

CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 689

Trust in times of AI

Francesco Bogliacino, Paolo Buonanno, Francesco Fallucchi, and Marcello Puca

October 2023



University of Naples Federico II



University of Salerno



Bocconi University, Milan

WORKING PAPER NO. 689

Trust in times of AI

**Francesco Bogliacino^{*}, Paolo Buonanno[†], Francesco Fallucchi[‡],
and Marcello Puca[§]**

Abstract

In an online, pre-registered experiment, we explore the impact of AI mediated communication within the context of a Trust Game with unverifiable actions. We compare a baseline treatment, where no communication is allowed, to treatments where participants can use free-form communication or have the additional option of using ChatGPT-generated promises, which were assessed in a companion experiment. We confirm previous observations that communication bolsters trust and trustworthiness. In the AI treatment, trustworthiness sees the most significant increase, yet trust levels decline for those who opt not to write a message. AI-generated promises become more frequent but garner less trust. Consequently, the overall trust and efficiency levels in the AI treatment align with that of human communication. Contrary to our assumptions, less trustworthy individuals do not show a higher propensity to delegate messages to ChatGPT.

JEL Classification: C93, D83, D84, D91.

Keywords: Artificial Intelligence, Trust Game, ChatGPT, Experiment.

Acknowledgements: We thank Martin Dufwenberg, Elena Manzoni, Daniele Nosenzo, and participants at the 10th BEEN meeting at the University of Trento for helpful comments.

^{*} University of Bergamo.

[†] University of Bergamo.

[‡] University of Bergamo.

[§] University of Bergamo and CSEF.

1 Introduction

Since OpenAI’s ChatGPT launch in November 2022, interactions with generative AI systems, including chatbots and automated customer services, have skyrocketed. The rise of “Large Language Models” (LLM) like ChatGPT has amplified productivity in many domains, including research (Charness et al., 2023). Yet, there’s growing concern about over-reliance on such technologies (Ahmad et al., 2023), as the LLMs’ potential to replace human tasks raises questions about personal engagement (see Gilardi et al., 2023; Charness and Grieco, 2023) and the broader job market’s future (Acemoglu et al., 2022; Acemoglu and Lensman, 2023). Compounding these issues is the ambiguity individuals face in distinguishing AI-generated content from human-created content (Menczer et al., 2023; Sadasivan et al., 2023), challenging the foundation of contracts and trust (Gillespie et al., 2023).

In this paper, we delve into this emerging challenge by examining the effects of AI-mediated communication within the context of the Trust Game (Berg et al., 1995), as adapted by Charness and Dufwenberg (2006). The interaction involves an individual, A, who can engage with B for potential mutual benefits. However, A may end up harmed if B, whose actions are hidden, chooses to respond in a purely self-interested manner. Before making a decision, B has the opportunity to communicate with A. At the core of our exploration is the hypothesis that while a communicated promise typically fosters trust, outsourcing such a commitment to AI—a medium physically and emotionally disconnected from the sender—might undermine its perceived sincerity, thereby reducing its trust-enhancing potential. In this study, we explore the efficacy and effects of the widely-used AI tool, ChatGPT, in a strategic communication setting. We initiate our investigation by prompting ChatGPT (in its 4.0 version)

with the instructions of the trust game employed in Charness and Dufwenberg (2006). After detailing the game’s payoffs, we tasked ChatGPT with generating a series of trustee messages, mirroring the promises typically made by trustworthy humans. Our findings align with other research, suggesting that ChatGPT effectively applies backward induction to solve the trust game and crafts persuasive messages (e.g., Shen et al., 2023).¹

Our first contribution is to document that, in a static Turing test (Turing, 1950), participants struggle to accurately identify AI-generated from human-written messages, often misattributing human-like characteristics to AI outputs (e.g., smileys). Based on these results, we retain AI-generated messages most frequently mistaken for human-written content. Subsequently, we conduct an online adaptation of the modified trust game by Charness and Dufwenberg (2006), introducing a treatment where trustees can choose to use these pre-identified, AI-generated promises.

Our second contribution is to confirm the pivotal role of pre-play communication in fostering trust and trustworthiness, as originally observed by Charness and Dufwenberg (2006), within an online setting and with a sufficiently powered test (1,082 participants). The dynamics in the AI-mediated communication treatment are nuanced. While trust levels decline for those not composing a message compared to the control, a significantly larger proportion of trustees delegate the task of communicating promises to AI. This delegation to AI results in a surge in the rate of promises. AI-induced promises are more frequent, yet less believed, leading to an overall trust level that is statistically indistinguishable between Communication and

¹While we recognize that a LLM cannot engage in actual strategic reasoning, we found that ChatGPT was still able to generate messages consistent with the subgame perfect equilibrium of the game.

AI treatments. Furthermore, our data suggests that promises, whether human-crafted or AI-generated, significantly enhance trustworthiness. Interestingly, the access to AI almost eliminates messages suggesting to play OUT, which we document (to the best of our knowledge, for the first time) in Communication.

Our third contribution is to provide evidence supporting guilt aversion (Charness and Dufwenberg, 2006) and commitment (Ellingsen and Johannesson, 2004) as the primary mechanisms behind our experimental results. Both are present in our sample. Traditional promises bolster trustworthiness, but when these commitments are delegated to AI, a unique dynamic emerges, as a result of two forces. First, the perceived detachment of an AI-generated promise does not evoke the same depth of trust as human communication. Second, the ease of using AI options leads to a notably higher propensity to make promises. In the end, the trust in promises remains the same across both human and AI-mediated communication.

We also contribute to the literature on the role of communication in strategic interactions. Numerous studies observed a remarkable impact of communication in resolving social dilemmas (Sally, 1995; Goeree and Yariv, 2011; Puca et al., 2023) or enhancing efficiency in bilateral negotiations (Brandts et al., 2019). However, communication introduces multiple causal mechanisms, such as clarifying contract terms, delivering promises, or conveying emotions to reduce social distance Brandts et al. (2016). Experimentally manipulating communication’s structure and content helps isolate these mechanisms. In principal-agent interactions like ours, several studies have contrasted basic promises with more comprehensive communication methods, finding that detailed messages often outperform limited ones (Ben-Ner and Putter-

man, 2009; Charness and Dufwenberg, 2010; Lundquist et al., 2009).² Theoretical literature, however, suggests that costless and unverifiable messages (Crawford and Sobel, 1982) should be unreliable (Farrell and Rabin, 1996), unless driven by psychological motivations, such as aversion to lying (Gneezy, 2005).

Our setting introduces a novel twist. LLMs can craft human-like messages, potentially delegating commitment and offering justifications for selfish actions. Consequently, we anticipated that when given a choice, individuals would delegate deceitful messages and assume responsibility for genuine ones. Contrary to our expectations, our data revealed that trustees do not differentiate based on message type: we had predicted that less trustworthy subjects would delegate promises (to avoid guilt), but the data contradicted this. The only noticeable negative effect is skepticism towards those who abstain from communication, especially since AI diminishes the cost of messaging.

The remainder of the paper is structured as follows: Section 2 describes the experimental design and the ensuing research hypotheses. Section 3 presents our experimental findings. Section 4 discusses the potential mechanisms underpinning our results, and Section 5 concludes.

2 Experimental Design and Hypotheses

Our study consists of two experiments. The main experiment is based on a trust game, as introduced by Berg et al. (1995) and modified by Charness and Dufwenberg

²A variation on the standard coding protocol is provided by Turmunkh et al. (2019), who categorize messages in terms of malleability, distinguishing between implicit v. explicit and conditional v. unconditional. They found that cooperation is highest (67%) when the pre-play communication is explicit and unconditional.

(2006). This game involves a sequential interaction between a trustor and a trustee. More specifically, the trustor (A) has two action choices: IN or OUT. If A chooses OUT, both participants receive the outside option of 1.25 GBP.³ If A chooses IN, the trustee (B) must decide on rolling a dice. Without rolling, A and B receive 0 and 3.5 GBP, respectively. If B decides to roll, there’s a 5/6 chance that A receives 3 and B gets 2.5 GBP, and a 1/6 chance that A receives 0 and B gets 2.5 GBP.⁴ The trustor’s initial decision to transfer is termed “trust”, as it can potentially benefit both but also grants the trustee control over the resources (Fehr, 2009). The trustee’s subsequent decision reflects her “trustworthiness”.

