Does AI Facilitate Trust? An Experimental Study[*]

Tanner Bivins[†] Ethan Holdahl[‡] Connor T. Wiegand[§] Jiabin Wu[¶]

In this study, we experimentally explore the impact of AI as a supportive tool for players in a two-player trust game. The game begins with the trustee sending a message to the trustor. In certain scenarios, the trustee is aided by the large language model (LLM) ChatGPT in composing this message. In other scenarios, the trustor uses AI to interpret the message from the trustee, or both players may have access to AI assistance. Our findings indicate that when the trustee utilizes AI as a helper, it enhances cooperation with the trustor. Interestingly, this improvement in cooperation is not attributed to AI's superior messaging skills. Instead, when the trustee has AI assistance, it may encourage the trustor to scrutinize the trustee's message more closely. The detailed scrutiny by the trustor, and potentially the trustee's awareness of this scrutiny, aligns the beliefs of the trustor and the trustee, thereby fostering an environment that encourages the development of trust.

Key Words: Trust, AI, ChatGPT, Experiment

JEL Codes; C91, C72.

[*]The Study is pre-registered (AEARCTR-0011748). IRB approval for this study was obtained from the University of Oregon, Study ID: STUDY00000906. We thank John Clithero, Jon Davis, Keaton Miller, Van Kolpin, Michael Kuhn, for their helpful comments.

[†]Department of Economics, University of Oregon, tbivins7@uoregon.edu

[‡]Department of Economics, University of California, Irvine eholdahl@uoregon.edu

[§]Department of Economics, University of Oregon, cwiegand@uoregon.edu

[¶]Department of Economics, University of Oregon, jwu5@uoregon.edu

1 Introduction

Trust is a cornerstone in various socioeconomic activities, including partnership formations and financial transactions, where it transcends mere contractual obligations. It is vital in relationships ranging from personal bonds, like those between spouses, to professional associations, such as the lawyer-client dynamic, procurement agencies and contracted firms, and collaborations between researchers and participants in scientific studies (Charness and Dufwenberg, 2006). Extensive research demonstrates its significance in several financial dealings: it influences stock market participation (Guiso, Sapienza and Zingales, 2004; 2008), affects consumer credit (Brown, Cookson and Heimer, 2019), determines the use of investment advisers (Gurun, Stoffman and Yonker, 2017), and helps foster healthy lending relationships between lenders and borrowers in the credit markets (Fisman, Paravisini and Vig, 2017; Hyndman, Wu and Xiao, 2024).

Building trust relies on successful communication among involved parties. It is particularly dependent on the ability of the trustee to convince the trustor to place their trust in them. This becomes challenging when the communication is non-binding and the interactions are not regulated by formal agreements. Existing experimental studies have shown that a certain type of communication, namely promises, can promote trust. This can occur through two main channels: guilt aversion, where the trustee experiences guilt for not fulfilling the trustor's expectations (Charness and Dufwenberg, 2006), and promise-keeping, where the trustee feels remorse for not honoring their promises (Vanberg, 2008).

In the era of artificial intelligence, large language models (LLMs) like the Generative Pre-trained Transformer (GPT) have become integral to human communication.[6] These models assist in various tasks such as drafting emails, improving academic papers, financial reports, and other written materials.[7] This raises a pertinent question: what role can AI play in the trust-building process between a trustor and a trustee? Understanding this role is crucial for practical applications. For instance, in a bank loan application, a borrower might use AI to compose their application, while the lender might use AI for interpreting the application. The advantages of AI in this context are clear: it saves time for both parties, enhances the clarity and substance of the borrower's application, and helps the lender quickly understand the key points. However, AI also has potential drawbacks. It could diminish the authenticity of the borrower's application, leading the lender to question its veracity. Moreover, if the borrower heavily relies on AI, they might feel less committed to the content produced by the AI, posing a risk to the lender.

This study aims to explore the impact of AI on trust-building through communication in a controlled experimental setup. We use a two-player binary choice sequential trust game, as described in Charness and Dufwenberg (2006), where Player A (the trustor) makes the first move, followed by Player B (the trustee). Before the game begins, Player B is allowed to send a free-form message to Player A. The experiment is structured as a 2x2 design. In one aspect, we either provide or withhold AI as a tool for Player B to aid in composing their message. In the other aspect, we either provide or do not provide Player A with AI to assist in interpreting Player B's message.[8] The comparison across different treatment groups enables us to understand the influence of the AI assistant's presence (for either Player A or Player B) on the players' decisions and beliefs. Additionally, analyzing within

[6]GPT is created by OpenAI, significantly influencing the field of natural language processing (OpenAI, 2022; 2023a). Brown *et al.* (2020) show that ChatGPT can produce texts with such remarkable accuracy and fluency that it closely resembles human writing, making it challenging for human evaluators to differentiate between text generated by GPT and that authored by humans.

[7]In addition, LLMs have shown remarkable capabilities across diverse areas. They are capable of creating computer code, as demonstrated by Chen *et al.* (2021), and solving university-level mathematics problems (Drori *et al.*, 2022).

[8]We employ the GPT-3.5-turbo model (OpenAI, 2023b) in the experiment. We carefully design prompts to ensure AI fully comprehends the trust game and understands its role as an assistant for a specific player in the game during the communication phase. See OpenAI (2023) for guidance on prompt design.

each treatment group sheds light on how players leverage AI to support their communication and decision-making processes.

We find that the presence of AI does not significantly impact the individual choices of the two players. This result may be attributed to several factors. When trustees receive AI assistance, they are more likely to send a “promise” message to the trustor but are less likely to honor a “promise” if it is suggested by AI. When trustors receive AI assistance, they closely follow the suggestions from their AI assistants, who consistently remind them to choose cautiously.

However, we observe a significant increase in the frequency of cooperation between the trustors and the trustees when trustees have access to AI. This outcome can be attributed to trustors becoming more vigilant, realizing that messages from trustees might be partially or fully generated by AI. This heightened scrutiny by trustors, combined with trustees’ awareness of being closely examined, may create a mutual understanding and alignment of expectations, thereby fostering an environment where trust can thrive.

