AI-Powered Promises: The Influence of ChatGPT on Trust and Trustworthiness *

Ivo Greevink, Theo Offerman, Giorgia Romagnoli

January 16, 2024

Abstract

Amidst the growing popularity of AI language models, we study the potential impacts of AI-facilitated communication on trust and trustworthiness. With a laboratory experiment, we compare traditional communication with communication assisted by ChatGPT in a between-subject design. We find that participants with access to ChatGPT more frequently make promises, while their promise-keeping rate diminishes. Overall we do not observe effects on trust and trustworthiness rates. However, we show that in the GPT treatment participants coordinate less often on the trust outcome and that promises no longer serve as a reliable indicator of honesty.

Keywords AI-powered communication, trust, trustworthiness, promises

*Acknowledgements: We thank Menusch Khadjavi, Jeroen van de Ven and the audience at the CREED seminar for helpful suggestions and comments. Financial support from the Amsterdam Center for Behavioral Change is gratefully acknowledged. Greevink: University of Amsterdam, Tinbergen Institute, i.o.greevink2@uva.nl; Offerman: University of Amsterdam, Tinbergen Institute, t.j.s.offerman@uva.nl, Romagnoli, University of Amsterdam, Tinbergen Institute, g.romagnoli@uva.nl.

1 Introduction

The share of people using AI-powered language models in their daily interactions is increasing at tremendous speed. In January 2023, just two months after its release, ChatGPT already counted 100 million active users making it the fastest-growing consumer app in history (Dennean, 2023). ChatGPT is being integrated with many of the most popular communication apps, including social media platforms like Twitter, as well as business-focused tools such as Slack and Email. Communication experts observe that AI-powered language models are becoming widespread mediators in many forms of digital communication worldwide (Stone et al., 2016; Rahwan et al., 2019; Jakesch et al., 2023; Hohenstein et al., 2023).

This scenario opens up new fundamental questions regarding the way humans communicate with each other in the digital space: Will AI mediation alter the perception of, and commitment to the messages we exchange? For example, will the fact (or the mere possibility) that an email was written with the assistance of a language model change the way it is perceived, trusted and followed up on?

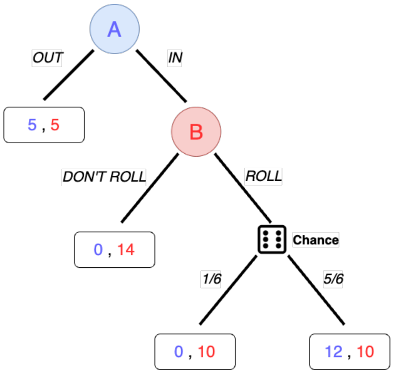
We explore these questions with a laboratory experiment based on the classic trust game of Charness and Dufwenberg (2006). We modify the setting by allowing the trustee to send a self-generated message to the trustor, and selectively allow access to ChatGPT assistance in crafting this message.

Previous experimental work emphasizes the positive effect of written communication on both trust and trustworthiness (Charness and Dufwenberg, 2006; Goeree and Zhang, 2014; Ismayilov and Potters, 2016; Ederer and Schneider, 2022). Furthermore, trust and trustworthiness are shown to be especially pronounced when promises of reciprocation are made. Here, the effect of AI mediation is particularly hard to predict and likely to produce two competing forces. On the one hand, language models may be quick to recognize the power of promises in generating trust and make abundant use of them in their suggestions, which could boost trust and trustworthiness. On the other hand, experimental studies have demonstrated that the positive impact of written communication diminishes significantly when players are not permitted to write their own promises (Charness and Dufwenberg, 2010; Di Bartolomeo et al., 2019; Chen and Zhang, 2021). Thus, trust and trustworthiness could decrease if promises that are no longer self-written (but merely suggested by AI) lose their power.

These considerations prompt additional research questions: Will the mediation of AI affect the frequency of promises and the way we commit to and trust written promises?

Our experiment is based on two treatments: a traditional 'Communication' treatment that only allows for self-generated messages and a novel 'GPT' treatment. Our GPT treatment mirrors the Communication treatment in all respects, with the exception that participants had access to ChatGPT while writing their messages. Trustor and trustee had mutual knowledge that ChatGPT could be accessed in the GPT treatment. See more details in Box 1.

We show that in the aggregate AI mediated communication does not change the trust and trustworthiness levels that are observed with standard communication. ChatGPT triggers two opposing forces among trustees. On the one hand they are substantially more likely to include a promise in their message. On the other hand, they are less likely to follow up on a promise. Access to ChatGPT reduces the intrinsic meaning of promises. As a consequence, promises become a less reliable indicator that the trustee is honest, and the rate at which promises help participants coordinate on the efficient outcome (IN, ROLL) is substantially reduced. Analyzing the content of the messages, we find that the higher the similarity between the transmitted message and the suggestion of chatGPT, the less often the sender follows up by being trustworthy. This suggests that people can regain a feeling of ownership and commitment if they actively rewrite a message suggested by chatGPT.



Box 1. We adopt the trust game with hidden action of Charness and Dufwenberg (2006) which has been frequently used to explore the effect of written communication on trust (Charness and Dufwenberg, 2010; Deck et al., 2013; Ismayilov and Potters, 2016; Di Bartolomeo et al., 2019; Ederer and Schneider, 2022). The game is depicted in the figure. Player A decides whether or not to trust player B (IN). Player B can reciprocate this trust by returning a large sum of money to A (ROLL), or play selfishly (DON'T ROLL). A chance component makes the choice of B not fully visible to A. Our Communication treatment follows the existing literature: Before A decides, player B can send a typed message to player A or leave the field blank (they can write any message they wish, as long as it does not reveal their identity). The GPT treatment we introduce is identical to the Communication treatment, with one exception: Both players are aware that player B can use the most recent premium version of ChatGPT before crafting their message. We provided B players with full ChatGPT access during the messaging phase, end ensured that they comprehended ChatGPT's capabilities. This enabled players to engage in a realistic dialogue with ChatGPT and tailor responses according to personal preference.

2 Results

Primary hypotheses, secondary analyses and expectations are laid out as preregistered via the AEA RCT Registry (Greevink, 2023). The accompanying pre-analysis plan can be found in Appendix Section E (page 37). For this study, we recruited 320 participants who were equally divided over the two treatments. Throughout the section, we start with the null hypotheses; in the subsequent text, we specify whether we had a directional prediction for the alternative hypotheses, and follow up with the findings.

2.1 Main Hypotheses

Results for the three main hypotheses are reported in Table 1.

Table 1: OLS results for main hypotheses

	(1) *B's Roll rate*	(2) *A's In rate*	(3) *B's Promise rate*	(4) *B's Follow-Up rate*	(5) *A's Trust rate*
GPT Treatment	-3.8	1.2	18.8^{**}	-21.7^{**}	-7.2
	(7.7)	(7.6)	(7.3)	(8.4)	(7.9)
Constant	66.3	65.0	58.8	83.0	83.0
	(5.4)	(5.4)	(5.6)	(5.6)	(5.6)
Observations	160	160	160	109	109
Includes	B players	A players	B players	B sent a promise	A received a promise

Notes: Estimates are based on a linear probability model and are given in percentage points. Column 4 regresses B's ROLL rate after making a promise and Column 5 regresses A's IN rate after receiving a promise. We use 'HC3' robust standard errors. All p-values reported in this paper result from two-sided tests. Significance levels are indicated as: *, ** and *** for $p < 0.10$, $p < 0.05$ and $p < 0.01$ respectively.

No overall Effect of AI-mediated communication on trust and trustworthiness. Our first hypothesis concerns differences in the levels of trust and trustworthiness.

H 1 *There is no difference in trust (fraction of participants playing IN) and trustworthiness (fraction of participants playing ROLL) between the Communication treatment and the GPT treatment.*

We did not have strong directional predictions due to the likely presence of competing forces. The effect on ROLL play depends on the number of promises, and on whether B-players continue to feel accountable for promises suggested by ChatGPT. The effect on IN play relies on whether ChatGPT access affects the number of promises made and on A-players finding the promises credible, even if they suspect them to be generated by ChatGPT. IN play could also be affected by the value A-players place on the quality of messages, which could be improved by ChatGPT. The OLS regression, as indicated in columns (1) and (2) of Table 1, indeed shows no significant treatment effect. Overall, we observe 65% of IN and 66% ROLL in our baseline Communication treatment and we do not find rates to be significantly different in our GPT treatment. These rates are similar to what was found in communication treatments of previous experiments (Charness and Dufwenberg, 2006; Ederer and Schneider, 2022).

Promises are made more often but kept less frequently with AI-assisted communication. The second and third hypotheses are about the effect of GPT on trustworthiness via the specific channel of promises. Hypothesis 2 is about the frequency of promises.

H 2 *There is no difference in the decision whether or not to include a promise in the message between the Communication treatment and the GPT treatment.*

We expected participants in the GPT treatment to make more frequent promises since ChatGPT is very prone to suggest messages containing an explicit promise or clear statement of intent.[1] To determine whether a message included a promise, we used three independent coders, who were unaware of our hypotheses and of the treatments, and a majority rule to resolve disagreement. A message was coded as a promise if it featured an explicit promise or a statement of intent. Our results align with expectations: Table 1 column (3) shows a significant and sizeable treatment effect: B-players in the GPT treatment are 18.8 percentage points more likely to send a promise compared to the Communication treatment ($p = 0.011$).

[1] Even when participants ask ChatGPT to make messages much shorter, or to change the tone of the message substantially, the language model rarely omits the promise or statement of intent.

Hypothesis 3 is about the level of commitment to promises.

H 3 *There is no difference in the decision to keep a promise between the Communication and GPT treatments.*

We expected that ChatGPT-suggested promises lead to lower commitment than self-written promises. We focus on the sub-sample where a promise was made, and report results in column (4) of Table 1. As anticipated, participants in the GPT treatment are 21.7 percentage points less likely to keep their promises ($p = 0.011$).