Subjects are randomly assigned to three treatments: in Communication, player B can communicate with A, sending a free-form message (as in Charness and Dufwenberg, 2006); in AI, the messages can be written or selected among those pre-written; in control, participants only play the stage game. To generate AI pre-written messages, we conducted a parallel experiment, detailed below.

2.1 Turing test experiment

To design credible AI-generated messages for the main study, we conducted a preliminary online experiment, randomly sampling participants from the UK population on Prolific. A total of 99 participants engaged in a message recognition task, inspired by Turing (1950). Participants received a show-up fee of 0.5 GBP, along with an additional 0.15 GBP for each correct response. All data were collected on May 23, 2023. To provide the necessary context, participants were initially briefed on the version of

³As the experiments’ participants were sampled from a British pool of volunteers, payoffs are denominated in pounds.

⁴The dice roll introduces a *hidden action* as the trustor observes the outcome but remains uncertain about the trustee’s choice.

the trust game used in the main experiment. Subsequently, they were presented with a randomized sequence of three human-written and three AI-generated messages and tasked with identifying the origin of each message.

The messages for this experiment were drawn from a pool of ten messages written by humans and ten generated by AI. The human messages were sourced from the data in Charness and Dufwenberg (2006).⁵ The AI messages were selected among those generated using the following prompt, where “[...]” indicates that we expect a reply from ChatGPT:⁶

[*after explaining the game*] Write 5 messages from B to convince A to play IN. Use different lengths and level of promises. [...] Write other 5 messages with different language tones [...] Imagine that A replies with “I still don’t trust you” can you write 5 messages that would surely convince A? [...] Rewrite these latter messages assuming that A and B never met and will never meet again after playing the game [...] Based on the following 10 messages, produce 10 more messages.

Results of the Turing test Table 1 displays the ten promises generated by ChatGPT, along with the percentage of participants who incorrectly believed each message was human-written. Messages highlighted in bold were chosen for the main experiment. Human-written messages were correctly identified 58% of the time, with recognition rates varying from a low of 18% to a high of 83%. For AI-generated messages,

⁵We selected messages S1-ID2, S1-ID5, S1-ID8, S1-ID9, S1-ID12, S1-ID18, S2-ID12, S3-ID4, S3-ID10, S3-ID19 from Table S.I in the Appendix of Charness and Dufwenberg (2006). All messages were associated with a trusting decision in the original paper and were classified as promises by independent coders.

⁶We report in Appendix 6.1 all the queries and answers by ChatGPT 4.0.

instead, the recognition rates ranged between 23% and 85%. The corresponding table for human messages can be found in Appendix 6.2.1. A Mann-Whitney U-test did not find a statistically significant difference in the median correct scores assigned to the two sets of messages ($z = 0.832$, $p = 0.43$).

Result 1 (Turing test). *People fail to distinguish between AI-generated and human-written messages.*

Before delving into the main experiment’s discussion, we conclude this subsection with two observations. First, the messages in the table reference the original paper, and the payoffs are denominated in USD. These were recalibrated to the new payoffs for inclusion in the main experiment. Second, although message (3) had a recognition rate of less than 50%, it alluded to some risk on the side of B, who in reality bears no risk. We thus decided to exclude this message to avoid potential confusion among the participants.

We compute the individual scores of participants by assigning one point to each correctly classified message. The observed mean score is 3.23 (the support is $[0, 6]$) and the expected value in case of random response is three. We test the null hypothesis that participants’ median score is not different from random guessing through a sign-rank test. We fail to reject the null hypothesis ($z = 1.668$, $p = 0.10$).

Table 1: List of messages generated by ChatGPT used for the Turing test

Message	Written by human (%)
(1) Trust me, if you choose IN, we both can win big! ;)	62%
(2) I promise I'll choose ROLL if you go for IN. We'll both have a higher chance of getting more money. Let's make it happen! 👍	75%
(3) I've been thinking a lot about our game, and I believe that if you choose IN and I choose ROLL, we could increase our combined earnings significantly. I know there's a bit of risk involved, but the potential reward for both of us is worth it. I assure you that I'll choose ROLL and stick to my word. Let's take this opportunity to maximize our profits together!	50%
(4) Trust me when I say I'll ROLL if you choose IN. This gives you a 5/6 chance of earning \$12, which is \$7 more than the alternative. There's no reason for me to lie – I assure you, I'll choose ROLL.	50%
(5) Even though we've never met and won't meet again, I promise you that I will choose ROLL if you pick IN. I have no reason to deceive you, as we'll never cross paths again. It's more beneficial for me to be honest in this one-time interaction. Let's make the most of it together.	35%
(6) Given that we don't know each other and have no past history, I can see why it's difficult for you to trust me. However, I genuinely want us both to succeed, and I'll choose ROLL if you go with IN. Let's take this opportunity to cooperate and maximize our potential gains.	46%
(7) I understand that it's difficult to trust someone you've never met, especially in a one-time game like this. But I want you to know that I'm putting myself out there and trusting you as well. By choosing ROLL, I'm taking a risk, too. I believe we can both benefit if we work together, even though we'll never meet again. Let's take a chance on each other.	54%
(8) By picking IN, you open up the possibility of earning more. I'll choose ROLL, and with a 5/6 chance, you could earn \$12. It's a great opportunity for both of us to benefit.	41%
(9) Greetings, fellow player! Feel free to make your own decision, but I believe that if you choose IN and I select ROLL, we can maximize our earnings together. With a 5/6 chance of success, it's worth a shot. Hoping you're doing well in this game of chance. Best wishes from a fellow player.	15%
(10) Trust me, choose IN and I'll ROLL. You have my word on this.	77%

2.2 Main experiment

The main experiment consists of three treatments. In the Control condition, participants simply play the modified trust game as outlined at the beginning of Section 2. In the Communication condition, B players can send a free-form written message to A players. A can read this message before deciding to play IN or OUT. In the AI condition, B players can opt to either write a message or select one of the five promises to ROLL, which were generated by ChatGPT. Both parties were informed about the availability of AI-generated messages. The instructions also mentioned that “in a previous study on Prolific, a representative sample of UK participants viewed these messages, and no more than 50% accurately identified them as AI-generated.” In both Communication and AI, B participants had the option to refrain from writing a message. In this case, A participants would simply see “No Message” on their screens.

We collected the trustors’ first-order beliefs about their own decision to choose IN and the trustees’ decision to roll the dice. Additionally, we gathered the trustees’ first-order beliefs about the trustors’ decision to choose IN and their second-order beliefs about the trustees’ decision to roll the dice (i.e., the trustees’ beliefs about the trustors’ beliefs about the trustees’ choice). All beliefs were obtained as estimates of the share of total subjects in each role, without incentives. In the AI treatment, Bs were also prompted to rate the likelihood that the received message was written by AI, using a four-item Likert scale. A post-experimental questionnaire collected self-reported socio-demographics and participants’ familiarity with ChatGPT.

To ensure attentiveness and deter bots, participants who failed a comprehension question on their first attempt after reading the instructions were excluded from

continuing. The instructions were modeled after the original study by Charness and Dufwenberg (2006), with minor modifications. Since the original experiment only had Communication and Control treatments, for the AI treatment, we closely mirrored the Communication treatment, adding the necessary sentences to explain the ChatGPT messages.⁷

Following the original protocol, we first collected all B decisions—asking them to assume that As chose IN—then gathered data from As in separate studies. Payment for As was processed immediately after the study’s completion (0.5 GBP base payment plus a bonus based on decisions made), while Bs received their base payment post-study, with bonuses disbursed within a week. On average, the experiment lasted almost 5 minutes, with an average payment of approximately 2.25 GBP. Data collection for the main experiment occurred between the 11th and 30th of July 2023.⁸

The experiment was programmed using Lioness (Giamattei et al., 2020). We conducted a power analysis with Gpower (Faul et al., 2007). The sample size was predetermined before the experiment, based on a power analysis. This ensured we could detect effects of the size taken from Charness and Dufwenberg (2006), with 90% power and an α of 0.05 for a test of proportions. Assuming the estimated treatment impact from the original study as the minimum detectable effect, we collected between 179 and 182 independent pairs per treatment.

2.3 Behavioral Hypotheses

Following our design, we formulate the following hypotheses for testing:

⁷The protocols can be found in Appendix 6.2.2.

⁸The University of Bergamo granted Institutional Review Board (IRB) approval for both the pilot and main experiments on the 19th of April 2023 (N4-2023). All participants provided informed consent.