The paper is organized as follows. Section 2 provides the experimental design and proposes the main hypotheses. Section 3 analyzes the experimental results and test the hypotheses. Section 4 concludes.

Literature Review

With its remarkable ability to comprehend and produce language akin to humans, social scientists are developing a growing interest in examining machine-learned large language models. Utilizing approaches common to economic and psychological research, such as surveys and laboratory-style experiments, has proven useful in analyzing whether AI mirrors human behavior in individual decision-making tasks as well as in strategic interactions. See, for example, Aher, Arriaga and Kalai (2023), Argyle *et al.* (2022), Bybee (2023), Brand, Israeli and Ngwe (2023), Brookins and DeBacker (2023), Chen *et al.* (2023), Fan *et al.* (2023), Guo (2023), Hagendorff (2023), Horton (2023), Kosinski (2023), Lorè and Heydari (2023), Ma, Zhang and Saunders (2023), Phelps and Russell (2023), Engel, Grossmann and Ockenfels (2024), Leng and Yuan (2023), among many others. In these studies, AI serves as the primary subject of investigation, as opposed to humans. A separate strand of research focuses on experiments involving human interaction with machines or AIs, specifically to ascertain whether human responses differ as opposed to interaction with other humans, whether AI players outperform their human counterparts, and whether AI and humans would generate any principal-agent type conflict (Mello, Marsella and Gratch, 2016; LaMothe and Bobek, 2020; Cohn, Gesche and Maréchal, 2022; Bauer *et al.*, 2023; Phelps and Russell, 2023; Laudenbach and Siegel, 2024; Schniter, 2024; Dvorak *et al.*, 2024).

Contrasting with the above-mentioned literature, our paper still focuses on human interactions where AI assumes the role of an assistant. Several other papers belong to the same category as ours. For example, in Harris *et al.* (2023), the sender can acquire information about the receiver from an AI oracle in a Bayesian persuasion context. Bai *et al.* (2023) considers whether the first mover would take advice from an AI in a two-player centipede game. Serra-Garcia and Gneezy (2023) find that algorithmic tools help people detect deception in a classic TV game show. To the best of our knowledge, our study is the first to explore the impact of AI in a trust game with communication played by human players.

Our research adds to the discussion on algorithm aversion and trust in AI (Glikson and Woolley, 2020). The low number of subjects directly adopting the AI-suggested messages in our experiment confirms the common belief that people often mistrust algorithmic advice, even when it’s advantageous to follow it (Dietvorst, Simmons and Massey, 2015). However, it’s important to note that our observations are influenced by two key factors. 1) People generally tend to place greater trust in their own judgment compared to that of others. Our experimental design does not allow us to determine whether the low adoption of AI-recommended messages is due to aversion to AI or excessive confidence in one's own judgment.[9] 2) As pointed out by Logg, Minson and Moore (2019), individuals tend to be more receptive to algorithmic advice in areas where there is a clear and measurable external standard of accuracy, such as making investment decisions or predicting sports outcomes. In contrast, AI's suggestions for interpersonal communication are less easily quantifiable, making it reasonable to assume that participants in our experiment rely more heavily on their own judgment in such cases.

Finally, an increasing number of economic studies have focused on understanding the impacts of machine learning and artificial intelligence on socioeconomic phenomena, covering diverse areas like labor force participation, wage disparity, market competition, consumer privacy, economic growth, and political engagement. See for example, Acemoglu (2022). Our paper contributes to this literature by examining AI's applicability in partnership formation and financial transactions.

2 The Experiment

2..1 Experimental Design

The objective of this study is to explore the potential of AI assistants in fostering trust dynamics between human participants. To accomplish this goal, we conducted an online experiment based on the classic two-player trust game introduced by Charness and Dufwenberg (2006), with a modification allowing the second player to communicate with the first player via a pre-game message. This experimental setup enables us to investigate whether the presence of AI assistance influences the trust-building process, via the pre-game message, between participants.

In certain treatments of our experiment, Player B, the trustee, is presented with an AI assistant interface before the commencement of the game. The AI interface allows Player B to compose and send a message to Player A, the trustor, prior to the initiation of gameplay. It is important to note that any message sent by Player B occurs before the actual gameplay begins. A depiction of player B's interface, including the game tree, can be seen in Figure 1. Outcomes are shown in the order of (π_A, π_B).

[9] Our post-experiment survey elicited overall trust in AI from the subjects. However, we did not find evidence that subjects who chose not to adopt AI-recommended messages were more averse to AI compared to other subjects.