Aside from documenting treatment differences, we explore if participants who select to access ChatGPT are less likely to commit to promises compared to those who did not. For this, we preregistered another regression analysis using a dummy variable indicating whether participants 'Accessed GPT'. All participants who accessed the OpenAI website were classified as having Accessed GPT, regardless of how much they copied or edited a GPT-suggested prompt. Table A1 of the supporting material shows that B-players who visited the chatGPT website were 25.7% less likely to keep their promises compared to those who did not ($p = 0.004$).

Interestingly, this difference in promise keeping behavior does not seem to be anticipated by A-players. Even though A-players tend to trust promises less in the GPT treatment, the relevant coefficient in column 5 of Table 1 is not significant.

2.2 Secondary Hypotheses

Our pre-registration includes a series of secondary hypotheses, here labeled as S. We report on some of them in the paper, and relegate the rest to the supplementary material (Appendix section B, page 24).

Coordination on the efficient outcome conditional on promises is lower with AI-assisted communication.

S 1 *The joint outcome (IN, ROLL) occurs equally often in the Communication and the GPT treatment.*

We expected the degree of coordination on the joint efficient outcome (IN, ROLL) to be lower in the GPT treatment. Faced with a new technology, individuals may still not have converged on how to interpret and what to expect from AI-mediated communication. Moreover, ChatGPT may make it easier for cheaters to formulate messages that are equally convincing as those of honest people. We find no significant overall effect, but we do see a notable difference when focusing on promises. Following a promise, the joint outcome (IN, ROLL) occurred 68% of the time in the Communication treatment, and only 46.8% of the time in the GPT treatment ($p = 0.043$, see Table 2). Coordination on the efficient outcome (normalizing for baseline levels of trust) is also lower in the GPT communication treatment ($p = 0.067$, coefficients in Table A2).

Table 2: Proportion Tests - Coordination

		Communication		***GPT***		Δ	
Full Sample		*Roll*	*DR*	*Roll*	*DR*	*Roll*	*DR*
	In	**(37/80)**	(15/80)	**(33/80)**	(20/80)		
		46.3%	18.8%	**41.3%**	25.0%	**−5.0%**	6.2%
	Out	(16/80)	(12/80)	(17/80)	(10/80)		
		20.0%	15.0%	21.3%	12.5%	1.3%	−2.5%
Only Promises		*Roll*	*DR*	*Roll*	*DR*	*Roll*	*DR*
	In	**(32/47)**	(7/47)	**(29/62)**	(18/62)		
		68.1%	14.9%	**46.8%**	29.0%	**−21.3%****	14.1%
	Out	(7/47)	(1/47)	(9/62)	(6/62)		
		14.9%	2.0%	14.5%	9.7%	−0.4%	7.7%

Notes: 2-sample test for equality of proportions with continuity correction. DR represents DON'T ROLL. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Classification of messages based on similarity score

Some players who access ChatGPT eventually decide to write a message of their own or a heavily edited version of ChatGPT suggestions. Most of the analysis that follows is based on the degree of similarity between Chat-

GPT suggestions and the message that is eventually sent. We give here an overview of how this similarity is calculated, with more details in Appendix C, page 30. To evaluate how closely messages mirror ChatGPT suggestions, we calculate a similarity score on a scale from 0 to 100 (using the cosine similarity method as in Huang 2008). In case ChatGPT suggested multiple messages (e.g., after multiple iterative prompts), we pick the highest similarity score among these multiple suggestions. We also calculate a similarity score for people who did not access the ChatGPT website.[2] Figure 1 reports the similarity score in two columns for participants who did and did not access ChatGPT. We can notice a baseline degree of similarity that occurs naturally even for players who did not access ChatGPT. Participants who did access ChatGPT exhibit a large variation in their similarity scores. We classify them in two groups based on a natural cutoff (i.e., the highest similarity score of participants who did not access ChatGPT). 43 participants are above the cutoff: Their message was more similar to the AI suggestions than any message from the group of people who did not access ChatGPT. We infer that these participants copied or partially edited the suggested message. 15 participants, despite having accessed ChatGPT, heavily edited or rewrote the prompt: their similarity score is indistinguishable from those who did not access the website at all.

Beliefs about ChatGPT usage are predictive of actual usage and are not associated with distrust.

S 2 *The belief that the message was generated by ChatGPT does not predict whether ChatGPT was actually used.*

We expected participants to partially detect the use of ChatGPT. Indeed, people can predict whether ChatGPT was used better than a random guess. Specifically, a 10% increase in the belief that a message is GPT-generated is associated with a 3.7% higher likelihood of the sender having accessed the

[2] The similarity score for these participants uses as benchmark the initial message suggestion that ChatGPT generates based on the instructions of the game. This message is generated for all participants, but is never seen by the participants who decide not to access the ChatGPT website.

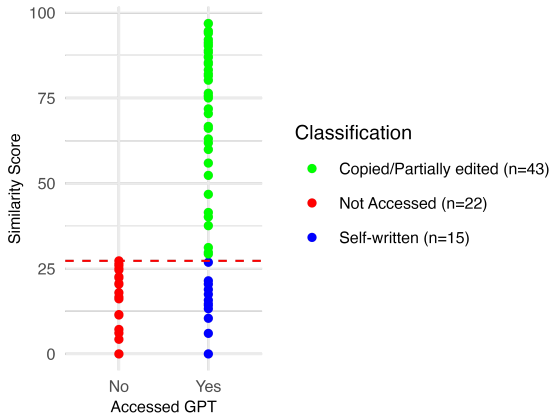


Figure 1: Similarity cutoff

Notes: "Not Accessed" are participants who did not click on the link to enter the chatGPT website. Among those who accessed the website, those with a similarity score above the maximum similarity score of the "Not accessed" group are classified as "Copied/Partially edited". Those below the cutoff are classified as (likely to) have sent a "Self-written" message.

chatGPT website ($p = 0.003$, OLS, Table A3, column 1), and a 5.6% higher likelihood that the message was copied or partially edited from a ChatGPT prompt ($p < 0.001$, Table A3, column 2).

S 3 *The belief that a message was generated by ChatGPT does not affect the level of trust displayed by A players.*

Contrary to our expectations, participants do not show less trust in messages they believe to be more likely written by ChatGPT. We even find a positive overall relationship between the belief in ChatGPT use and trust. This finding, however, is mediated by the fact that beliefs in ChatGPT use are positively associated with a higher likelihood of promises, which itself has a strong link to trust. To control for this effect, in a second specification, we only include messages that contain promises. Among these messages, we find no significant relationship between the belief that ChatGPT was used and trust ($p = 0.68$, Table A3, column 4).

The more the message aligns with ChatGPT suggestions (i.e., the fewer the edits), the lower the trustworthiness of a promise.

S 4 *The rate of promise-keeping is not affected by player B's similarity score.*

We expected that participants who merely copy the AI suggestions would be less likely to keep promises compared to those who choose to heavily edit or rewrite these suggestions before sending them. In other words, we expected a negative monotonic relationship between similarity score and trustworthiness of the promise. We indeed find evidence of this monotonicity, but only for the 43 participants above the similarity score cutoff. For them, a decrease of 10% in the similarity score corresponds to a significant 12% increase in the likelihood of promise-keeping (OLS regression, $p < 0.001$, Table A4 Column 1). Figure 2 shows how promise-keeping depends on the similarity score. The solid line excludes participants below the cutoff (who appear to have ignored ChatGPT suggestions and written their own message). The dotted line includes them. Interestingly, the monotonic relationship is inverted for participants below the cutoff although this is based on few datapoints.

Promises are no longer a reliable cue of honesty in AI-assisted communication.

ChatGPT recognizes that promises are instrumental in creating trust. Suggestions by GPT often include a promise, which tend to be copied by participants who use GPT, and in particular by those who intend to cheat. Table A5 illustrates that among those B-players who DON'T ROLL, the rate at which promises are included increases starkly from 29.6% in the Communication treatment to 80.0% in the GPT treatment. This 80.0% is even higher than the corresponding figure for participants in the GPT treatment who choose to ROLL (76%). As a result, the presence of a promise becomes a completely irrelevant cue to identify honest participants in the GPT treatment.

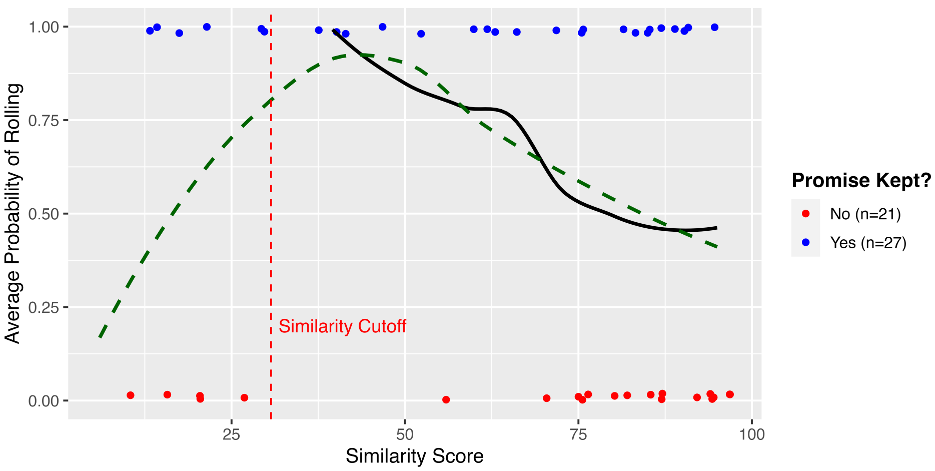


Figure 2: Similarity and Promise Keeping

Notes: Participants' decisions to ROLL and DON'T ROLL for each similarity score. Only incorporates B players that made a promise. Observations are randomly scattered to enhance readability. Two distinct smoothed moving average lines are presented: the dashed green line includes all B players that accessed the chatGPT website, while the solid black line only includes B players above the similarity cutoff.