H_0 : Participants unable to communicate with their partners will exhibit significantly lower levels of trust than those who can communicate.

H_1 : Participants aware they might receive pre-written messages from an AI will exhibit significantly lower levels of trust than those engaging in free-form communication.

H_2 : Access to pre-written messages by the trustee will not affect the level of trustworthiness in the game compared to the baseline.

H_3 : Trustworthy participants will be more inclined to write their own messages rather than use pre-written ones.

Aiming to replicate the main finding of Charness and Dufwenberg (2006), H_0 suggests that when participants can communicate, they will trust more than when they cannot, even in an online setting. Following the original article, we assume this to be the case because messages may contain promises. H_1 and H_2 are the main hypotheses of our study. H_1 is based on the idea that a promise, being a form of commitment, increases the trust of the recipient. This increase might be because the person making the promise does not want to feel guilty later (as in Charness and Dufwenberg, 2006) or because they genuinely want to keep their word (like in Vanberg, 2008; Ederer and Stremitzer, 2017).⁹ However, we hypothesize that if an AI makes the promise, it will not have the same effect on trust. This is not just because the promise is outsourced, but because it is given to a machine, which might make it seem less sincere. H_2 hypothesizes that trustees, aware of the indistinguishability between AI

⁹Note that our study was not designed to separate these two mechanisms, as both lead to the same outcome.

and human messages, will anticipate the same trust levels as in the condition without communication. Lastly, H_3 infers that intrinsically trustworthy individuals will favor personal messages, speculating that AI-generated messages might signal lower trustworthiness. Therefore, they might signal their trustworthiness more effectively with a personalized message than with a standardized one.

We preregistered these hypotheses on May 5, 2023.¹⁰ We will now present the results of our empirical analysis.

3 Results

In Section 3.1 we analyze the main experimental results and the overall differences across treatments. In Section 3.2 we focus on how the composition of messages changed across the two treatments with communication, and in Section 3.3, we analyze efficiency.

3.1 Trust and Trustworthiness

Result 2. *Free-form communication significantly increases trust and trustworthiness compared to the control.*

Result 3. *Access to AI messages increases trustworthiness above control and communication, but does not significantly increase trust compared to the control.*

Table 2 reports the levels of trust and trustworthiness by treatment. In the control treatment, 83 out of 182 B players chose to roll the dice (46%). When allowed to write a message, 118 out of 178 decided to roll (66%). The increase is even more pronounced

¹⁰Pre-registration is available at https://aspredicted.org/DH7_RJ4, ID:131199.

in the AI condition, where 137 out of 181 chose to roll (76%). A non-parametric test of equality of distribution rejects the null hypothesis for Communication and Control ($\chi^2(1) = 15.62, p < 0.001$), for AI and Control ($\chi^2(1) = 34.41, p < 0.001$), and for AI and Communication ($\chi^2(1) = 3.85, p = 0.050$).

Table 2: Rate of Trust and Trustworthiness across treatments.

Treatment	A's IN Rate	B's ROLL Rate	(IN, ROLL)
Control	42%	46%	19%
Comm	56%	66%	39%
AI	51%	76%	42%

In the Control, 77 out of 182 (42%) chose IN. When Bs are allowed to send messages, we observe an increase in trust to 56% (100 out of 178). In the AI condition, 93 out of 181 (51%) select IN. We formally test the null hypothesis that messages will not affect trust and reject the null for the communication treatment ($\chi^2(1) = 6.93, p = 0.008$), but not for the AI treatment ($\chi^2(1) = 3.00, p = 0.083$). Even though Communication exhibits a higher level of trust than AI, we cannot statistically distinguish between the two treatments ($\chi^2(1) = 0.83, p = 0.362$).

Comparing the results with our pre-registered hypotheses, the data support our claim that communication increases trust (i.e., we verify H_0). Additionally, we find that AI does not change the level of trust compared to the baseline, although barely above the threshold of statistical significance (i.e., we partially verify H_1). Furthermore, the data indicate that under AI, the trustworthiness of a trustee is not diminished; in fact, it exceeds the control condition (i.e., we partially verify H_2). However, our expectations differed regarding the use of messages by trustworthy participants.

We did not find consistent evidence that they utilize personalized messages (i.e., we do not verify H_3), as detailed in the next subsection.

3.2 Classification of the messages

Result 4. *In Communication, more than 50% of trustees write promises. AI significantly increases the rate of promises among messages.*

Result 5. *A non-negligible share (15%) of trustees in Communication recommends OUT.*

In Table 3, we classify messages and the corresponding behavior based on sending or receiving a message. A limited number of participants refrained from sending a message to As: eight in the Communication and 32 in the AI condition.

Table 3: Rate of trust and trustworthiness, by alternative message classifications

Classification	Communication		AI	
	IN (%)	ROLL (%)	IN (%)	ROLL (%)
No Message	57%	57%	28%	66%
Any Message	56%	67%	56%	78%
Promise	62%	79%	57%	78%
Non promise	49%	49%	27%	64%
Recommend “out”	37%	37%	$N = 1$	$N = 1$
Non recommend “out”	60%	72%	52%	76%
Use of AI message	NA	NA	57%	78%
Non AI (Written or No message)	56%	66%	38%	69%

We categorized the messages to identify those containing promises. By design, the messages prepared by ChatGPT (detailed in Section 2.1) are included in this category. Previous studies show that promises enhance trust (Charness and Dufwenberg, 2006;

Vanberg, 2008). Our data align with this hypothesis. While trust in the control stands at 42%, it rises to 62% when human-written communication includes a promise ($\chi^2(1) = 10.01, p = 0.002$). This rate is also (marginally) significantly higher than the trust level associated with non-promise messages (49%, $\chi^2(1) = 3.14, p = 0.076$). Trustworthiness also increases, reaching 79%, compared to the control’s 46% ($\chi^2(1) = 29.44, p < 0.001$) and the 49% for non-promises ($\chi^2(1) = 16.49, p < 0.001$).

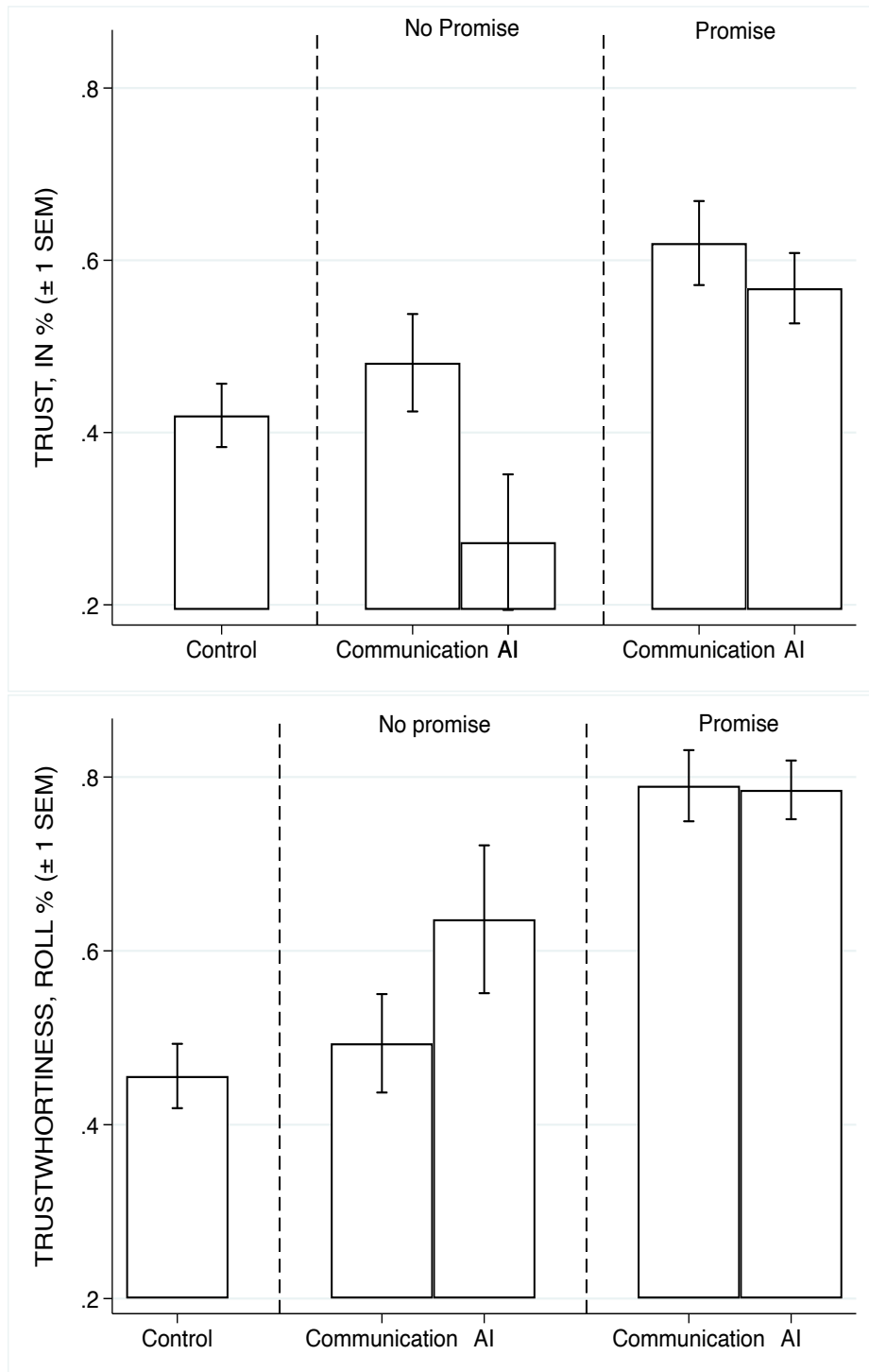
In the AI treatment, the trust level for promises is 57%, exceeding the control’s level ($\chi^2(1) = 6.82, p = 0.009$) and the 27% for non-promises ($\chi^2(1) = 9.39, p = 0.002$). Trustworthiness reaches 78%, compared to the control’s 46% ($\chi^2(1) = 36.62, p < 0.001$) and 64% for non-promises ($\chi^2(1) = 3.19, p = 0.074$).