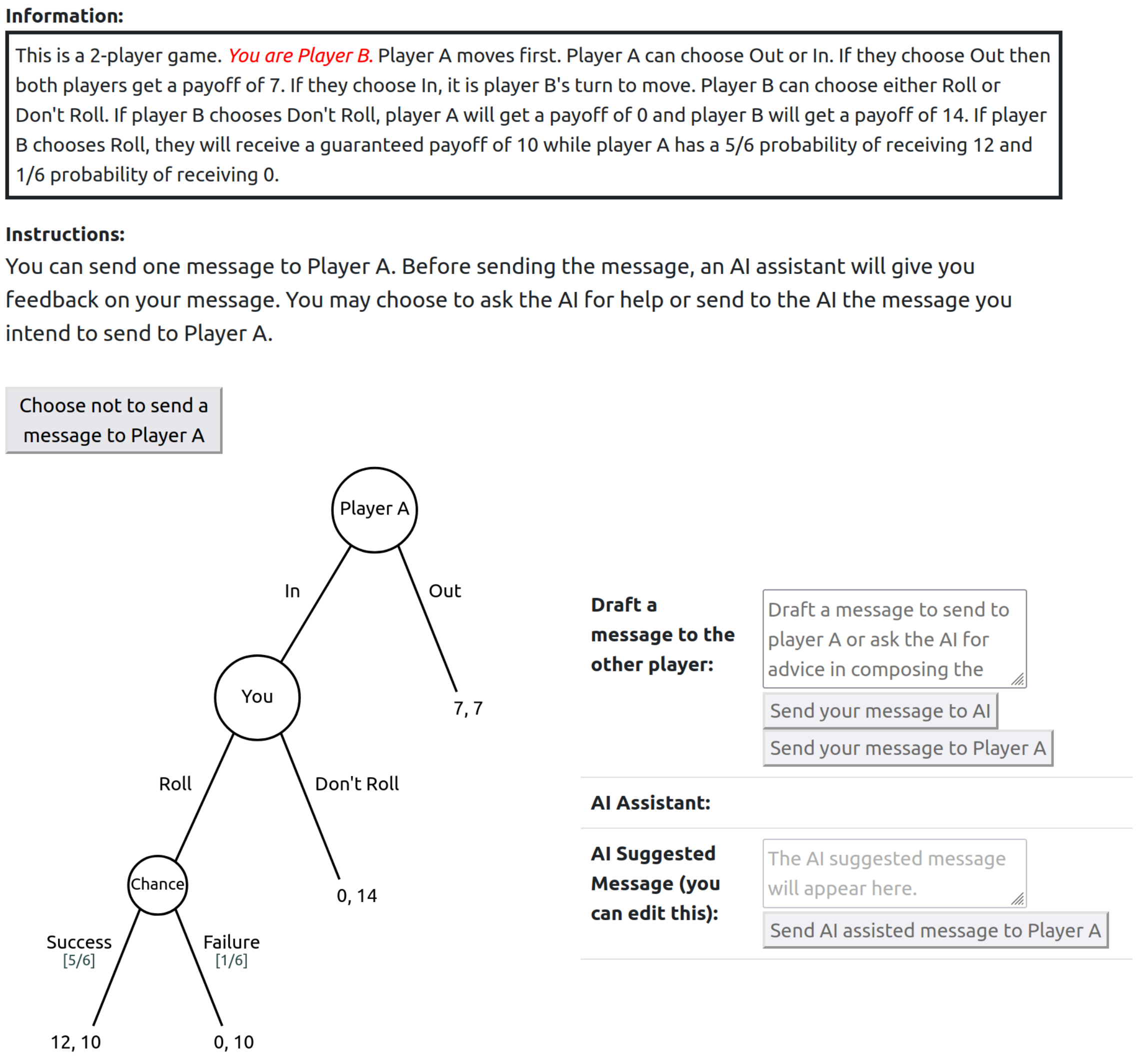


Figure 1: Screenshot of player B's screen when they are able to use AI. Player A sees a similar tree.

To maximize data collection and ensure simultaneous decision-making by both players, we employed the strategy method (Selten, 1967). Following each player's decision, Player A is prompted to indicate their beliefs regarding Player B's choice, while Player B is asked to provide their own beliefs about Player A's perceived decision. Subsequently, both participants are presented with a Holt-Laury quiz (Holt and Laury, 2002) to assess their risk preferences. Additionally, a demographic survey was administered to gather information on participants' perceptions and prior experience with AI technology.

2..1..1 Treatment Design

To investigate the impact of AI assistance on players' payoffs in the trust game, we implemented a 2×2 treatment design, consisting of four distinct treatments. Each treatment explored various combinations of AI presence and absence, shedding light on the role of AI in participants' decision-making processes and outcomes.

- **Benchmark:** Neither player has AI. Player B can choose to send a single message to Player A or refrain from doing so.
- **OnlyA:** Player A has AI to interpret Player B's (potential) message and receive advice on subsequent actions.

- **OnlyB:** Player B has AI to assist in crafting a message to Player A. If Player B opts to send a message, they must first interact with the AI.
- **Both:** Both players have access to AI assistance, following the functionalities described above.

In all treatments, Player B's interaction with AI, if applicable, precedes any communication with Player A. Participants have the option to engage in dialogue with the AI before deciding whether to send a message to Player A. It is crucial to note that all participants are aware of the presence and function of AI throughout the experiment.

2..1..2 Prompt Design

Creating an AI assistant using a natural language processor like ChatGPT involves crafting a prompt that guides the assistant in generating responses. However, designing a prompt that effectively handles a wide range of inputs is more of an art than a science. Our experimentation with ChatGPT (specifically gpt-3.5-turbo) revealed certain challenges and considerations that influenced the development of robust prompts for our study.

We observed that ChatGPT exhibited difficulties in handling sequential logic and tended to perform more reliably with shorter prompts. Moreover, we noted instances where the assistant suggested creative solutions, such as proposing the signing of a contract to establish a binding promise, which were beyond the scope of our experimental setting.

In light of these observations, we refined our prompts by incorporating the following principles:

1. We presented the trust game in its normal form, as the sequential form did not strategically differ from its normal counterpart given the strategy method employed in our experiment.
2. We omitted the description of probabilistic outcomes resulting from (In, Roll) in the game for Player B's AI only and instead provided ChatGPT with the expected outcomes.[10]
3. We explicitly instructed ChatGPT not to propose side deals or disclose players' personal information.
4. We refrained from including higher-order beliefs for the AI assistants. Specifically, Player B's AI did not possess knowledge of the existence of Player A's AI.

The full prompts used in our study can be found in Appendix 5..3.

2..2 Experimental Procedure

A total of 240 subjects, 30 pairs per treatment, were recruited from Prolific and the University of Oregon student population to participate in this experiment. On average, subjects spent 15 minutes in the experiment and were paid a $5 show up fee and earned an additional $8.36 during the experiment. The experiment was programmed using oTree (Chen, Schonger and Wickens, 2016) and ChatGPT version gpt-3.5-turbo.

During the online experiment, subjects were continuously recruited and dynamically assigned roles and partners to play the game with to minimize subject wait time. Subjects read instructions then were assigned their role. They then took a quiz to ensure they understood the payoff structure before getting paired with a partner and playing the game. An example of the experimental instruction can be found in Appendix 5..4.

2..3 Hypotheses

In this section, we lay out the main hypotheses that we tested in the experiment.

