to charity, the impacts we make can last a lifetime and impact generations to come. As of now, addressing climate change is the most important thing any of us could do. If we are in a place to donate, putting our money towards charities that help take care of the planet will impact future generations of children and animals.

(One of the post was randomly drawn and shown to the participant.)

1. ReadingReactHuman: How would you react if you saw this post on your social media timeline? [Like / Dislike / I would not react to this post]

2. GuessWriterHuman: What is your best guess for who wrote this text? [Human / Traditional rule-based bot / Conversational Gen-AI bot]

On the next page, participants were shown the other post, while being informed that it is drawn from a pool of four posts from 1 climate-conscious humans, and 3 GPT-powered Gen-AI bots

1. ReadingReactAI: How would you react if you saw this post on your social media timeline? [Like / Dislike / I would not react to this post]

2. GuessWriterAI: What is your best guess for who wrote this text? [Human / Traditional rule-based bot / Conversational Gen-AI bot]

Attention Check

1. AttnCheckExplanation: Participants were instructed to ignore the next question and instead write "I read the instructions" to verify attention.

2. AttnCheck: Please describe how you think AI is impacting your daily life. [open text]

Writing Task

Participants were randomly assigned to an identified or anonymous condition and asked to write a social media post responding to the following prompt:

> How can you and other students in Zurich contribute to combating climate change in their daily lives?

Posts were added to a database used in a follow-up survey, with composition varying by treatment condition.

Willingness to Pay to Withdraw Post

1. WTP: What is the maximum amount (out of your CHF 10 participation payment) you would be willing to pay to remove your post from the database? [slider from 0 to 10 CHF in 0.25 increments]

2. WTPLong: Please explain the reasoning behind your decision. [open text]

Perceptions and Mechanisms

1. GenAIEffective: Compared to ETH/UZH students, do you think Gen-AI bots are more or less effective at writing posts that attract likes? [More effective / Equally effective / Less effective / Unsure]

2. PerceiveHuman: How likely do you think the audience will believe your post is AI-generated? [Very unlikely / Somewhat unlikely / Neutral / Somewhat likely / Very likely]

3. SignalValue: How important is it to you that the audience perceives you positively? [Not important at all / Slightly important / Moderately important / Very important / Extremely important]

Experimenter Demand Check

1. StudyPurposeLong: If you had to guess, what would you say our research project is about? [open text]

Final Comments

1. GeneralFeedbackLong: Please share any additional feedback or comments. [open text]