Notably, we find messages suggesting to play OUT. They account for 15% of all messages in the communication treatment and have not been documented in similar experimental contexts. Unsurprisingly, they lead to significantly lower rates of trust and trustworthiness, both at 37%. These messages may be compatible with the theory of aversion to inequality (Fehr and Schmidt, 1999), or agents bearing the cost of overcoming temptations, as in the “deal me out” concept Dufwenberg et al., 2017. However, the introduction of AI eliminates these messages, suggesting that the presence of AI-generated promises may introduce additional framing effects.

Result 6. *AI-generated promises are less believed even though they are trustworthy, but they are also more frequent. Hence, the overall level of trust is the same between Communication and AI.*

As we stated in our pre-registration, we expected AI generated promises to wipe out the effect of messages, leading to levels of trust that cannot be distinguished from the baseline but systematically lower than in the free-form communication treatment.

Figure 1: Share of Trust and Trustworthiness across treatments.



Evidence for this hypothesis is mixed: the level of trust in the AI treatment is intermediate between and not statistically different from the other two treatments.

The effect of access to Chat GPT generates complex effects on trust. This is a description of the effects in our sample, as summarized in Figure 1, where we distinguish messages in promises and other messages, and we report their frequency and the associated level of trust and trustworthiness.

AI induced a more prominent use of promises (82%), possibly because of a lower writing cost. Among those who promised, only one of ten chose to write their own message, as explained in the previous subsection. The level of trust associated with those promises tends to be higher than in the control, but lower than promises within the communication treatment. Only one Chat GPT message produced a rate of IN higher than promises in communication (by 2 percentage points, *pp*), while the other ones have a deficit of trust that can be quantified from 5 to 15*pp*. Notably, the deficit of trust in AI promises spills over to human-elaborated promises in the AI treatment (58% against 62% of promises in Communication). In other words, promises generated with AI have a lower return, but are also less expensive to adopt. Moreover, they are not less worthy as the trustworthiness increases between Communication and AI. More promises but less believed. The net effect is zero.

Three notable observations. First, in the AI treatment, we witness a decline in trust among senders when recipients abstain from sending a message. Specifically, only 9 out of 32 chose IN (28%). This highlights a key distinction between human-written and AI-mediated communication: trustors appear more sensitive to the absence of messages in the AI context, possibly because the perceived cost of sending a message is lower. However, it's crucial to note that the limited number of observations

might exaggerate these differences. Second, using one of the five promises generated by ChatGPT is linked with higher trust and trustworthiness compared to the control and other messages within the same treatment. Specifically, trustworthiness levels are 78%, 46% ($\chi^2(1) = 33.39, p < 0.001$), and 69% ($\chi^2(1) = 1.65, p = 0.198$), while trust levels are 57%, 42% ($\chi^2(1) = 6.16, p = 0.013$), and 38% ($\chi^2(1) = 4.87, p = 0.027$). This effect might be another manifestation of the promise effect or a reflection of the perceived quality of the message. In fact, senders view the messages as more trustworthy ($\chi^2(2) = 17.85, p < 0.001$). Third, one particular message (#2) garnered less trust. This could be attributed to the message’s potential to mislead regarding the game of chances. Intriguingly, in our pre-experiment, this message was perceived as human-generated 75% of the time. We interpret this as participants attributing potential interpretation errors more to humans than to artificial intelligence.

3.3 Efficiency

To gain a deeper understanding of the results, it’s instructive to examine the outcome of the contract. The only Subgame Perfect Equilibrium of the Stage Game is (OUT, Don’t Roll if IN). This equilibrium is observed 31% of the time in the Control, 16% in Communication, and 15% in AI. The distribution across treatments is significantly different, as confirmed by the chi-squared test ($\chi^2(2) = 17.03, p < 0.001$). When we incorporate the psychological game with Guilt Aversion as described by (Charness and Dufwenberg, 2006), another equilibrium emerges, which is the most efficient (In, Roll if IN). This is observed 19% of the time in the Control, 39% in Communication, and 42% in AI. Once again, the distribution is significantly different across treatments ($\chi^2(2) = 26.13, p < 0.001$).

4 Mechanisms

4.1 Commitment or guilt aversion?

Result 7. *Promises increase trustworthiness through the guilt aversion and the commitment effect. AI-generated promises induce a commitment effect comparable to free-form communication. Less trustworthy participants do not preferentially opt for pre-written messages.*

The introduction of AI-mediated communication resulted in a 30pp increase in trustworthiness compared to the control, an effect larger than standard communication by 10pp. Participants leverage the access to communication to deliver promises. Promises increase trustworthiness, a phenomenon previously documented in the literature (Charness and Dufwenberg, 2006). Notably, promises are equally maintained whether participants craft them independently in the communication treatment or select them from AI-generated options.

Contrary to our expectations, put forth in our pre-registered hypotheses, participants' use of pre-written messages was not based on their trustworthiness levels. Our claim was grounded on the hypothesis that participants want to appear more genuine when making a commitment. However, the data refute this: within the AI treatment, promises generated by ChatGPT exhibit a trustworthiness level of 78.29%, equal to the 78.95% of human written promises in the AI treatment ($\chi^2(1) = 0.00$, $p = 0.949$), and nearly mirroring the 79.00% in the Communication treatment ($\chi^2(1) = 0.02$, $p = 0.897$).

Two primary, non-competing explanations for participants honoring their word are guilt aversion and commitment (Vanberg, 2008; Di Bartolomeo et al., 2019, 2023;

Ellingsen and Johannesson, 2004). The commitment theory suggests an intrinsic preference to uphold one’s word, while guilt aversion implies decisions are swayed by potential guilt from betraying legitimate expectations. Table 4 consolidates evidence supporting both guilt aversion and commitment. The data leans towards an expectation-based rationale: in both communication and AI conditions, second-order beliefs (B’s expectation of A’s expectation that B will roll) contingent on promises stand at 60.59% and 62.59%, respectively. Second-order beliefs are significantly larger than in the control ($z = -2.33$, $p = 0.020$ and $z = -3.19$, $p = 0.001$). Moreover, second-order beliefs are considerably higher for promises than for other messages or their absence in both communication (56.29%) and AI (54.21%) treatments. The latter are not statistically different than the control’s ($z = -0.69$, $p = 0.487$ and $z = 0.03$, $p = 0.979$).