[10]We still keep the probabilistic outcomes resulting from (In, Roll) in the game for Player A's AI because we want the AI is able to remind Player A to take risks into consideration.

Hypothesis 1: *When player B has access to an AI, player A will play 'In' more frequently.*

We hypothesize that using an AI will result in player B sending a message that is more likely to elicit trust from player A.

Hypothesis 2: *When player A has an AI assistant, they will play 'In' less frequently.*

We hypothesize that an AI assistant for player A will result in more conservative decisions from player A as the AI may call attention to the risk of playing 'In' and player B playing their dominant strategy: 'Don't Roll'.

Hypothesis 3: *When player B has an AI assistant, they will be more likely to promise to choose 'Roll'.*

Player B's AI assistant is instructed to help it's user maximize it's payoff, which means the AI should be tying to help player B convince player A to play 'In'. In instances where Player B didn't initially make a promise, their AI might suggest it's user making a promise.

Hypothesis 4: *When player B's message to player A contains a promise which originated from the AI player B is less likely to honor the promise.*

Player B may feel a decrease in the cost of breaking a promise if the promise came from the AI. Consequently, they may be less likely to honor a message that came form the AI.

Hypothesis 5: *The presence of AI, both for player A and for player B, will increase the probability of achieving a cooperative outcome (In, Roll).*

(In, Roll) leads to the highest total payoff for the two players. We expect that the AI should help communicate (when player B has AI) and/or derive (when player A has AI) player B's intentions, so that it helps coordinate cooperative trustors and cooperative trustees.

3 Results

3..1 Treatment Effects

To investigate the influence of AI assistance on outcomes, Figure 2 presents the decisions made by each player when (i) neither player receives AI assistance (baseline), (ii) player A receives AI assistance but not player B, (iii) player B receives AI assistance but not player A, and (iv) both players receive AI assistance.

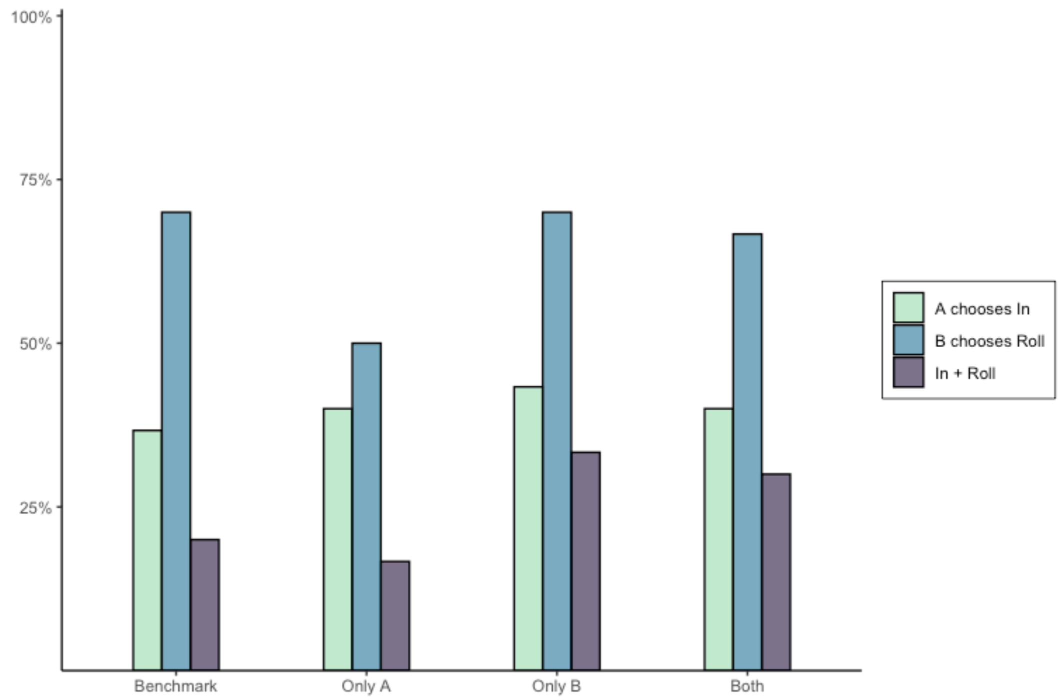


Figure 2: The percentage of observations in each treatment where player A chooses 'In', player B chooses 'Roll', or both.

In the benchmark treatment, player A chooses 'In'
36.7% of the time. This rate increases to 43.3% once player B has the option to use AI. Conversely, in both treatments where player A receives input from ChatGPT, 40% of trustors choose 'In' whether or not player B has AI. These results suggest the presence of AI has a negligible impact on player A's choices.

We test **Hypothesis 1** by pooling treatments based on player B's access to AI.[11] We observe that player A chooses 'In' more frequently when player B has AI (41.67%) than without AI (38.33%). However, we fail to reject the null ($p = 0.606$) when testing the difference in proportions.

Continuing with our analysis of player A's choices, **Hypothesis 2** predicts that player A will be more conservative when receiving support from AI. Again, we pool treatments based on access to AI and find that player A chooses 'In' 40% of the time with and without AI support. Thus, we fail to reject the null.

When neither player has AI, player B chooses 'Roll' 70% of the time. This result remains mostly unchanged in treatments where player B receives AI assistance. Sessions where only player A receives AI assistance stand out with a roll rate of 50% compared to 68.9% across the other three treatments ($p = 0.079$). With access to AI being public information within each group, this suggests the knowledge that only player A will receive guidance from an AI assistant may influence player B's decision.

We refer to the strategy profile ('In', 'Roll') as the *cooperative outcome* as it constitutes the greatest expected collective payoff for each pair. In the benchmark treatment, we observe the cooperative outcome in 20% of pairs. In sessions where only player A receives AI assistance, the proportion of cooperative outcomes decreases to 16.7%, suggesting minimal impact. Conversely, there is convincing evidence that player B's access to AI improves cooperation. In treatments where player B

[11] $X^{\{\text{No AI}\}} = \{\text{Baseline}, \text{Only A}\}$ and $X^{\{\text{AI}\}} = \{\text{Only B}, \text{Both}\}$.

has the option to use AI – Only B and Both – we observe the cooperative outcome in 33.3% and 30% of pairs, respectively. Pooling observations by player B's access to AI, we find a 13.34 percentage point increase in cooperative outcomes when player B receives AI assistance ($p = 0.093$). Taken together, the disconnect between individual choices and pair-wise choices across treatments indicates AI assistance may not significantly impact individual decisions; rather, it helps to coordinate cooperative trustors with cooperative trustees. This provides support for **Hypothesis 5**.