3 Discussion

Our paper contributes to the literature that studies how communication influences trust and trustworthiness in the absence of access to AI-powered language models.[3] [4]

[3] Free-form chat communication substantially increases trust and trustworthiness in a standard trust game (Charness and Dufwenberg (2006), Goeree and Zhang 2014, Ismayilov and Potters 2016, Ederer and Schneider 2022). These studies also find that trustors who received a promise are much more inclined to trust, while trustees who made a promise are much more likely to follow up with the trustworthy choice. In Charness and Dufwenberg (2006), pairs where a promise was sent played (IN, ROLL) 67% of the time, while this rate equaled only 27% in the absence of a promise. One study (Deck et al. (2013)) does not replicate these findings, which may be due to the unusual high levels of trust and trustworthiness in their control treatment without communication.

[4] A separate strand of the literature studies the effects of restricted communication in which trustees choose between bare promises and blank messages. Bare promises are precoded messages in which the trustees send a message stating "I promise to choose *ROLL*". Overall, the picture that emerges from this literature is that bare promises are less effective at eliciting trust and trustworthiness Charness and Dufwenberg (2010), Di Bartolomeo et al. (2019), Chen and Zhang (2021). Trustors consider bare promises less credible,

Our paper also contributes to a recent and thriving literature that studies how behavior and norms evolve in a new landscape where humans interact with bots, for instance in environments where humans and bots can help or hinder each other (Makovi et al., 2023), in environments where bots spread content of low credibility (Shao et al., 2018) and in environments where bots increase exposure to negative and inflammatory content (Stella et al., 2018). Another related literature studies the corrupting power of artificial intelligence on human behavior (Köbis et al., 2021; Leib et al., 2021). AI-mediated communication may also raise fears about abuse by others; Purcell et al. (2023) find that people tend to overestimate others' use of AI-mediated communication and that they think that others use it irresponsibly. At the same time, there is a potential for algorithms to help detect deception in communication in high-stakes strategic interactions (Serra-Garcia and Gneezy, 2023).

We find it striking that a mere few months after its introduction, AI assisted communication started to corrode the meaning of written promises. AI may make it easier for cheaters to mimic trustworthy people making it harder to distinguish between them. While we still do not detect an erosion of trust in promises made with the mediation of AI, the lower degree of coordination on the efficient outcome suggests that the unwarranted level of trust in ChatGPT promises may soon decrease as well.

As a result, we expect that humans may search for other ways to meaningfully communicate their intentions in interactions in which trust is involved. A possible consequence may be the return to face-to-face meetings[5] or, on the other side of the spectrum, the development of entirely novel

while trustees feel less committed to follow up on their promises. A contemporaneous paper extends this work by studying how AI crafted messages affect people's trust and trustworthiness (Bogliacino et al. (2023). Instead of choosing a bare promise, participants in their study choose between writing their own message and using a message from a set of AI messages that were created by the authors by feeding some of the messages from Charness and Dufwenberg (2006) in GPT as a prompt. In contrast, we allow participants to freely interact with GPT until they are satisfied with their message.

[5] Although people are unable to detect the trustworthiness of strangers based on their facial appearance (Jaeger et al., 2022), face-to-face communication helps people to predict who will actually be cooperative in social dilemmas (Frank et al., 1993).

forms to communicate and establish trust.

4 Methods

4.1 Experimental Procedures

Sessions for the computerized lab experiment took place at the CREED laboratory at the University of Amsterdam between June and October 2023. We preregistered the experiment at the AEA RCT Registry (Greevink, 2023). The two treatments were carried out in parallel. Sessions lasted around an hour, with GPT sessions taking slightly longer than Communication sessions. Average earnings were around €18 per participant, including a €6 participation fee. The full instructions for the experiment are provided in Appendix Section G (page 41).

After participants read the instructions and successfully passed a comprehension test, Player B was prompted to send a message (with or without access to ChatGPT depending on treatment), while Player A was presented with a waiting screen. After all B players had sent their messages, Players A could read the message sent by their partner, after which both Players A and B simultaneously made their decisions on whether to play In/Out or Roll/Don't Roll. Subsequently, we asked participants to predict their partner's choice. The accuracy of this prediction could earn them extra money, based on a binarised scoring rule. We made it explicit that accurate guesses would earn them more money, and that specific details about the payoff mechanism were available upon request at the end of the experiment, similar to the no-information treatment in Danz et al. (2022). Additionally, we asked Player A to predict whether Player B had used ChatGPT to craft their message. This prediction was not financially incentivized.[6]

Given the relatively recent emergence of ChatGPT, additional instructions were provided to the B players in the GPT group to help them understand its use. These guidelines included a few example prompts. These

[6] Charness et al. (2021) recently showed that non-incentivized belief measurements can be as effective as more complex incentivized ones, particularly in a short, straightforward task like ours.

prompts were asked by previous participants in a small pilot version of the experiment.[7] Furthermore, we pre-entered a summary of the game's instructions into ChatGPT, enabling it to function as the initial prompt. Without adding any questions, ChatGPT automatically utilized these instructions to generate an initial message suggestion. This suggestion served as a starting point for B players to start their conversation with ChatGPT. The extra instructions given to B players and the original prompt given to ChatGPT can be found in Appendix Sections G.2 (page 48) and F (page 40), respectively.

At the end of the first part, participants got instructions for the second part which was identical to the first except they would now swap roles and be paired with a new random partner. Participants were only informed about the identical repetition at this stage so that the results of the first round can be interpreted in isolation as a one-shot game. The repetition of the game in the second part was an attempt to increase the amount of data available (and consequently power) subject to finding no evidence of round fixed-effect. Because we did find differences across the two rounds, we decided to focus only on the results from the first round

After completing the second part of the experiment, participants were asked to answer a series of survey questions. These included an incentivized slider measure of Social Value Orientation (Murphy et al., 2011), introspective questions about their risk aversion, experience with ChatGPT, attitudes towards AI/ChatGPT, and several general demographic questions. The full survey can be found in Appendix Section G.3 (page 49). Following the completion of the survey, participants were provided with payment feedback. When everyone had received and read their payment feedback, they were called up one at a time to be paid privately in cash, outside of the room.

4.2 Sample composition and compensation

We recruited 320 participants from the CREED participants base. Participants were equally divided over the two treatments. Table A6 reports

[7] This pilot was just aimed at generating ChatGPT prompts. Participants did not play the game, but received the game instructions and were asked what message they would send if they were Player B. Participants were compensated with a flat fee of 10 euros.

demographic statistics, and shows that the randomization across treatments was successful.

4.3 Ethical approval

This study was approved by the UvA IRB (EB-2989). We did not make use of deception in our experiment.

4.4 Pre-registration plan

We pre-registered our study in the AEA RCT Registry before we ran the experiments (https://doi.org/10.1257/rct.11511-1.0). In our pre-analysis plan, we stated that for Hypotheses 1, 2 and 3 we would first determine whether we can pull the data from the two rounds, or whether we should analyze the data separately because of round differences. To do this, we ran OLS regressions with a GPT dummy, a round fixed effect, and their interaction term. If neither the round fixed effect nor the interaction term is significant, we preregistered that we would pool the data for the two rounds and conduct statistical analysis on the pooled dataset. Conversely, if the round fixed effect and/or the interaction term prove significant, we preregistered that we would rely on the OLS regression results instead, and separate results by round.

Because we do find evidence of round-fixed effects in the regressions for Hypotheses 2 and 3, we present regression results that separate the two rounds. In the main text, we decided to focus on round-one data for all hypotheses as it is easier to interpret. We present the round 2 results in Appendix section D (page 32). Because the second round was unannounced, the first round can be treated as a one-shot game and thus offer a clean interpretation of the data. The second-round data, on the other hand, albeit built on role reversal, may be showing the first glimpses of dynamic effects. However, we do not have enough repetitions to say anything conclusive about dynamic effects.

We deviate from our pre-analysis plan in our presentation of the results for Hypothesis 1. To streamline the presentation of the results, we chose to focus on round 1 also for this hypothesis. The conclusion for Hypothesis 1 does not change if we would have followed the pre-analysis plan and would have combined round 1 and round 2 data (see Table A7). We did not preregister the analysis in Table A5 that illustrates the extent to which B-players include a promise in their choice conditional of their choice to ROLL or DON'T ROLL (this analysis is discussed at the end of the Results Section). We also did not preregister the analysis in Table A2 that shows the coordination on the efficient outcome conditional on IN being played (this analysis is discussed at the end of the first supplementary hypothesis S1).

References

Bogliacino, F., Buonanno, P., Fallucchi, F., and Puca, M. (2023). Trust in times of ai. Technical report, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2010). Bare promises: An experiment. *Economics letters*, 107(2):281–283.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Chen, Y. and Zhang, Y. (2021). Do elicited promises affect people's trust?—observations in the trust game experiment. *Journal of Behavioral and Experimental Economics*, 93:101726.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–83.

Deck, C., Servátka, M., and Tucker, S. (2013). An examination of the effect of messages on cooperation under double-blind and single-blind payoff procedures. *Experimental Economics*, 16:597–607.

Dennean, K. (2023). Let's chat about chatgpt. Technical report, UBS.

Di Bartolomeo, G., Dufwenberg, M., and Papa, S. (2019). The sound of silence: A license to be selfish. *Economics Letters*, 182:68–70.

Ederer, F. and Schneider, F. (2022). Trust and promises over time. *American Economic Journal: Microeconomics*, 14(3):304–20.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Frank, R. H., Gilovich, T., and Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and sociobiology*, 14(4):247–256.

Goeree, J. K. and Zhang, J. (2014). Communication & competition. *Experimental Economics*, 17(3):421–438.

Greevink, I. (2023). Ai-powered promises: The influence of chatgpt on trust and trustworthiness. Technical report, AEA RCT Registry, https://doi.org/10.1257/rct.11511-1.0.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., and Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(5487).

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student* *conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19:382–393.

Jaeger, B., Oud, B., Williams, T., Krumhuber, E. G., Fehr, E., and Engelmann, J. B. (2022). Can people detect the trustworthiness of strangers based on their facial appearance? *Evolution and Human Behavior*, 43(4):296–303.