However, distinguishing between commitment and guilt aversion remains challenging, and the former’s contribution to the results cannot be ruled out. Notably, second-order beliefs for promises are statistically consistent across both message treatments ($z = -0.44$, $p = 0.660$). Consequently, beliefs are unlikely the sole factor behind the 10pp rise in AI’s trustworthiness over communication. Part of this effect stems from composition: while the behavior post-promise remains consistent across treatments, the propensity to make a promise is roughly 26pp higher in AI.¹¹

Among participants who refrained from sending a promise in the Communication treatment, approximately half of Bs opted to ROLL, a trustworthiness level mirroring the control’s ($\chi^2(1) = 0.42$, $p = 0.515$). In the AI treatment, 21 out of 33 chose to

¹¹For promises, second-order beliefs align with first-order beliefs in communication (62.85% vs. 60.59%, $z = .063$, $p = 0.525$), but diverge significantly in AI (62.72% vs. 56.66%, $z = 1.95$, $p = 0.050$)

roll, a proportion (64%) that seems to exceed both the control’s and communication’s, though not statistically distinct ($\chi^2(1) = 3.64, p = 0.057$ and $\chi^2(1) = 1.74, p = 0.188$ respectively).

Table 4: Promises and beliefs, by treatment

	Comm	AI	Control	Comm v. AI
Promise	56%	82%		$\chi^2 = 27.51, p < 0.001$
$ROLL Promise$	79%	78%		$\chi^2 = 0.01, p = 0.907$
$E_B[E_A[ROLL]]$	63%	64%	49%	$z = 0.05, p = 0.954$
$E_B[IN]$	55%	60%	55%	$z = 1.90, p = 0.056$
$E_B[E_A[ROLL] Promise]$	61%	63%		$z = -0.44, p = 0.660$
$E_A[ROLL Promise]$	63%	57%		$z = 1.89, p = 0.058$
$E_B[IN Promise]$	57%	63%		$z = 1.89, p = 0.058$

4.2 Treatment effect decomposition

The main factor influencing the observed level of trustworthiness is the varying propensity to make a promise between Communication and AI. The reduced cost of promising in the AI treatment led to a surge in the rate of promises from 56% to 82%, statistically significant ($\chi^2(1) = 27.51, p < 0.001$).

To distinguish among the three components, we introduce the following decomposition. Let $\mathbb{E}[ROLL^{AI} - ROLL^C]$ represent the treatment effect of AI on trustworthiness compared to communication. Using the law of iterated expectations, we can rewrite $\mathbb{E}[ROLL]$ as $\mathbb{E}[\mathbb{E}[ROLL|message]] = p \cdot \mathbb{E}[ROLL|promise] + (1 - p) \cdot \mathbb{E}[ROLL|empty]$,¹² with p^j being the probability of making a promise in treatment j .

¹²We follow Charness and Dufwenberg (2006) and define “empty” the non promises.

With some algebra, we get

$$\begin{aligned}
& \mathbb{E}[ROLL^{AI} - ROLL^C] = \\
& \mathbb{E}[ROLL^{AI} - ROLL^C | \text{empty}] + \\
& p^C (\mathbb{E}[ROLL^{AI} - ROLL^C | \text{promises}] - \mathbb{E}[ROLL^{AI} - ROLL^C | \text{empty}]) + \\
& (p^{AI} - p^C) (\mathbb{E}[ROLL^{AI} | \text{promises}] - \mathbb{E}[ROLL^{AI} | \text{empty}])
\end{aligned} \tag{1}$$

The first term represents the baseline difference, which could be a spillover effect (i.e., we frame participants on the promise even if they do not make one). This term positively contributes to trustworthiness by 13.64*pp*. The second term is the differential between promise and non-promise across treatments, indicating that the contribution of promises to trustworthiness is more pronounced in communication than in AI, with a difference of $-8.01pp$. The third term captures the composition effect, representing the increase attributed to the changed likelihood of delivering a promise, amounting to 3.77*pp*.

Applying the formula in Equation (1), the difference in trust between AI and communication ($-4.80pp$) can be decomposed into three components. The first component is negative ($-21.45pp$), indicating that the absence of a promise in the AI treatment diminishes trust. The second component is positive (9.10*pp*), suggesting that the difference in trust between promise and non-promise messages is more pronounced in AI than in Communication, mostly driven by the *drop* in trust in the absence of a promise. Lastly, the composition effect, attributed to a higher frequency of promises in AI, is positive (7.55*pp*).

5 Concluding remarks

This study provides insights into the role of AI-mediated communication in shaping trust dynamics. While there are prevalent concerns about the potential dilution of genuine human interactions with the advent of AI, our findings offer a more optimistic view. Specifically, AI-generated promises have shown their capability to foster trust, aligning closely with the outcomes of human-written promises. This observation underscores the potential of AI to complement, and in certain contexts, even enhance the trustworthiness inherent in traditional communication methods. As we integrate AI into communication platforms, it's essential to approach its implementation with careful consideration, especially given the intricacies observed, such as trustors' heightened sensitivity to the absence of AI messages. Our research not only contributes to the current discourse on AI and trust but also sets the stage for future investigations into the multifaceted interactions between AI and human communication.

As AI's societal imprint grows, understanding these dynamics becomes crucial. Our research serves a dual purpose. First, we aim to understand AI-mediated interactions in economic transactions, building on prior studies of human-robot interactions (Cominelli et al., 2021). Second, we explore how to shape democratic inputs for Artificial Intelligence. While AI's convenience might encourage a culture of delegation, guiding AI toward efficient outcomes could foster successful partnerships.

References

- Acemoglu, Daron and Todd Lensman**, “Regulating Transformative Technologies,” Technical Report, National Bureau of Economic Research 2023.
- , **David Autor, Jonathon Hazell, and Pascual Restrepo**, “Artificial intelligence and jobs: evidence from online vacancies,” *Journal of Labor Economics*, 2022, *40* (S1), S293–S340.
- Ahmad, Sayed Fayaz, Heesup Han, Muhammad Mansoor Alam, Mohd Rehmat, Muhammad Irshad, Marcelo Arraño-Muñoz, Antonio Ariza-Montes et al.**, “Impact of artificial intelligence on human loss in decision making, laziness and safety in education,” *Humanities and Social Sciences Communications*, 2023, *10* (1), 1–14.
- Ben-Ner, Avner and Louis Putterman**, “Trust, communication and contracts: An experiment,” *Journal of Economic Behavior and Organization*, 5 2009, *70*, 106–121.
- Berg, Joyce, John Dickhaut, and Kevin McCabe**, “Trust, Reciprocity, and Social History,” *Games and Economic Behavior*, 1995, *10* (1), 122–142.
- Brandts, Jordi, David J. Cooper, and Christina Rott**, *Communication in laboratory experiments*, Edward Elgar Publishing,
- , **Matthew Ellman, and Gary Charness**, “LET’S TALK: HOW COMMUNICATION AFFECTS CONTRACT DESIGN,” *Journal of the European Economic Association*, 8 2016, *14*, 943–974.
- Charness, Gary and Daniela Grieco**, “Creativity and Artificial Intelligence,” Technical Report, Working paper 2023.
- **and Martin Dufwenberg**, “Promises and Partnership,” *Econometrica*, 2006, *74* (6), 1579–1601.
- **and** – , “Bare promises: An experiment,” *Economics Letters*, 5 2010, *107*, 281–283.
- , **Brian Jabarian, and John List**, “Generation Next: Experimentation with AI,” Technical Report, The Field Experiments Website 2023.
- Cominelli, Lorenzo, Francesco Feri, Roberto Garofalo, Caterina Giannetti, Miguel A Meléndez-Jiménez, Alberto Greco, Mimma Nardelli,**

- Enzo Pasquale Scilingo, and Oliver Kirchkamp**, “Promises and trust in human–robot interaction,” *Scientific reports*, 2021, 11 (1), 9687.
- Crawford, Vincent P. and Joel Sobel**, “Strategic Information Transmission,” *Econometrica*, 1982, 50 (6), 1431–1451.
- Di Bartolomeo, Giovanni, Martin Dufwenberg, and Stefano Papa**, “Promises and partner-switch,” *Journal of the Economic Science Association*, 2023, 9 (1), 77–89.
- , – , – , and **Francesco Passarelli**, “Promises, expectations & causation,” *Games and Economic Behavior*, jan 2019, 113, 137–146.
- Dufwenberg, Martin, Maroš Servátka, and Radovan Vadovič**, “Honesty and informal agreements,” *Games and Economic Behavior*, 3 2017, 102, 269–285.
- Ederer, Florian and Alexander Stremitzer**, “Promises and expectations,” *Games and Economic Behavior*, 2017, 106, 161–178.
- Ellingsen, Tore and Magnus Johannesson**, “Promises, Threats and Fairness,” *The Economic Journal*, 2004, 114 (495), 397–420.
- Farrell, Joseph and Matthew Rabin**, “Cheap Talk,” *Journal of Economic Perspectives*, September 1996, 10 (3), 103–118.
- Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner**, “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,” *Behavior Research Methods*, 2007, 39 (2), 175–191.
- Fehr, Ernst**, “On the Economics and Biology of Trust,” *Journal of the European Economic Association*, 2009, 7 (2/3), 235–266.
- and **Klaus M. Schmidt**, “A theory of fairness, competition, and cooperation,” *Quarterly Journal of Economics*, 1999, 114 (3), 817–868.
- Giamattei, Marcus, Kyanoush Seyed Yahosseini, Simon Gächter, and Lucas Molleman**, “LIONESS Lab: a free web-based platform for conducting interactive experiments online,” *Journal of the Economic Science Association*, 2020, 6 (1), 95–111.
- Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli**, “Chatgpt outperforms crowd-workers for text-annotation tasks,” *arXiv preprint arXiv:2303.15056*, 2023.