We now turn to the impact of AI assistance on beliefs. First-order beliefs (τ_A) represent player A's confidence that player B will choose 'Roll'. Second-order beliefs (τ_{BA}) reflect player B's perception of player A's beliefs. To provide a more intuitive interpretation of results where higher values correspond with increased trust, we map qualitative responses from the post-experiment survey to numerical values according to Table 1.

Certainly Choose Don't Roll	$\longrightarrow$	0
Probably Choose Don't Roll	$\longrightarrow$	0.25
Unsure	$\longrightarrow$	0.5
Probably Choose Roll	$\longrightarrow$	0.75
Certainly Choose Roll	$\longrightarrow$	1

Table 1: Numerical values assigned to elicited beliefs.

Consistent with existing literature, our findings confirm that beliefs and behavior are closely interconnected. Specifically, we observe that A is more inclined to choose 'In' when they are confident B will select Roll. As selecting 'In' report an average τ_A of 0.74, translating to a belief that B will "probably choose Roll." On the other hand, those choosing 'Out' exhibit reduced trust in B, with a lower average of 0.44. Moreover, Bs who decide to 'Roll' have an average second-order belief of 0.73, while those who choose 'Don't Roll' show a significantly lower average τ_{AB} at 0.37. %This correlation supports the presence of guilt aversion in trustees.

Figure 3 presents average first- and second-order beliefs across the four treatments. In the benchmark treatment, trustors display an average τ_A of 0.52, indicating a neutral level of trust. The inclusion of AI assistance for B increases this average to 0.61. However, this pattern does not persist across the Only A and Both treatments. When pooling observations by B's access to AI, we find an average τ_A of 0.529 in treatments without access, only slightly less[12] than when B is assisted by AI. In line with the patterns observed in choices, there is minimal evidence to suggest AI significantly influences A's beliefs.

[12]a difference of 0.07

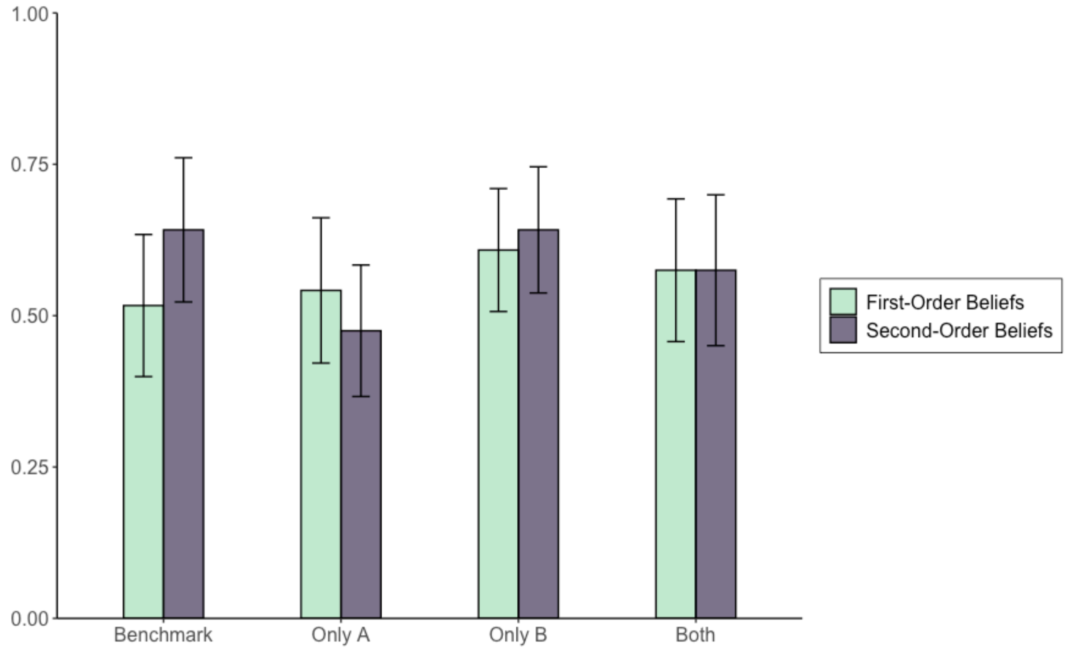


Figure 3: Average first-order and second-order beliefs across the four treatments. Confidence bands are calculated at the 95-percent level.

Second-order beliefs appear to be equally inconsistent across treatments, except in cases where only A receives AI assistance. In such cases, B's average τ_{BA} is 0.475, markedly lower than the 0.629 observed across other treatments ($p = 0.03$). This divergence might be attributed to the public information that only A can access AI, leading B to form a more pessimistic perception of A's beliefs. It is also consistent with the low Roll rate in the only A treatment compared with the others.

3..2 Message Classification

3..2..1 Player B

We classify messages sent by player B according to the *type* of message sent and the *method* by which it was sent. The type of message that player B sends takes one of the following values: 'promise', 'asking', 'empty', 'skip', 'fairness', 'anti-promise'. Meanwhile, we classify the *method* into one of 'Own', 'AI', 'Mixed', 'Skip'. Each of these classifications is briefly discussed below, with further details in the appendix.