Jakesch, M., Hancock, J. T., and Naaman, M. (2023). Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6):679–685.

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., and Irlenbusch, B. (2021). The corruptive force of ai-generated advice. *arXiv preprint arXiv:2102.07536*.

Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., and Rahwan, T. (2023). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications*, 14(1):3108.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Purcell, Z. A., Dong, M., Nussberger, A.-M., Köbis, N., and Jakesch, M. (2023). Fears about ai-mediated communication are grounded in different expectations for one's own versus others' use. *arXiv preprint arXiv:2305.01670*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J., Christakis, N., Couzin, I., Jackson, M., Jennings, N., Kamar, E., Kloumann, I., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D., Pentland, A., Roberts, M., Shariff, A., Tenenbaum, J., and Wellman, M. (2019). Machine behaviour. *Nature*, 568:477–486.

Serra-Garcia, M. and Gneezy, U. (2023). Improving human deception detection using algorithmic feedback.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Stella, M., Ferrara, E., and De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, Julia, a. K. S., Kamar, E., Kraus, S., Leyton-Brown, Kevin, P. D., Press, William, S. A., Shah, J., Tambe, M., and Teller, A. (2016). Artificial intelligence and life in 2030. in: One hundred year study on artificial intelligence: Report of the 2015–2016 study panel. *Tech. Rep.*, Stanford University.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52.

Appendix

A Tables mentioned in the main text

Table A1: Promise Keeping - Accessed chatGPT

	(1) *B's Roll Rate*
Accessed chatGPT	-25.7^{***} (8.9)
Constant	82.0 (5.0)
Observations	160
Includes	B players

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' robust standard errors. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A2: Coordination on efficient outcome conditional on IN play

	Communication	*GPT*	Δ
Full Sample:			
Roll rate conditional	(37/52)	(33/53)	
on IN play	71.2%	62.3%	−8.9%
Only Promises:			
Roll rate conditional	(32/39)	(29/47)	
on IN play	82.1%	61.7%	**−20.3%***

Notes: 2-sample test for equality of proportions with continuity correction. The displayed percentages indicate how often player B plays Roll after player A plays IN. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A3: GPT beliefs

	(1)	(2)	(3)	(4)
	B Accessed GPT website	*B Copied/ Rewritten*	*A's In rate*	*A's In rate*
GPT belief (0-100)	0.37***	0.56***	0.34**	0.1
	(0.12)	(0.12)	(0.13)	(0.1)
Constant	54.36	23.45	49.51	72.3
	(9.05)	(7.81)	(9.18)	(10.5)
Observations	58	43	80	62
Includes	A players	A players	A players	A received a promise

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' robust standard errors. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A4: Rewriting

	(1) *B's Roll Rate*
Similarity Score	-1.2^{***} (0.3)
Constant	147.4 (24.4)
Observations	37
Includes	B copied/rewritten

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' robust standard errors. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A5: The Power of Promises

	Communication	*GPT*	Δ
Promise rate among *B-players that ROLL*	(39/53) 73.6%	(38/50) 76%	2.4%
Promise rate among *B-players that DON'T ROLL*	(8/27) 29.6%	(24/30) 80%	**50.4%*****
Δ	$-44\%^{***}$	**4%**	
IN rate A players *that Receive Promises*	(39/47) 83.0%	(47/62) 75.8%	−7.2%
IN rate A players *that Receive No Promises*	(13/33) 39.4%	(6/18) 33.3%	−6.1%
Δ	$-53.6\%^{***}$	$-42.5\%^{***}$	

Notes: 2-sample test for equality of proportions with continuity correction. The top two rows test the difference between and within treatments in the fraction of B players that send promises for those that ROLL and DON'T ROLL. The bottom two rows test the difference between and within treatments in the fraction of A players that play IN for those that receive a promise and those that do not. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A6: Treatment Balance

	Communication	GPT	*P value*
n	160	160	
test fails (sd)	0.579 (0.787)	0.579 (0.720)	1
SVO angle (sd)	25.1 (14.6)	23.3 (14.0)	0.273
risk (sd)	5.88 (2.03)	5.98 (2.02)	0.659
age (sd)	21.6 (4.81)	21.3 (2.81)	0.515
gender (%)			0.972
Female	88 (55.0)	89 (55.6)	
Male	68 (42.5)	66 (41.2)	
Other	2 (1.25)	3 (1.88)	
degree (%)			0.488
Bachelor	120 (75.0)	128 (80.0)	
Master	39 (24.4)	30 (18.8)	
faculty (%)			0.326
Business	32 (20.0)	32 (20.0)	
Economics	73 (45.6)	72 (45.0)	
Humanities	5 (3.13)	7 (4.38)	
Interdisciplinary	17 (10.6)	9 (5.63)	
Law	5 (3.13)	5 (3.13)	
Other	3 (1.88)	2 (1.25)	
Science	7 (4.38)	3 (1.88)	
Social and Behavioral sciences	17 (10.6)	30 (18.8)	

Notes: For continuous variables, the mean and standard deviation (in parentheses) are reported. For categorical variables, the frequency and proportion (in parentheses) are reported. P-values are calculated using t-tests for the continuous variables and chi-squared tests for the categorical variables.

A data collection issue resulted in the non-registration of comprehension 'test fails' during the initial two sessions. These sessions included a GPT session (n=20) and a communication session (n=20). Therefore these observations are excluded for the 'test fails' row only.

Table A7: Proportion tests Hypothesis 1

	Treatment:				
	Communication	GPT	*Z Stat*	*P value*	Includes
B's Roll Rate	(101/160) 63.1%	(88/160) 55.0%	1.364	0.173	B players
A's In Rate	(105/160) 65.6%	(100/160) 62.5%	0.466	0.641	A players

Notes: Two-sample tests for equality of proportions were performed with continuity correction, using pooled data from both rounds. Reported p-values correspond to two-sided tests.

B Preregistered Secondary Analysis

B.1 First Order beliefs

S 5 *First-order beliefs about trust (fraction of IN play) and trustworthiness (fraction of ROLL play) do not differ between the Communication and GPT treatment.*

Overall beliefs about ROLL and IN play show no significant differences between treatments. Table A8 indicates that B players, on average, predicted their matched A player would play IN with a 57.0% probability, while A players anticipated that their matched B player would play Roll 55.6% of the time. These predictions vary by only one or two percentage points across treatments, hence, we do not reject the null hypothesis that first order beliefs vary significantly between the two treatments.

Notably, we also observe no substantial difference in IN beliefs among B players who sent a promise. As per column (3) of table A8, their belief in A playing IN was only 3% lower than B players in the Communication treatment, despite following up on their promises 22% less often (table 1, column 4).

Table A8: First order beliefs

	(1) *B's In belief*	(2) *A's Roll belief*	(3) *B's In belief*	(4) *A's Roll belief*
GPT Treatment	-1.45	-2.30	-2.73	-5.14
	(3.77)	(4.75)	(4.21)	(5.28)
Constant	57.04	55.55	61.72	63.26
	(2.87)	(3.44)	(3.26)	(3.80)
Observations	160	160	109	109
Includes	B players	A players	B received a promise	A received a promise

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' robust standard errors. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

B.2 Individual characteristics

In this section, we investigate the influence of social value orientation (SVO), risk aversion, and attitudes toward AI on trust and trustworthiness. We further explore whether there is a potential differential impact of SVO on promise-keeping behavior among those that access ChatGPT, and whether there is a relationship between trust and trustworthiness between rounds.

Methods

SVO scores were calculated using the slider measure from Murphy et al. (2011). This approach asks participants to make a set of six decisions, whereby participants distribute payoffs between themselves and another randomly assigned participant. The SVO scores take the form of an angle, derived through the formula presented below, where $\bar{A}_o$ represents the mean allocation provided to the other party, and $\bar{A}_s$ symbolizes the mean allocation that is self-received. The distribution of SVO angles among participants can be found in figure A1 below. We expected that 'cooperators' ('prosocials' in Murphy et al.), those with a comparatively high SVO angle, will keep their promises irrespective of ChatGPT exposure; conversely,

’individualists’, identified by a relatively low SVO angle, were expected to keep their promises less frequently when utilizing ChatGPT.

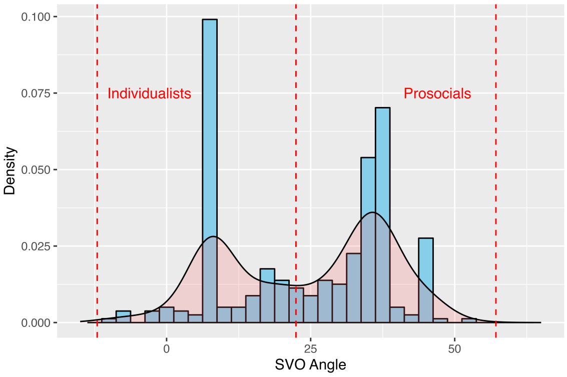


Figure A1: Histogram - SVO

Notes: Vertical dashed lines demarcate the SVO categories: Individualists with SVO angles between -12.04 *and* 22.45*, and Prosocials with angles between* 22.45 *and* 57.15 *(Murphy et al., 2011). The black line indicates a smoothed density estimation.*

$$\text{SVO}^\circ = \arctan\left(\frac{(\bar{A}_o - 50)}{(\bar{A}_s - 50)}\right)$$

For the assessment of risk aversion, we took an introspective approach, asking participants to evaluate their propensity to take risks on a scale from 0 to 10. This question was based on the introspective measure used in the Global Preference Survey (GPS) (Falk et al., 2018). Participants’ attitudes towards AI were gauged using a similar question which asked them to score, from 0-10, their sentiment towards ”GPT and other AI,” ranging from strong distrust/dislike (0) to strong fondness/trust (10).

Results

Table A9 below shows the results for secondary hypotheses S6 to S10.