- Gillespie, Nicole, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari**, “Trust in Artificial Intelligence: A global study,” 2023.
- Gneezy, Uri**, “Deception: The Role of Consequences,” *American Economic Review*, March 2005, *95* (1), 384–394.
- Goeree, Jacob K and Leeat Yariv**, “An experimental study of collective deliberation,” *Econometrica*, 2011, *79* (3), 893–921.
- Lundquist, Tobias, Tore Ellingsen, Erik Gribbe, and Magnus Johannesson**, “The aversion to lying,” *Journal of Economic Behavior and Organization*, 5 2009, *70*, 81–92.
- Menczer, Filippo, David Crandall, Yong-Yeol Ahn, and Apu Kapadia**, “Addressing the harms of AI-generated inauthentic content,” *Nature Machine Intelligence*, 2023, pp. 1–2.
- Puca, Marcello, Krista Jabs Saral, and Simone M. Sepe**, “The Value of Consensus. An Experimental Analysis of Costly Deliberation,” CSEF Working Papers, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy 2023.
- Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi**, “Can ai-generated text be reliably detected?,” *arXiv preprint arXiv:2303.11156*, 2023.
- Sally, D**, “Conversation and Cooperation in Social Dilemma,” *Rationality and Society*, 1995, *7*, 58–92.
- Shen, Xinyue, Zeyuan Chen, Michael Backes, and Yang Zhang**, “In chatgpt we trust? measuring and characterizing the reliability of chatgpt,” *arXiv preprint arXiv:2304.08979*, 2023.
- Turing, Alan Mathison**, “Mind,” *Mind*, 1950, *59* (236), 433–460.
- Turmunkh, Uyanga, Martijn J. Van den Assem, and Dennie Van Dolder**, “Malleable lies: Communication and cooperation in a high stakes TV game show,” *Management Science*, 10 2019, *65*, 4795–4812.
- Vanberg, Christoph**, “Why Do People Keep Their Promises? An Experimental Test of Two Explanations¹,” *Econometrica*, 2008, *76* (6), 1467–1480.

6 Online Appendix

6.1 ChatGPT query

[Prompt in bold]

Consider the following sequential game: Players label: A, B.

A chooses OUT: A receives \$5 and B receives \$5.

A chooses IN, B chooses DON'T ROLL: A receives \$0 and B receives \$14.

A chooses IN, B chooses ROLL, if die = 1: A receives \$0 and B receives \$10.

A chooses IN, B chooses ROLL, if die = 2, 3, 4, 5, or 6: A receives \$12 and B receives \$10.

To analyze this sequential game, let's first construct the game tree and then use backward induction to determine the optimal strategy for each player.

The game tree is as follows:

1. A chooses OUT:

- A receives \$5
- B receives \$5

2. A chooses IN:

a. B chooses DON'T ROLL:

- A receives \$0
- B receives \$14

b. B chooses ROLL:

i. If die = 1:

- A receives \$0
- B receives \$10

ii. If die = 2, 3, 4, 5, or 6:

- A receives \$12
- B receives \$10

Now, let's use backward induction to find the optimal strategy for each player:

1. First, consider B's decision, given that A chose IN:

a. If B chooses DON'T ROLL, B receives \$14.

b. If B chooses ROLL, B receives \$10 with a probability of 1 (since the die can only result in a value between 1 and 6).

Since $\$14 > \10 , B's optimal strategy is to choose DON'T ROLL if A chooses IN.

2. Now, consider A's decision:

- a. If A chooses OUT, A receives \$5.
- b. If A chooses IN, B will choose DON'T ROLL (based on the previous step), and A will receive \$0.

Since $\$5 > \0 , A's optimal strategy is to choose OUT.

In conclusion, the optimal strategy for player A is to choose OUT, and the optimal strategy for player B is to choose DON'T ROLL if A chooses IN. This results in both players receiving \$5 each.

Write 5 messages from B to convince A to play IN. Use different lengths and level of promises

1. Short and vague: "Trust me, if you choose IN, we both can win big! ;)"
2. Medium and direct: "Hey A, I promise I'll choose ROLL if you go for IN. We'll both have a higher chance of getting more money. Let's make it happen! 👍"
3. Long and convincing: "Dear A, I've been thinking a lot about our game, and I believe that if you choose IN and I choose ROLL, we could increase our combined earnings significantly. I know there's a bit of risk involved, but the potential reward for both of us is worth it. I assure you that I'll choose ROLL and stick to my word. Let's take this opportunity to maximize our profits together!"
4. Short and persuasive: "Go for IN, and I'll choose ROLL. We can earn up to \$22 combined! Let's take the chance."
5. Long and emotional: "A, as we embark on this game together, I want you to know that I value our partnership. I truly believe that if you choose IN and trust me to make the right decision, we can both come out ahead. I understand the risks, but I'm willing to do what it takes for our mutual benefit. So, please, have faith in me and choose IN, and I promise you won't regret it. Let's make a memorable move together!"

Write other 5 messages with different language tones

1. Formal and polite: "Dear Player A, I kindly request that you consider selecting the IN option in our game. Rest assured, I intend to make a choice that will prove advantageous for both of us. Your cooperation is greatly appreciated."
2. Casual and friendly: "Hey A, let's give IN a shot, alright? I've got a good feeling about it, and I'll do my best to make sure we both benefit from it. Let's go for it, buddy!"
3. Humorous and light-hearted: "Hey A, let's spice things up! Go for IN and I'll ROLL the die. Worst case, we'll have a good laugh, and best case, we'll be swimming in cash. What do you say? :D"
4. Enthusiastic and energetic: "A, let's do this! Choose IN and let's ride the wave"

of fortune together! I'm super pumped to make the right call for both of us. Ready, set, IN!"

5. Sincere and heartfelt: "A, I genuinely believe that if you choose IN, we'll both have a better chance of success. I'll do everything in my power to make sure we come out ahead. Please trust me on this, and let's make the most of this opportunity together."

Imagine that A replies with "I still don't trust you" can you write 5 messages that would surely convince A?

1. Appealing to logic and fairness: "I understand your hesitation, A. But let's consider the options rationally. If we both choose to cooperate and I select ROLL, our combined potential earnings will be higher. It's in my best interest to maximize the total payout, and I'm sure it is in yours as well. Let's trust each other and work together for the best outcome."

2. Offering assurance: "A, I know trust is hard to build, especially in a game like this. But I promise you that I will choose ROLL if you pick IN. To help build trust, let's play a few practice rounds where I demonstrate my commitment to choosing ROLL. Once you see that I am true to my word, we can proceed with confidence."

3. Acknowledging past behavior: "A, I know that in the past, I may not have always made the best decisions or lived up to my promises. But I've learned from my mistakes, and I'm committed to making the right choice this time. I genuinely want us both to succeed, and I'll choose ROLL if you go with IN. Let's start a new chapter of trust and cooperation."

4. Expressing vulnerability: "A, I understand that it's difficult to trust someone in this situation. But I want you to know that I'm putting myself out there and trusting you as well. By choosing ROLL, I'm taking a risk, too. I believe we can both benefit if we work together, and I'm willing to take that leap of faith. I hope you'll join me."