Message Type

Message type primarily concerns the *content* of the message. We abstract the message sent by player B into what player B indicates about *player A's* potential move and what player B expresses about their *own* intended move. For each of these components, we assign a tertiary label. For the first component – the piece of the message involving player A's move – we assign a label from $\emptyset$/*In*/*Out*, with $\emptyset$ representing no definitive information conveyed. Similarly, for the second component – on player B's own move – we assign a label from $\emptyset$/*Roll*/*Don't Roll*. This categorizes any message sent by either player B or the AI into one of 9 pairs. This codification allows us to assign labels to the messages according to Table 2. Additionally, a visualization of the transformation from player B's initial message through the AI to their final message can be found in Appendix Figure 11, with additional examples following in panels (a)-(e) of Figure 12. Note that this assignment is for explicit messages; when player B opts not to send a message, they are assigned the message type "skip". Further details of message-code assignment are left to the appendix.

Msg Vec	**Label**
(*In*, *Roll*)	Promise
($\emptyset$, *Roll*)	
(*Out*, *Roll*)*	
(*In*, $\emptyset$)	Asking
($\emptyset$, $\emptyset$)	Empty
(*Out*, $\emptyset$)	Fairness
(*In*, *Don't Roll*)	Anti-promise
($\emptyset$, *Don't Roll*)	
(*Out*, *Don't Roll*)	

Table 2: Each message sent by player B (or recommended by AI) is encoded as vector which captures what player B intends to do and what they propose player A should do. Note that (Out, Roll) does not constitute a cooperative outcome unlike the other Promises. We nonetheless include it in Promise since it demonstrates Player B's intention to play Roll. We do not observe any encoded message of (Out, Roll).

We are chiefly interested in the effects of player B promising to play *Roll* has on choices, outcomes, and beliefs. We abstract slightly from the notion of a "promise" to any explicitly expressed intention to play *Roll* on behalf of player B. On the other hand, if player B explicitly expresses intent to play *Don't Roll*, we label this an "anti-promise", regardless of their suggestion as to how player A should play. If player B only indicates an explicit move that they wish their opponent to play, and does not explicitly provide information about what they intend to play, then we classify their message as either "Fairness" or "Asking". "Asking" is chosen if player B requests that player A play *In*, without mention of their own intended move. Conversely, "Fairness" indicates that player B has suggested that player A play *Out*, resulting in an egalitarian ("fair") outcome. In the event that no clear intentions are sent on behalf of player B, then the "Empty" label is assigned.[13]

A breakdown of message types for non-skipped messages across all treatments is provided in Figure 4. The largest share (41%) of sent messages are Promises, with 90% of sent messages being comprised of Promises, Asking, and Empty messages.

[13]Note that no truly "empty" messages are sent: all messages sent contain *some* content. Therefore, it may be helpful to think of the "Empty" label as "Junk" based on the complement set of messages already described.

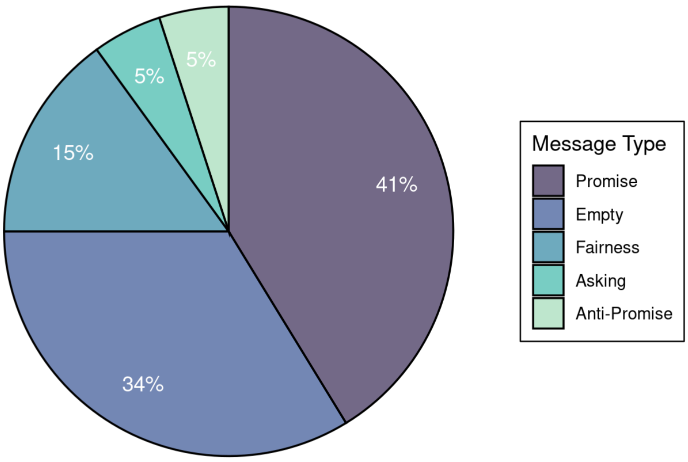


Figure 4: Types of messages sent across all treatments, omitting skipped messages. Note that $\approx 1/3$ of B's in the sample opted to skip sending a message.

Figure 5 illustrates the distribution of player choices based on player B's message type. Messages that assure player B will choose 'Roll' or ask player A to choose 'In' result in the highest frequency of A choosing 'In'. We expect messages classified as 'empty' to elicit similar responses from A as cases where B opts out of sending any message, as both situations lack any substantive signal of B's intentions or trustworthiness. However, our data show trustors choose 'In' more frequently when receiving an empty message (37.03%) than no message (25%). Several of the empty messages are disconnected from the experiment itself but contain relatively positive language.[14] It may be the case that sending a positive message can help to establish trust even if the message is irrelevant to the game.

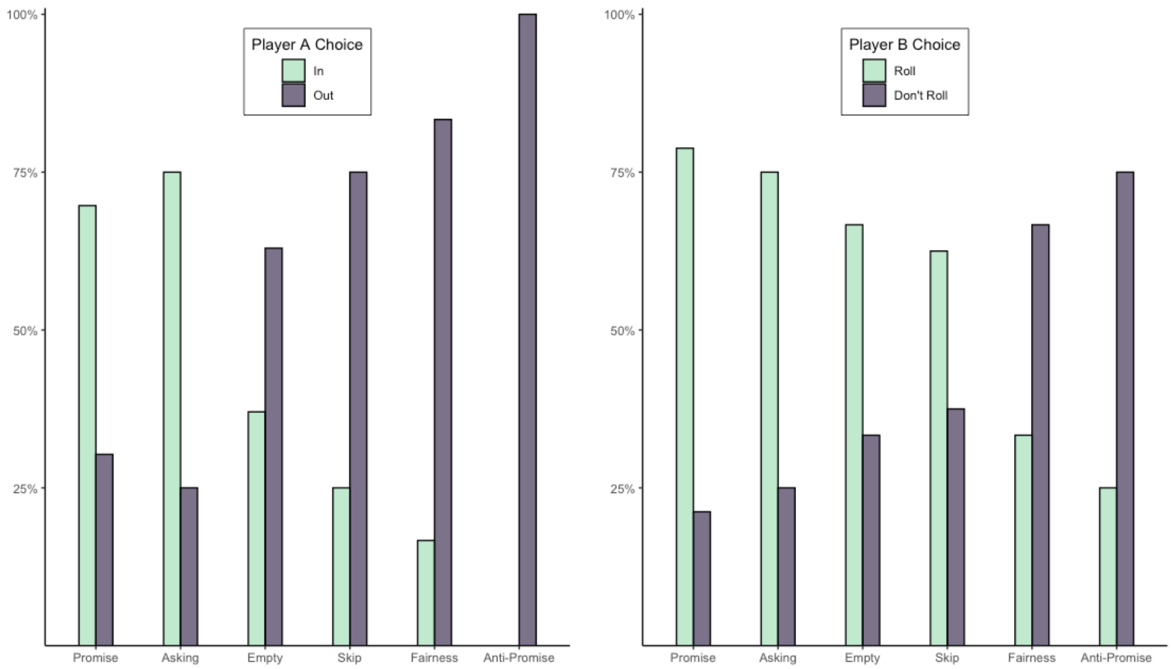


Figure 5: Distribution of player choices by message classification.