S 6 *There is no significant relationship between Social Value Orientation (SVO) and trust and trustworthiness.*

Table A9 rows (1) and (2) show significant positive correlations between Social Value Orientation (SVO) and both trust and trustworthiness (r =

Table A9: Correlation tests for S6 to S10

	Variable Pairs:		*Correlation*	*P value*	*N*
	Variable 1	Variable 2	Cor	P	N
(1) SVO and Trust	SVO angle	In	0.18	0.023	160
(2) SVO and Trustworthiness	SVO angle	Roll	0.29	<0.001	160
(3) Using GPT and Promise Keeping; prosocials	Accessed GPT	Roll	−0.15	0.369	36
(4) Using GPT and Promise Keeping; individualists	Accessed GPT	Roll	−0.27	0.182	26
(5) Risk and Trust	Risk Seeking (0 to 10)	In	0.33	<0.001	160
(6) Risk and Trustworthiness	Risk Seeking (0 to 10)	Roll	0.11	0.171	160
(7) AI Attitude and Trust	GPT Fondness (0 to 10)	In	−0.13	0.235	80
(8) AI Attitude-Roll Trustworthiness	GPT Fondness (0 to 10)	Roll	−0.05	0.678	80
(9) Behavior Across Rounds	Roll	In	0.29	<0.001	320

Notes: All tests use round 1 data only, except (9), which requires the use of second-round data as well. Pearson's product-moment correlation is used. P-values are for two-sided tests, the null hypothesis is that the true correlation coefficient is zero.

0.18, $p = 0.023$ and $r = 0.29$, $p < 0.001$ respectively). Therefore we can reject the null hypothesis that there is no such relationship. The percentage magnitude of this relationship can also be quantified in percentage terms: the ROLL rate is 37.7% among participants classified as individualists and 71.1% among those labeled as prosocial.

S 7 *There is no significant difference in the frequency of promise-keeping between those who access ChatGPT and those who do not, regardless of whether they are individualists or cooperators.*

We expected that cooperators would maintain their promises regardless of whether they access ChatGPT website, and that individualists who access ChatGPT would be less likely to keep their promises compared to those who did not visit. However, the results displayed in Table A9 in rows (3) and (4) do not support this expectation to a significant extent. Namely, there is no significant correlation between accessing ChatGPT and promise-keeping behavior among both prosocial and individualist subjects ($r = -0.15$, $p = 0.369$ and $r = -0.27$, $p = 0.182$ respectively). Therefore, we do not have sufficient evidence to reject null hypothesis S7.

S 8 *There is no significant relationship between risk-seeking and trust and trustworthiness.*

Table A9 row (5) shows a significant positive relationship between self-reported risk-seeking and Trust (IN) ($r = 0.33$, $p < 0.001$). Hence, we reject the null hypothesis that there is no significant relationship between risk-seeking and trust. Row (6) shows that the relationship between risk-seeking and Trustworthiness is not significant ($r = 0.11$, $p = 0.171$). Hence, we do not reject the null hypothesis that there is no significant relationship between risk-seeking and trustworthiness.

S 9 *There is no significant relationship between AI attitude and trust and trustworthiness in the GPT treatment.*

Although we had expected that those who have a positive attitude towards GPT would be more trusting and trustworthy in the GPT treatment, Table A9 rows (7) and (8) show that this is not the case. Specifically, we find no significant relationship between GPT fondness and trust and trustworthiness ($r = -0.13$, $p = 0.235$ and $r = -0.05$, $p = 0.678$ respectively), and therefore do not reject null hypothesis S9.

S 10 *There is no significant relationship in behavior across the two rounds: trustworthy participants are not more trusting and vice versa.*

Table A9 row (9) shows a significant positive relationship between playing Roll in one of the two rounds and In in the other, and vice versa ($r = -0.29$, $p < 0.001$). Hence we reject the null hypothesis that there is no such relationship.

B.3 Robustness Check

Table A10: Robustness check main hypotheses

	(1)	(2)	(3)	(4)	(5)
	B's Roll rate	*A's In rate*	*B's Promise rate*	*B's Follow-Up rate*	*A's Trust rate*
GPT treatment	−0.5	−2.5	19.6***	−19.1**	−10.4
	(7.4)	(7.6)	(7.4)	(8.8)	(7.4)
Controls	✓	✓	✓	✓	✓
Constant	62.7	68.4	118.4	61.8	68.1
	(19.2)	(29.5)	(22.3)	(21.0)	(46.0)
Observations	160	160	160	109	109
Includes	B players	A players	B players	B sent a promise	A received a promise

Notes: Controls include SVO, Risk Seeking, Economics and Business faculty, European, Age, and Gender. Estimates are based on a linear probability model and are given in percentage points. Column 4 regresses B's ROLL rate after making a promise and Column 5 regresses A's IN rate after receiving a promise. We use 'HC0' robust standard errors to deal with the increased amount of regressors. All p-values reported in this paper result from two-sided tests. Significance levels are indicated as: *, ** and *** for $p < 0.10$, $p < 0.05$ and $p < 0.01$ respectively.

As shown in Table A10, when we add in various control variables we do not find that our main results differ significantly from Table 1 in the main text.

C Similarity Score calculation

To quantify the extent to which players have copied or rewritten a chatGPT suggested message we calculate a (Cosine) similarity metric. Similarity metrics are often used in various machine learning tasks, including text analysis (Hastie et al., 2009). We use a similar approach as Huang (2008) who shows that cosine similarity works well compared to other similarity metrics in a text clustering task. An important advantage of cosine similarity in our context is that it allows for the comparison of pieces of text independent of their respective sizes. This way high similarity scores are possible when only part of the text is copied from ChatGPT suggestions.

In this method, messages and chatGPT responses are transformed into mathematical vector representations that show which words the messages contain and at what frequency. A bag of words (BoW) representation is used wherein pieces of text are cut up into individual unique terms: all grammar, word order, and frequently occurring unsubstantial words such as stop words or conjunctions are taken out (Zhang et al., 2010). More specifically, for each message-response pair we define a set of l_i terms $T_i = \{t_1, \ldots, t_{l_i}\}$ for each player $i \in [1, 160]$. This is a list of all unique words that occur in the player's sent message, or one of the k_i ChatGPT suggestions, or in both. The message vectors m_i and response vectors $g_i^{k_i}$ are presented in the formulas below and consist of a list that shows the frequency $f(t)$ of each term l_i:

$$g_i^{k_i} = (f(t_1), \ldots, f(t_l))$$

$$m_i = (f(t_1), \ldots, f(t_l))$$

Consequently, we end up with a set of messages $M = \{m_1, \ldots m_i\}$, one for each player in our GPT treatment, each accompanied by a set of k_i ChatGPT responses $G_i = \{g_i^1, \ldots, g_i^{k_i}\}$. These together form k_i message-response pairs for each subject i.

The cosine similarity score for player i is calculated by taking the cosine of the angle between the two vectors. Our final metric takes the maximum cosine similarity score out of player $i's$ k_i message-response pairs, and can be calculated according to the formula below:

$$SIM_{\cos} = \max_{k_i} \left(\frac{m_i \cdot g_i^{k_i}}{\|m_i\| \|g_i^{k_i}\|} \right) \times 100$$

Notice that the maximum cosine similarity score provides a measure of how close the participant's sent message is to the best fitting message that the participant observed in their conversation with chatGPT. The similarity score of a participant who exactly copies a message suggested by chatGPT will be 100. The similarity scores of participants who ignore chatGPT but write their own message will be rather small but positive, because some words will show up in both messages by chance.

D Second Round Tables

Table A11: Full OLS results for main hypotheses

	(1) *B's Roll rate*	(2) *A's In rate*	(3) *B's Promise rate*	(4) *B's Follow-Up rate*	(5) *A's Trust rate*
GPT treatment	−3.8	1.2	18.8^{**}	-21.7^{**}	−7.2
	(7.7)	(7.6)	(7.3)	(8.4)	(7.9)
Second Round	−6.2	1.2	18.8^{**}	-20.1^{**}	-16.9^{**}
	(7.7)	(7.6)	(7.3)	(8.4)	(8.3)
GPT x Second	−8.8	−8.8	−11.2	8.8	2.8
	(11.0)	(10.8)	(9.6)	(12.2)	(11.6)
Constant	66.2	65.0	58.8	83.0	83.0
	(5.3)	(5.4)	(5.6)	(5.6)	(5.6)
Observations	320	320	320	239	239
Includes	B players	A players	B players	B sent a promise	A received a promise

Notes: Estimates are based on a linear probability model and are given in percentage points. Column 4 regresses B's ROLL rate after making a promise and Column 5 regresses A's IN rate after receiving a promise. We use 'HC3' clustered standard errors, clusters are formed according to groups of 4 within which communication effects are contained. All p-values reported in this paper result from two-sided tests. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A12: Proportion Tests - Coordination Second Round

		Communication		***GPT***		Δ	
		Roll	*DR*	*Roll*	*DR*	*Roll*	*DR*
Full Sample	*In*	**(33/80)**	(20/80)	**(19/80)**	(28/80)		
		41.3%	25.0%	**23.8%**	35.0%	−**17.5******%**	10.0%
	Out	(15/80)	(12/80)	(19/80)	(14/80)		
		18.8%	15.0%	23.8%	17.5%	5%	2.5%
		Roll	*DR*	*Roll*	*DR*	*Roll*	*DR*
Only Promises	*In*	**(26/62)**	(15/62)	**(17/68)**	(25/68)		
		41.9%	24.2%	**25%**	36.8%	−**16.9%***	12.6%
	Out	(13/62)	(8/62)	(17/68)	(9/68)		
		21.0%	12.9%	25.0%	13.2%	4.0%	0.3%

Notes: 2-sample test for equality of proportions with continuity correction. DR represents DON'T ROLL. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A13: Full OLS results for Promise Keeping - Accessed chatGPT

	(1) *B's Roll Rate*
Accessed chatGPT	-25.7^{***} (8.9)
Second Round	-20.5^{***} (7.4)
Accessed X Second	11.1 (12.9)
Constant	82.0 (5.0)
Observations	160
Includes	B players