5. Sharing a personal story or connection: "A, I remember when we first met, and we shared stories about our lives and experiences. I felt a genuine connection with you. Since then, I've always believed that we can rely on each other, even in difficult situations. I know this game might be different, but I still trust you and hope you can trust me, too. Let's choose the path that has the potential to benefit both of us the most."

Rewrite these latter messages assuming that A and B never met and will never meet again after playing the game

1. Appealing to logic and fairness: "I understand your hesitation since we don't know each other. But let's consider the options rationally. If we both choose to cooperate and I select ROLL, our combined potential earnings will be higher. It's in my best interest to maximize the total payout, and I'm sure it is in yours as well."

Let's trust each other just this once and work together for the best outcome."

2. Offering assurance: "Even though we've never met and won't meet again, I promise you that I will choose ROLL if you pick IN. I have no reason to deceive you, as we'll never cross paths again. It's more beneficial for me to be honest in this one-time interaction. Let's make the most of it together."

3. Acknowledging the lack of history: "Given that we don't know each other and have no past history, I can see why it's difficult for you to trust me. However, I genuinely want us both to succeed, and I'll choose ROLL if you go with IN. Let's take this opportunity to cooperate and maximize our potential gains."

4. Expressing vulnerability: "I understand that it's difficult to trust someone you've never met, especially in a one-time game like this. But I want you to know that I'm putting myself out there and trusting you as well. By choosing ROLL, I'm taking a risk, too. I believe we can both benefit if we work together, even though we'll never meet again. Let's take a chance on each other."

5. Focusing on the shared experience: "Even though we've never met and will never meet again after this game, we find ourselves in this unique situation where we can choose to cooperate for mutual benefit. We don't have any reason to deceive each other, as there's nothing to gain after this interaction. Let's choose the path that has the potential to benefit both of us the most, and create a positive memory of this one-time encounter."

Based on the following 10 messages, produce 10 more messages: 1. Choose in, I will roll dice, you are 5/6 likely to get 2,3,4,5, or 6, → \$12. This way both of us will win something. 2. If you will choose "In", I will choose to roll. This way, we both have an opportunity to make more than \$5! :) 3. The fairest thing to do is if you opt "IN". Then I will proceed to choose "roll." That way you and I have 5/6 chances to make money for the both of us. That's much better than just making \$5 each. Increases both our chances. Thanks. 4. Choose In and I will Roll You have my word 5. Hey. OK I think that the best way for both of us to make a profit is for you to choose IN and for me to roll. That way we both make some money. There's no point in me not rolling because that would give you and me less profit. So I'm a roller if you're in :). 6. Choose "In" so we can both make some \$\$ What are the chances me rolling a 1? I'll try my best. 7. Hello fair stranger, anonymous partner Choose whatever you want. Far be it from me to influence your decision, but I think you should choose "in" and I should choose "roll" and we should take the chance at both earning as much as we can. 5 chances out of 6 say it'll work, and I'm totally broke, looking to rake in stray cash however I can. I feel the luck in the air. E In R I don't really have much

else to say. Hope you're doing well, whoever you are. Yes. That's all. Random note from random human 8. If you choose in then I'm going to choose roll. This gives you a $5/6$ chance of getting 12 dollars. That is 7 more than if you choose out. Since the money is free anyway – why not believe me. I'm don't lie – I promise I will choose roll. 9. Hi, well I'm going to Roll so you have at least a shot for more money. I hope it works out. 10. I'm choosing ROLL, which gives you a chance to get \$12 instead of \$5, so stay. It's a risk, but you could end up getting a lot more.¹³

1. By picking IN, you open up the possibility of earning more. I'll choose ROLL, and with a $5/6$ chance, you could earn \$12. It's a great opportunity for both of us to benefit.
2. If you go for IN, I'll definitely ROLL the dice. This way, we have a chance to earn more than just \$5 each. Let's maximize our potential earnings together! :)
3. Opting for IN and having me ROLL the dice increases our chances of making more money. With a $5/6$ probability of success, it's a much better option than just earning \$5 each. Let's work together to achieve the best outcome.
4. Trust me, choose IN and I'll ROLL. You have my word on this.
5. Hey there! I think the best strategy for both of us is for you to go IN and for me to ROLL. This way, we can make some money together. I have no reason not to ROLL, as it benefits us both. So, if you're IN, I'll ROLL. ;)
6. Go ahead and pick IN so we can both earn some cash! What are the odds of me rolling a 1? I'll do my best to make sure we both win.
7. Greetings, fellow player! Feel free to make your own decision, but I believe that if you choose IN and I select ROLL, we can maximize our earnings together. With a $5/6$ chance of success, it's worth a shot. Hoping you're doing well in this game of chance. Best wishes from a fellow player.
8. Trust me when I say I'll ROLL if you choose IN. This gives you a $5/6$ chance of earning \$12, which is \$7 more than the alternative. There's no reason for me to lie – I assure you, I'll choose ROLL.
9. Hello! I plan on choosing ROLL, so you have a chance to earn more money. Fingers crossed for a favorable outcome!
10. I'm going to ROLL, giving you an opportunity to earn \$12 instead of just \$5. So, take a chance and choose IN – it might be a risk, but the potential payoff is worth it.

¹³These messages are taken from Charness and Dufwenberg (2006).

6.2 Experimental instructions

6.2.1 “Turing test” experiment instructions

[Messages and content in Italic]

Welcome Thank you for participating in our study. Including the time for reading these instructions, the study will take about 5 minutes to complete. During this activity, please do not close this window or leave the web pages in any other way. If you do close your browser or leave the study, you will not be able to re-enter and we will not be able to pay you! For participating you will earn a guaranteed fee of £0.50.

Instructions In this study, you will be evaluating whether certain messages were generated by Artificial Intelligence (AI) or written by humans. The messages are related to a specific situation that other subjects have encountered and responded to, as described below. Your objective is to classify each message as either “AI-generated” or “written by humans” based on your judgment.

This is the Situation the messages refer to: You will be paired with another person. However, no participant will ever know the identity of the person with whom he or she is paired. In each pair, one person will have the role of A, and the other will have the role of B. The amount of money you earn depends on the decisions made in your pair. On the designated decision sheet, each person A will indicate whether he or she wishes to choose IN or OUT. If A chooses OUT, A and B each receive \$5. If A chooses IN, each person B will indicate whether he or she wishes to choose ROLL or DON’T ROLL (a die). Note that B will take this decision without knowing A’s

decision. If A has chosen IN and B chooses DON'T ROLL, then B receives \$14 and A receives \$0. If B chooses ROLL, B receives \$10 and rolls a six-sided die to determine A's payoff. If the die comes up 1, A receives \$0; if the die comes up 2–6, A receives \$12. This information is summarized in the chart below:

	A Receives	B Receives
A chooses OUT	\$5	\$5
A chooses IN, B chooses DON'T ROLL	\$0	\$14
A chooses IN, B chooses ROLL, die = 1	\$0	\$10
A chooses IN, B chooses ROLL, die = 2, 3, 4, 5, or 6	\$12	\$10

Before A's decision concerning IN or OUT, B has the option to send a message to A.

[Task repeated six times]

Please read the following message and indicate whether you believe it was written by a human or generated by an AI system.

[Random message selected from the list]

In your opinion, this message has been:

- written by a human
- generated by an AI system

[Demographics questions]

Feedback Thank you for completing the study. You have answered correctly to XX questions. You will receive £0.15 for each correct question on top of the £0.50 fee for participating.

6.2.2 Main experiment, Role B

[Messages and content in Italic]

Stage 1: Welcome Thank you for participating. This study aims to understand how people make decisions in a particular situation.

The study should take a few minutes and you will receive £0.50 for completing it.
You may also receive additional money, depending on the decisions made by you and another person participating in another study on Prolific you are paired with.

At the end of this study, you will receive a code to collect your payment via Prolific. Upon completion of both studies, this additional amount will be paid to you as a bonus.

Stage 2: Decision Task In each pair, one person will have the **role of A**, and the other will have the **role of B**. The amount of money you earn depends on the decisions made in your pair.

YOU have the role of B

Each person A will indicate whether he or she wishes to choose **IN** or **OUT**. **If A chooses OUT, A and B each receive a bonus of £1.25.** Each person B will indicate whether he or she wishes to choose **ROLL** or **DON'T ROLL** (a virtual die).

Note that B will not know whether A has chosen IN or OUT; however, since B's decision will only make a difference when A has chosen IN we ask B's to presume (for the purpose of making this decision) that A has chosen IN.