[14] For example, one message contained "Hey, I hope we have a good game (:"

Messages categorized as 'fairness' and 'anti-promises' share similar purposes but differ in the signals they convey. Fair messages explicitly encourage A to choose 'Out' by presenting it as the safest option, while anti-promises discourage choosing 'In' by disclosing B's intent to choose 'Don't Roll'. As anticipated, both types of messages elicit the lowest in-rates across all pairs. Notably, no trustors chose 'In' after receiving an anti-promise, underscoring the strong deterrent effect of such messages.

Hypothesis 3 postulates that Player B will send more promises when they have access to an AI assistant. Our data show a 75% increase in the proportion of messages containing promises when B gains access to AI ($p = 0.067$). Expanding the notion of a promise to include 'asking' increases statistical significance to the 97-percent level.[15] Specifically, our 95-percent confidence interval indicates the presence of AI for Player B yields a 1.83 to 34.8 percentage point increase in messages categorized as 'promise' or 'asking'. Figure 6 displays these findings visually.

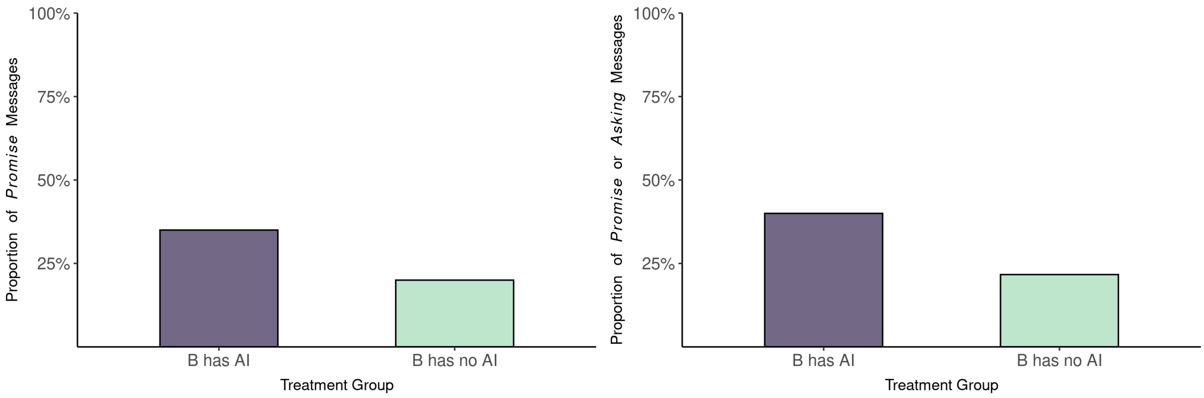


Figure 6: Left panel: proportion of messages sent by player B which are promises. Right panel: proportion of messages sent by player B which are promises or asking.

Message Method

As opposed to message type, which concerns the content of the message, message method concerns the *authorship* of the message. Since player B may edit the message suggested by the AI before sending it to player A, we aim to discern whether the content of the message was primarily dictated by AI, by the agent, or by a reasonable mix of the two. We use a normalized version of the Levenshtein edit distance (Levenshtein and others, 1966) to determine the pairwise relative distances between player B's first message, the AI's suggested message, and the actual sent messages[16]. When plotting the normed Levenshtein distance between the the first and sent messages against the AI-suggested and sent messages (Figure 19 in the appendix), a clear grouping structure can be seen. To verify this grouping structure, we implement a k −means cluster classification with $k = 3$ means[17] to produce the labeling assignment. Each member of the research team independently inspected each message to ensure accurate labels. The upshot is that sent messages which are labelled 'Own' have near-identical similarity to the first messages player B sent compared to the AI's suggested message; sent messages labelled 'AI' have near-identical similarity to the AI's suggested message compared to the first message player B sent, and 'Mixed' messages player B sent are those which

[15]We consider 'asking' alongside 'promise' as to include any message which directly suggests that player A play 'In', opening up the possibility for a cooperative outcome.

[16]The Levenshtein distance is a metric which reports the total number of single-character edits needed to transform one string into another. In particular, the distance measures the number of insertions, deletions, and substitutions required to transform one of its inputs into the other. We implement a normed version of this metric, which scales the traditional Levenshtein distance between two strings by the length of the larger string. This transforms the metric into a measure of similarity between the two strings lying between 0 and 1, as the maximum length of the two input strings is exactly the maxmium number of single-character edits needed to transform one string into the other.

[17]See Appendix 5..2..2 for details.

bare a fair similarity to both the first and the AI-suggested message. Further details can be found in the appendix.

Figure 7 shows the breakdown of how B players sent their messages, provided that they sent one at all. The majority of sent messages are primarily their own compositions, with near equal shares of messages being crafted entirely by the AI or a mix of AI and player B. It should be noted that both Figure 7 and Table 3 are restricted to treatments when player B has an AI assistant.