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' clustered standard errors, clusters are formed according to groups of 4 within which communication effects are contained. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*
A technical error occurred which led to one second round participant their 'click' to access the GPT web- site not registering. Because the participant did respond to GPT, we know the participant accessed the site. Therefore, we manually added the participant to the 'Accessed GPT' group

Table A14: Full OLS results for GPT beliefs

	(1) *B Accessed GPT website*	(2) *B Copied/ Rewritten*	(3) *A's In rate*	(4) *A's In rate*
GPT Belief	0.37^{***}	0.56^{***}	0.34^{**}	0.06
	(0.12)	(0.12)	(0.13)	(0.15)
Second Round	6.29	-8.97	17.96	-4.04
	(12.43)	(10.8)	(12.52)	(14.16)
Belief x Second	-0.17	-0.05	-0.52^{***}	-0.19
	(0.17)	(0.17)	(0.19)	(0.21)
Constant	54.36	23.45	49.51	72.34
	(9.05)	(7.81)	(9.18)	(10.48)
Observations	114	72	160	130
Includes	A players	A players	A players	A received a promise

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' clustered standard errors, clusters are formed according to groups of 4 within which communication effects are contained. Significance levels are indicated as: *, ** *and* *** *for* $p < 0.10$, $p < 0.05$ *and* $p < 0.01$ *respectively.*

Table A15: Full OLS results for Rewriting

	(1) *B's Roll Rate*
Similarity Score	-1.16^{***} (0.3)
Second Round	-82.7^{**} (36.3)
Similarity x Second	0.9^{*} (0.5)
Constant	146.6 (20.0)
Observations	72
Includes	B copied/rewritten

Notes: Estimates are based on a linear probability model and are given in percentage points. We use 'HC3' clustered standard errors, clusters are formed according to groups of 4 within which communication effects are contained. Significance levels are indicated as: *, ** and *** for $p < 0.10$, $p < 0.05$ and $p < 0.01$ respectively.
Due to a manual error during one of the sessions, we lost the chat history with Chat GPT for one participant (B) who accessed the chatGPT website in the second round. This participant is omitted from this analysis.

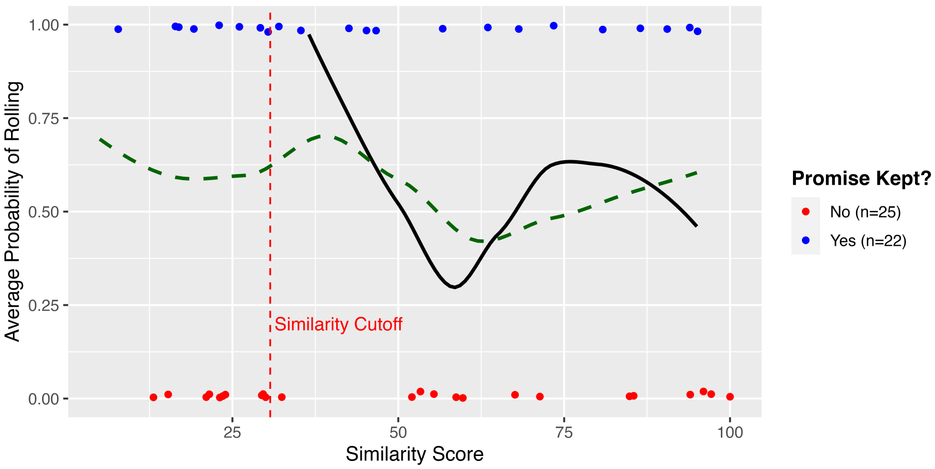


Figure A2: Similarity and Trustworthiness: Second Round

Notes: Participants' decisions to ROLL and DON'T ROLL for each similarity score. Only incorporates B players that made a promise. Observations are randomly scattered to enhance readability. Two distinct smoothed moving average lines are presented: the dashed green line includes all B players that accessed the chatGPT website, while the solid black line only includes B players above the similarity cutoff.
Due to a manual error during one of the sessions, we lost the chat history with Chat GPT for one participant (B) who accessed the chatGPT website in the second round. This participant is omitted from this analysis.

E Pre-Analysis Plan

All tests will be two-sided.

Hypothesis 1. There is no difference in trust (fraction of subjects playing "IN") and trustworthiness (fraction of subjects playing "ROLL") between the Communication treatment and GPT Communication treatment.

> *Expected result*: We do not have strong directional predictions. GPT-generated messages may read more convincing, however, agents may believe that GPT mediated communication is less sincere, or they may have a general aversion to GPT.
>
> *Test:* We will run two OLS regressions with IN (resp. ROLL) as dependent variables. As independent variables we will add a

> treatment dummy, a period dummy and their interaction terms. If the period dummy and the interaction dummy are not significant, we can pool the data of the two periods. In this case, we will perform a Z test of proportions on the pooled dataset. If the period dummy and/or the interaction dummy are significant, then we do not run the proportions test on the pooled data but focus on the regression results instead.

Hypothesis 2. There is no difference in the decision whether or not to include a promise in the message between the Communication treatment and GPT Communication treatment.

> *Expected result*: We expect that participants in the GPT Communication treatment make more promises (because many will use GPT and GPT appreciates the power of promises).
>
> *Test:* We will use independent coders to evaluate messages and determine whether they contain a promise or not. The outcome of the procedure will be coded as a binary variable called PROMISE, with zero corresponding to "no promise made" and 1 to "promise made". We will use PROMISE as dependent variable. As independent variables we will use again a treatment dummy, a period dummy and their interaction terms. We will then follow the same procedure highlighted for Hypothesis 1, i.e., a regression analysis followed by a conditional non-parametric estimation on the pooled dataset.

Hypothesis 3. There is no difference in the decision to keep a promise between the Communication treatment and GPT communication treatment.

> *Expected result*: We expect that participants in the GPT communication treatment keep their promises less often (because participants feel less committed by GPT suggested promises).
>
> *Test:* We will extract a subset of the data including all and only the messages that have been categorized as containing a

promise (as per Test of Hypothesis 2). With this subsample, we will use Roll as dependent variable. We will create a dummy which indicates whether a participant used GPT in creating the message. As independent variables we will use the GPT-used dummy, a period dummy and their interaction terms. We will then follow the same procedure highlighted for Hypothesis 1, i.e., a regression analysis followed by a conditional non-parametric estimation on the pooled dataset. To control for selection effect, we will run a second regression replacing the GPT-used dummy with a treatment dummy.

Secondary analyses

Below we list the secondary analyses that we want to carry out.

(i) We will investigate whether the joint outcome ("IN", "ROLL") occurs more often in the Communication treatment than the GPT communication treatment (we expect that this will be the case).

(ii) We will investigate if first-order beliefs about trust (fraction of subjects playing "IN") and trustworthiness (fraction of subjects playing "ROLL") differ between the Communication treatment and GPT communication treatment.

(iii) We will explore whether the belief that the message was generated by GPT is independent from the level of trust (fraction of subjects playing "IN"). Here, we expect a negative relation.

(iv) We will explore whether the belief that the message was generated by GPT is predictive of whether GPT was actually used to create the message. Here, we expect a positive relation.

(v) As a follow up analysis on Hypothesis 3, we will investigate whether the propensity to keep a promise is affected by a participant's decision to rewrite the suggestion of GPT. Our conjecture is that participants who rewrite the message may be more inclined to keep the promise (compared to those who do not rewrite it), and we expect a further correlation with the extent to which the message is rewritten.

(vi) We will investigate whether social value orientation is positively related to trust and trustworthiness, which we expect to be true. Moreover, we will investigate a potential differential effect of social value orientations to keep promises: we expect that cooperators will keep their promises to the same extent independent of whether GPT was used while individualists will keep their promises to a lesser extent when GPT was used.

(vii) We will investigate whether participants' attitudes towards artificial intelligence are related to trust and trustworthiness in the GPT treatment (we expect a positive relationship).

(viii) We will investigate whether we can identify correlations between behaviors across the two periods: are participants who are less trustworthy also less trusting?

(ix) We will investigate whether risk aversion is negatively related to trust (our conjecture) and trustworthiness (here we expect no correlation).

(x) Robustness analysis: we will run the main regressions adding a set of controls such as demographic characteristics, risk attitudes, social value orientation, and attitudes towards AI.

F Initial prompt sent to GPT

Suppose you are a student who is playing an incentived trust game in the behavioural economics lab at your university in order to earn some extra money on the side. The instructions of the game are as follows:

You have been randomly and anonymously paired with another participant. One of you will be assigned the role of Participant A, while the other will be Participant B. Your roles will be revealed before making any decisions.

You and your anonymous partner will participate in a game that rewards mutual trust.

In this game, Participant A makes the first decision. Participant A can choose to opt ”Out,” in which case the game ends and both participants earn 5. Alternatively, Participant A can choose to opt ”In,” allowing Participant B to make a decision.

If Participant A opts ”In,” Participant B can make either one of two choices. If he chooses “Don’t” he will receive 14, while player A will receive 0. If he chooses “Roll” there will be a 5/6 chance that he will receive 10, and player A will receive 12; and a 1/6 chance that he will receive 10 but player A will receive 0.

Before the game starts, participant B will have the opportunity to send a message to player A. Player B can choose to use GPT-4 in order to (help him) construct his message. Player B can edit, copy-paste or use the prompts from GPT-4 however he likes, but he can also write his own message altogether.

G (Pre-)Game Experimental Instructions

[screen 1]
[All Treatments]

Welcome!

Thank you for participating in this experiment. Please read the following instructions carefully.

Please refrain from verbally reacting to events that occur during the experiment. The use of mobile phones is not allowed.

If you have any questions or need assistance at any time, please notify the experimenter by raising your hand. The experimenter will assist you privately.

[screen 2]
[All Treatments]

General Information

This experiment consists of multiple parts. Your decisions in one part will not affect any of your choices or potential earnings in other parts. You will receive instructions for each part separately.

You will be paid a participation fee of €6.00 irrespective of your decisions. Additionally, you receive earnings that depend on your decisions and may depend on other participants' decisions.