If A has chosen IN and B chooses DON'T ROLL, then B receives £3.50 and A receives £0.

If B chooses ROLL, B receives £2.50 and the result of a roll of a six-sided die run by a computer determines A's payoff. If the die comes up 1, A receives £0;

if the die comes up 2–6, **A receives £2.5** (all of these amounts are in addition to the £0.5 participation fee).

This information is summarised in the Table below:

	A Receives	B Receives
A chooses OUT	£1.25	£1.25
A chooses IN, B chooses DON'T ROLL	£0	£3.50
A chooses IN, B chooses ROLL, die = 1	£0	£2.50
A chooses IN, B chooses ROLL, die = 2, 3, 4, 5, or 6	£0	£2.50

Comprehension question (if you fail to answer this question correctly we will not be able to pay you):

Suppose that A chooses IN. Which of the following statements is FALSE?

- A can get £0
- A gets £1.25 and B gets £1.25¹⁴
- B gets at least £2.50
- B gets either £2.50 or £3.50

Stage 3: Choice

- *Treatment: Communication*

Before A decides whether to go with IN or OUT, you (as player B) can send a message to player A.

Go ahead and write your message clearly and meaningfully. If your message is not logical we cannot forward it to player A. Remember, player A will read your message before making their choice. Write your message to A below:

[Message text]

¹⁴This is the correct answer.

- *Treatment: AI*

Before A decides whether to go with IN or OUT, you (as player B) can send a message to player A.

You can send pre-written promise messages to ROLL. These messages have been generated by ChatGPT. In a previous study on Prolific, a sample of UK participants was shown these messages, and no more than 50% of them correctly identified them as being generated by Artificial Intelligence. (A is aware that generated messages are available, but won't know if you, as B, chose one of them). You can also write your own message by selecting "Write own message." In that case, go ahead and write your message clearly and meaningfully. Remember, player A will read your message before making their choice.

- I promise I'll choose ROLL if you go for IN. We'll both have a higher chance of getting more money. Let's make it happen!
- Trust me, if you choose IN, we both can win big! ;)
- Trust me, choose IN and I'll ROLL. You have my word on this.
- Trust me when I say I'll ROLL if you choose IN. This gives you a 5/6 chance of earning £3, which is £1.75 more than the alternative. There's no reason for me to lie – I assure you, I'll choose ROLL.
- I've been thinking a lot about our game, and I believe that if you choose IN and I choose ROLL, we could increase our combined earnings significantly. I know there's a bit of risk involved, but the potential reward for both of us is worth it. I assure you that I'll choose ROLL and stick to my word. Let's take this opportunity to maximize our profits together!
- Write your own message:

[Message text]

As a reminder, here a summary of possible earnings:

	A Receives	B Receives
A chooses OUT	£1.25	£1.25
A chooses IN, B chooses DON'T ROLL	£0	£3.50
A chooses IN, B chooses ROLL, die = 1	£0	£2.50
A chooses IN, B chooses ROLL, die = 2, 3, 4, 5, or 6	£0	£2.50

If A plays IN, what would you choose?

- ROLL
- DON'T ROLL

Stage 4: Guess-1 We now ask you, as B, to make guesses about the percentage of A's who chose IN.

I guess that the percentage of A's who chose IN is __%

[slider]

Stage 5: Guess-2 We will ask A's to make guesses about the percentage of B's who chose ROLL. We ask you now to guess some of the average guesses made by those A's who chose IN.

For A's who chose IN, I guess that the average guess about the percentage of B's who chose ROLL is __%

[slider]

Stage 6: Demographics *[Basic demographic questions]*

How often have you used ChatGPT over the last three months?

- Never
- Rarely
- Sometimes
- Often
- Regularly

How well does the following statement describe you as a person: “As long as I am not convinced otherwise, I assume that people have only the best intentions.” Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.

Stage 7: End This is the end of this study. The computer will randomly match you with a player A recruited in a different study on Prolific. Your choices and the choices made by the other player will determine your payment. To be eligible to receive your payment please copy the following code and paste it into Prolific.

[Prolific code]

Once you have done that, you can close this window. Thank you for your participation!

6.2.3 Main experiment, Role A

[Messages and content in Italic]

Stage 1: Welcome Thank you for participating. This study aims to understand how people make decisions in a particular situation.

The study should take a few minutes and you will receive £0.50 for completing it. You may also receive additional money, depending on the decisions made by you and another person participating in another study on Prolific you are paired with. At the end of this study, you will receive a code to collect your payment via Prolific. Upon completion of both studies, this additional amount will be paid to you as a bonus.

Stage 2: Decision Task In each pair, one person will have the **role of A**, and the other will have the **role of B**. The amount of money you earn depends on the decisions made in your pair.

YOU have the role of A

Each person A will indicate whether he or she wishes to choose **IN** or **OUT**. **If A chooses OUT, A and B each receive a bonus of £1.25.** Each person B will indicate whether he or she wishes to choose **ROLL** or **DON'T ROLL** (a virtual die). Note that B will not know whether A has chosen IN or OUT; however, since B's decision will only make a difference when A has chosen IN we ask B's to presume (for the purpose of making this decision) that A has chosen IN. **If A has chosen IN and B chooses DON'T ROLL, then B receives £3.50 and A receives £0.** **If B chooses ROLL, B receives £2.50** and the result of a roll of a six-sided die run by a computer determines A's payoff. If the die comes up 1, A receives £0; if the die comes up 2–6, A receives £2.5 (all of these amounts are in addition to the £0.5 participation fee).

This information is summarised in the Table below:

	A Receives	B Receives
A chooses OUT	£1.25	£1.25
A chooses IN, B chooses DON'T ROLL	£0	£3.50
A chooses IN, B chooses ROLL, die = 1	£0	£2.50
A chooses IN, B chooses ROLL, die = 2, 3, 4, 5, or 6	£0	£2.50

Comprehension question (if you fail to answer this question correctly we will not be able to pay you.)

Suppose that A chooses IN. Which of the following statements is FALSE?

- A can get £0
- A gets £1.25 and B gets £1.25
- B gets at least £2.50
- B gets either £2.50 or £3.50

Stage 3: Info

- *Treatment: Communication*

Before you (in the role of A) decide whether to go with IN or OUT, B can send you a message. Please click Continue to see the message and to make your decision.

- *Treatment: AI*

Before you (in the role of A) decide whether to go with IN or OUT, B can send you a message. B can send pre-written promise messages to ROLL. These messages have been generated by ChatGPT. In a previous study on Prolific, a representative sample of UK participants was shown these messages, and no more than

50% of them correctly identified them as being generated by an Artificial Intelligence.

B can also write his own message. Please click Continue to see the message and to make your decision.

Stage 4: Choice

- *Treatment: Communication & AI*

This is the message to you from B:

[Message text]

As a reminder, here a summary of possible earnings:

	A Receives	B Receives
A chooses OUT	£1.25	£1.25
A chooses IN, B chooses DON'T ROLL	£0	£3.50
A chooses IN, B chooses ROLL, die = 1	£0	£2.50
A chooses IN, B chooses ROLL, die = 2, 3, 4, 5, or 6	£0	£2.50

What would you choose?

- IN
- OUT

Stage 4: Guess-1 We now ask you, as A, to make guesses about the percentage of other A's in this study who chose IN.

I guess that the percentage of other A's in this study who chose IN is __% *[slider]*

Stage 5: Guess-2

I guess that the average guess about the percentage of B's who chose ROLL is __%

[slider]

Stage 6: Demographics [*Basic demographic questions*]

How often have you used ChatGPT over the last three months?

- Never
- Rarely
- Sometimes
- Often
- Regularly

How well does the following statement describe you as a person: "As long as I am not convinced otherwise, I assume that people have only the best intentions." Please use a scale from 0 to 10, where 0 means "does not describe me at all" and a 10 means "describes me perfectly". You can also use the values in-between to indicate where you fall on the scale.

Stage 7: Results The computer randomly matched you with a player B recruited in a different study on Prolific. Your choices and the choices made by the other player determined your payment. As a result of your and Participant B's choices, you will receive $\pounds Payoff$, in addition to your participation fee of $\pounds 0.5$.

Stage 8: Study completed This is the end of this study. To be eligible to receive your payment please copy the following code and paste it into Prolific.

[*Prolific code*]

Once you have done that, you can close this window. Thank you for your participation!