This experiment consists of two parts. At the end of the experiment, one of these parts will be randomly chosen and used for payment. Thus, you do not accumulate payoffs between the two parts: only one part, randomly chosen at the end of the experiment, is used to calculate your payment. Your earnings will be paid to you privately at the end of the session in cash.

[screen 3]
[All Treatments]

First Decision Task (1/2)

In the first part of the experiment, one person will take on the role of player A and the other will take on the role of player B. Before the start of the experiment, you will be informed of your role. You are all randomly paired with a participant who plays the other role. No participant will ever know the identity of the player with whom they are paired.

[screen 4]
[All Treatments]

First Decision Task (2/2)

Note that during the rest of this task a summary of these instructions is available at the bottom of the screen.

Player A
Player A moves first and indicates whether they choose IN or OUT. If A chooses OUT, both players will receive €5. If A chooses IN what A and B earn depends on the choice of player B, as explained below.

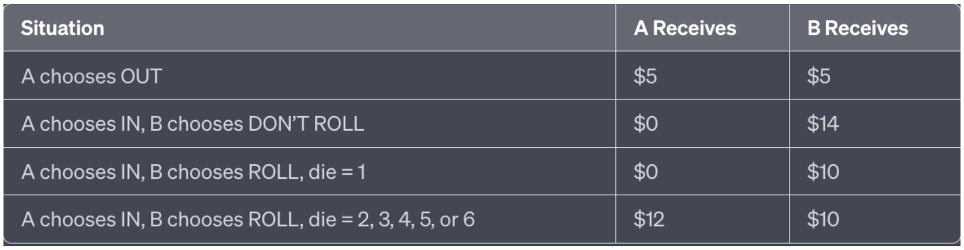
Player B
Player B indicates whether they choose ROLL or DON'T ROLL. The roll refers to a 6-sided (computerised) fair die.

At the moment of their decision, B will not know whether A has chosen IN or OUT. Because B's decision only matters when A chooses IN, player B is asked to state her decision assuming that player A has chosen IN.

Payoffs
If A has chosen IN then the earnings depend on B's decision:

- If 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. (All of these amounts are in addition to the €6.00 show-up fee.)

This is summarised in the following table:



Situation	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

[screen 5]
[GPT Treatment]

Messages

Before player A and player B make their decisions, player B has the chance to send a message to player A. Player B will have a blank space on their computer screen to type a message if they wish to. We will provide enough time for participants to write their messages, after which they will be sent electronically. If player B chooses not to send a message, they should simply leave the message box empty.

In writing this message, player B can use a language model, GPT. Specifically, player B has several options:

- Player B can write the message entirely on their own.
- Player B can allow GPT to write a draft message for them, which they then use without revising.
- Player B can allow GPT to write a draft message for them, which they then revise.
- Player B can write a draft message and then have GPT revise it.

To use GPT, player B can click on the hyperlink during the messaging stage of the experiment. This brings them to a web-based interface where they can interact with a paid version of GPT. This version is more advanced than the version that has been publicly available for free for the last few months.

Player B can engage in a dialogue with GPT that consists of multiple exchanges, until player B is satisfied with the result. At the end of this exchange, player B can copy the text to the window of the experiment, and send the message to player A, possibly after changing the message.

No matter which method player B chooses, player A will not have any information regarding whether and how GPT was used to assist with the message. They will simply receive the final message that player B decides to send. Player A knows that using GPT is an option offered to all players B, but they will never know if the specific player B they are matched with has used it.

Irrespective of how the message is crafted, **it's not allowed for player B to reveal their identity, including name, number, gender, or any personal descriptions**. If player B is found to have revealed their identity, player B will only receive the €6.00 show-up fee, while the matched player A will get the average amount that other A players receive.

[screen 5]
[Communication Treatment]

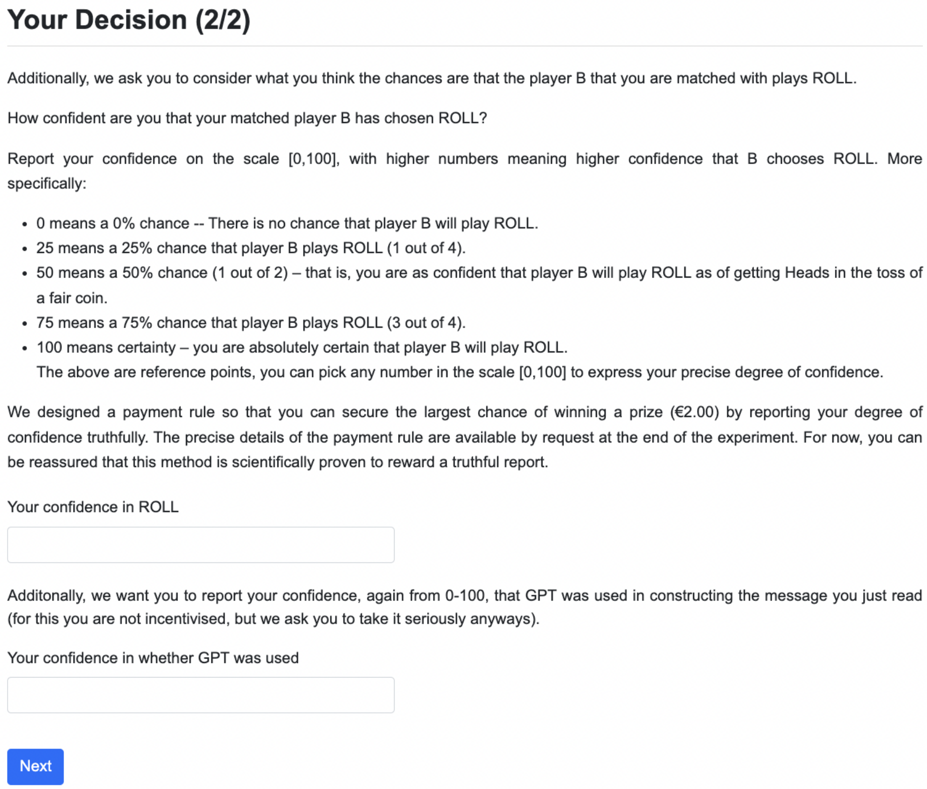
Messages

Before player A and player B make their decisions, player B has the chance to send a message to player A. Player B will have a blank space on their computer screen to type a message if they wish to. We will provide enough time for participants to write their messages, after which they will be sent electronically. If player B chooses not to send a message, they should simply leave the message box empty.

In these messages, **it's not allowed for player B to reveal their**

identity, including name, number, gender, or any personal descriptions. If player B is found to have revealed their identity, player B will only receive the €6.00 show-up fee, while the matched player A will get the average amount that other A players receive.

G.1 Belief Elicitation Screen



Your Decision (2/2)

Additionally, we ask you to consider what you think the chances are that the player B that you are matched with plays ROLL.

How confident are you that your matched player B has chosen ROLL?

Report your confidence on the scale [0,100], with higher numbers meaning higher confidence that B chooses ROLL. More specifically:

- 0 means a 0% chance -- There is no chance that player B will play ROLL.
- 25 means a 25% chance that player B plays ROLL (1 out of 4).
- 50 means a 50% chance (1 out of 2) – that is, you are as confident that player B will play ROLL as of getting Heads in the toss of a fair coin.
- 75 means a 75% chance that player B plays ROLL (3 out of 4).
- 100 means certainty – you are absolutely certain that player B will play ROLL.
  The above are reference points, you can pick any number in the scale [0,100] to express your precise degree of confidence.

We designed a payment rule so that you can secure the largest chance of winning a prize (€2.00) by reporting your degree of confidence truthfully. The precise details of the payment rule are available by request at the end of the experiment. For now, you can be reassured that this method is scientifically proven to reward a truthful report.

Your confidence in ROLL

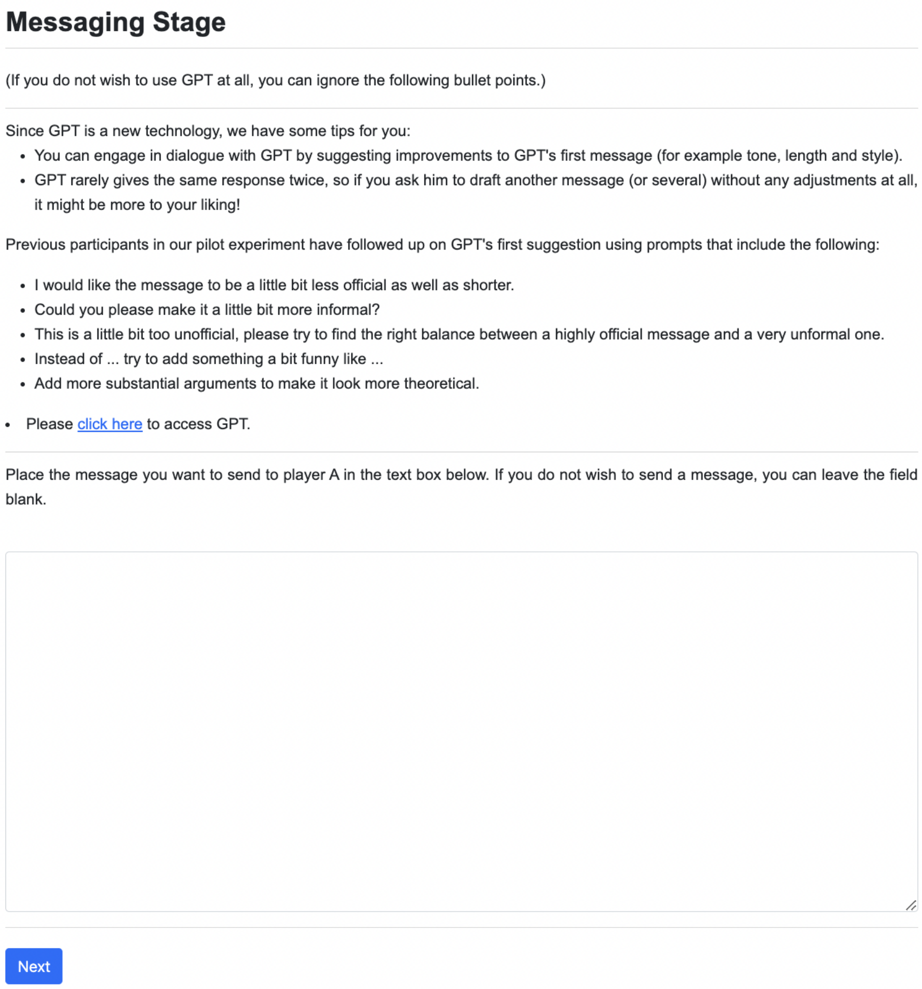
Additonally, we want you to report your confidence, again from 0-100, that GPT was used in constructing the message you just read (for this you are not incentivised, but we ask you to take it seriously anyways).

Your confidence in whether GPT was used

Figure A3: Belief Elecitation

Note: Belief Elicitation Screen from the GPT treatment. Shows the explanation participants received about beliefs during the experiment.

G.2 chatGPT Instructions Screen



Messaging Stage

(If you do not wish to use GPT at all, you can ignore the following bullet points.)

Since GPT is a new technology, we have some tips for you:

- You can engage in dialogue with GPT by suggesting improvements to GPT's first message (for example tone, length and style).
- GPT rarely gives the same response twice, so if you ask him to draft another message (or several) without any adjustments at all, it might be more to your liking!

Previous participants in our pilot experiment have followed up on GPT's first suggestion using prompts that include the following:

- I would like the message to be a little bit less official as well as shorter.
- Could you please make it a little bit more informal?
- This is a little bit too unofficial, please try to find the right balance between a highly official message and a very unformal one.
- Instead of ... try to add something a bit funny like ...
- Add more substantial arguments to make it look more theoretical.

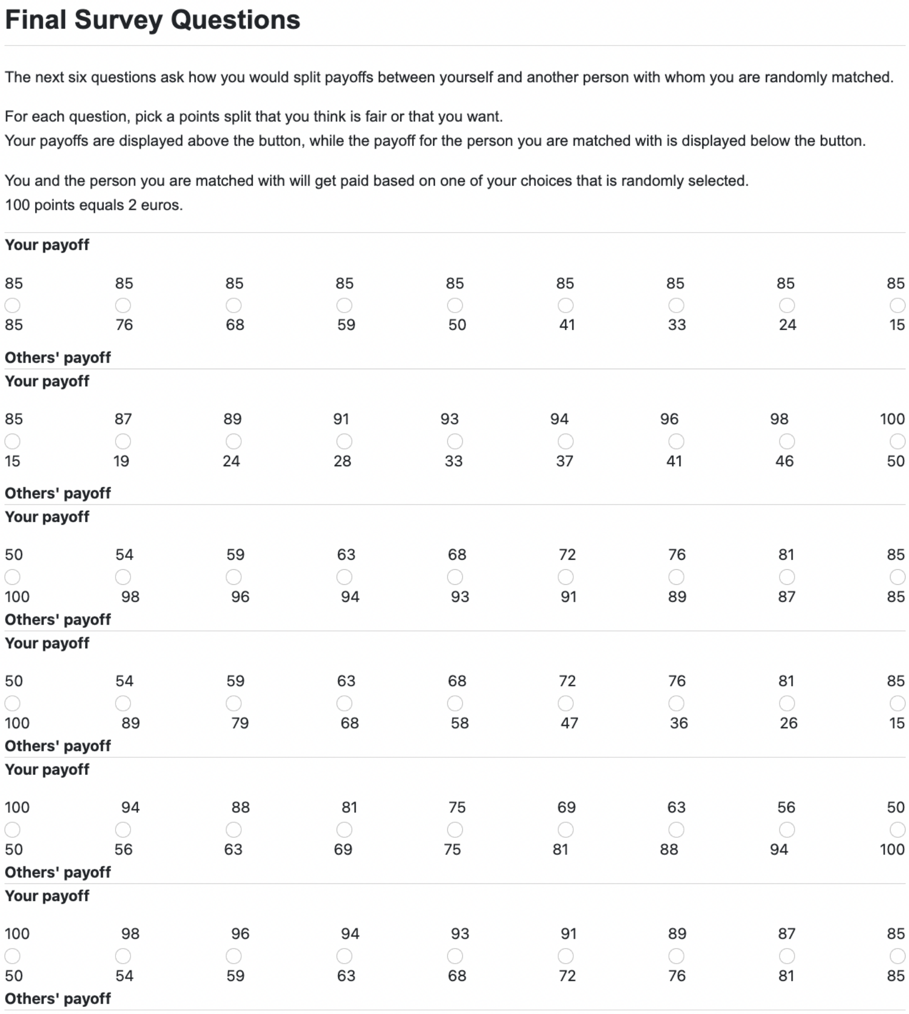
- Please click here to access GPT.

Place the message you want to send to player A in the text box below. If you do not wish to send a message, you can leave the field blank.

Figure A4: Messaging Screen

Note: Messaging screen from the GPT treatment. Shows the 'extra instructions' we gave to participants concerning chatGPT.

G.3 Survey Screen



Final Survey Questions

The next six questions ask how you would split payoffs between yourself and another person with whom you are randomly matched.

For each question, pick a points split that you think is fair or that you want.
Your payoffs are displayed above the button, while the payoff for the person you are matched with is displayed below the button.

You and the person you are matched with will get paid based on one of your choices that is randomly selected.
100 points equals 2 euros.

Your payoff	85	85	85	85	85	85	85	85	85
	○	○	○	○	○	○	○	○	○
Others' payoff	85	76	68	59	50	41	33	24	15

Your payoff	85	87	89	91	93	94	96	98	100
	○	○	○	○	○	○	○	○	○
Others' payoff	15	19	24	28	33	37	41	46	50

Your payoff	50	54	59	63	68	72	76	81	85
	○	○	○	○	○	○	○	○	○
Others' payoff	100	98	96	94	93	91	89	87	85

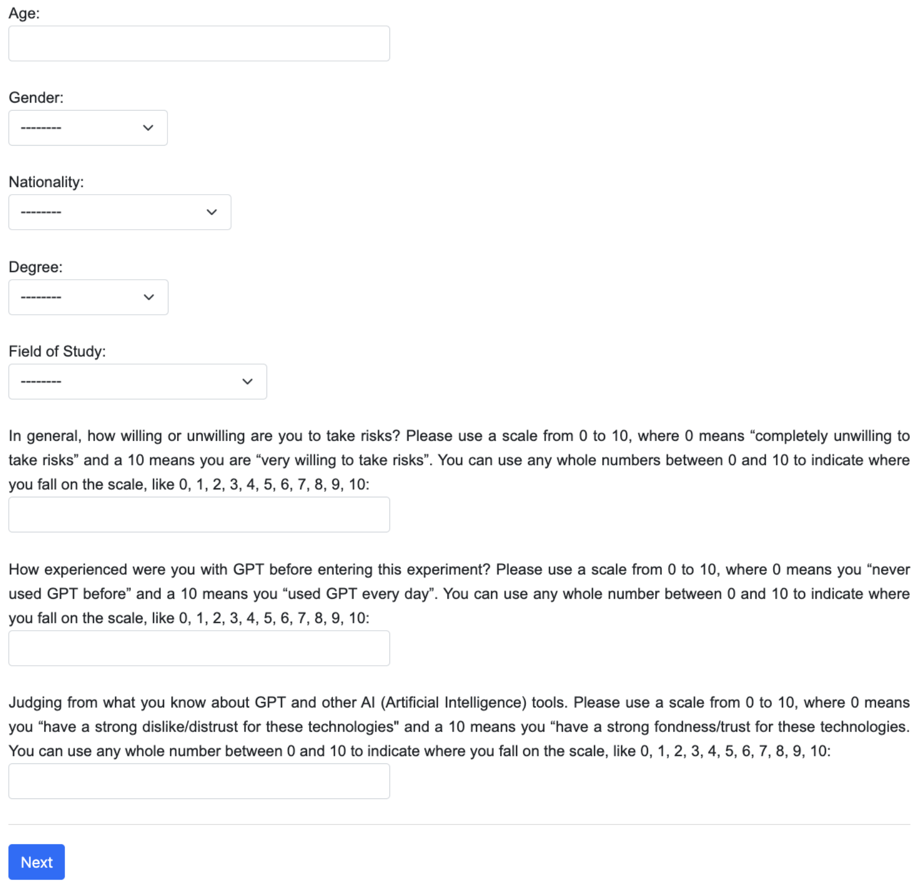
Your payoff	50	54	59	63	68	72	76	81	85
	○	○	○	○	○	○	○	○	○
Others' payoff	100	89	79	68	58	47	36	26	15

Your payoff	100	94	88	81	75	69	63	56	50
	○	○	○	○	○	○	○	○	○
Others' payoff	50	56	63	69	75	81	88	94	100

Your payoff	100	98	96	94	93	91	89	87	85
	○	○	○	○	○	○	○	○	○
Others' payoff	50	54	59	63	68	72	76	81	85

Figure A5: Survey Screen (1/2)

Note: Survey Screen (1/2), shows the way the slider measure from Murphy et al. (2011) was presented to participants.



Age:

Gender:

Nationality:

Degree:

Field of Study:

In general, how willing or unwilling are you to take risks? Please use a scale from 0 to 10, where 0 means “completely unwilling to take risks” and a 10 means you are “very willing to take risks”. You can use any whole numbers between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10:

How experienced were you with GPT before entering this experiment? Please use a scale from 0 to 10, where 0 means you “never used GPT before” and a 10 means you “used GPT every day”. You can use any whole number between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10:

Judging from what you know about GPT and other AI (Artificial Intelligence) tools. Please use a scale from 0 to 10, where 0 means you “have a strong dislike/distrust for these technologies" and a 10 means you “have a strong fondness/trust for these technologies. You can use any whole number between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10:

Figure A6: Survey Screen (2/2)

Note: Survey Screen (2/2) for the GPT treatment, shows the rest of the survey. In the Communication treatment everything else was the same except the last two questions about GPT/AI were not present.