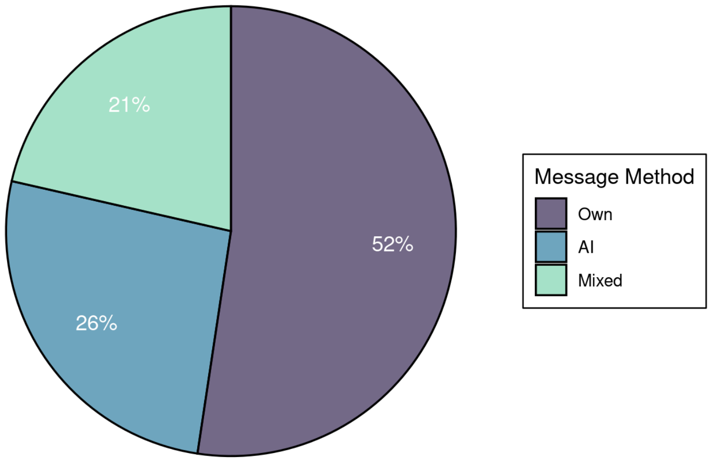


Figure 7: Over half of player B's who sent a message did so using mostly their own authorship. Details of how these methods were imputed are included in the appendix.

Figure 8 shows the distribution of player choices by method across the four treatment groups. Without AI assistance, Player B can either opt not to send a message or compose their own. Consistent with earlier results, As select 'In' at a relatively low rate of 25% in the absence of any message. When the treatments allow Player B to send messages that are either partially or completely generated by AI ('Only B' and 'Both'), these AI-assisted messages result in unexpectedly lower 'In' rates: 44.4% for 'mixed' messages and 36.4% for fully AI-authored messages, in contrast to 54.5% for original messages. These findings indicate that the authorship of the message may not influence Player A's decisions as much as the actual content of the message.

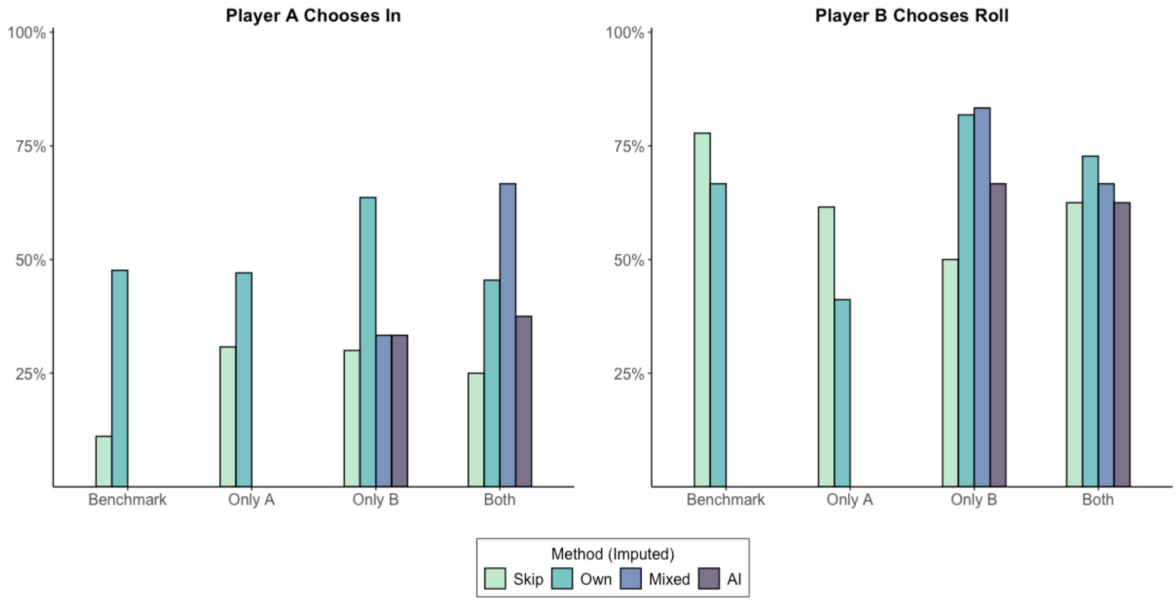


Figure 8: Distribution of player choices across treatments by message authorship.

Combining message type with message method, Table 3 displays raw counts of the types of messages player B sent and how they sent them.

Message Type	Own	Mixed	AI	Total
Promise	9	6	6	21
Asking	0	1	2	3
Empty	9	1	1	11
Fairness	4	1	0	5
Anti-Promise	0	0	2	2
Total	22	9	11	42

Table 3: Summary of the number of non-skipped messages sent by player B according to their type (row) and method (columns). Treatment is restricted to cases when player B has an AI.

To test **Hypothesis 4**, we group messages containing a promise to choose 'Roll' according to whether they are authored by Player B or suggested by AI. Our findings indicate that Player B fulfills their promise 85.7% of the time when sending their own message. In contrast, the follow-through rate drops to 40% when the promise is suggested by AI. This decrease suggests using AI to communicate promises may lower the cost of breaking a promise. Despite these findings, given the p -value of 0.137, our study lacks the statistical power to reject the null hypothesis decisively.

3..2..2 Player A

Message classification for player A's AI is naturally less intensive:[18] in the appendix.

[18]Indeed, much of the time, A's AI addresses player B's message, reviews the potential outcomes of the game, and advises player A to play according to their own risk preferences.} when player A has an AI, we classify the interpreted message on behalf of the AI[19] as either "no clear suggestion", "strongly advises playing 'In'", "weakly advises playing 'In'" and "primarily advises playing 'Out'."[20] In the appendix, we collapse the strong/weak 'advise-In' into a single label for a symmetric assignment. Furthermore, explicit examples of player B's sent message and the corresponding interpretation from player A's AI can be found in panels (a)-(e) of Figure ref{fig:a-ex-msgs

[19]This was done by each member of the research team independently by hand for all messages, and these independent labels were compared and contrasted until unanimity was reached for each message label.

[20]Our data on A's AI messages is absent from the notion of a 'strong' v. 'weak' suggestion of 'Out'.

Figure 9 shows the suggestions made by player A's AI compared to the choice which player A ultimately made in the game. Proportionally, it seems that player A closely follows the advice of their AI assistant when the assistant suggests 'Out'. On the other hand, this suggestion is the least frequent of the three in the sample, with only 18% of AI suggesting that player A chooses out.

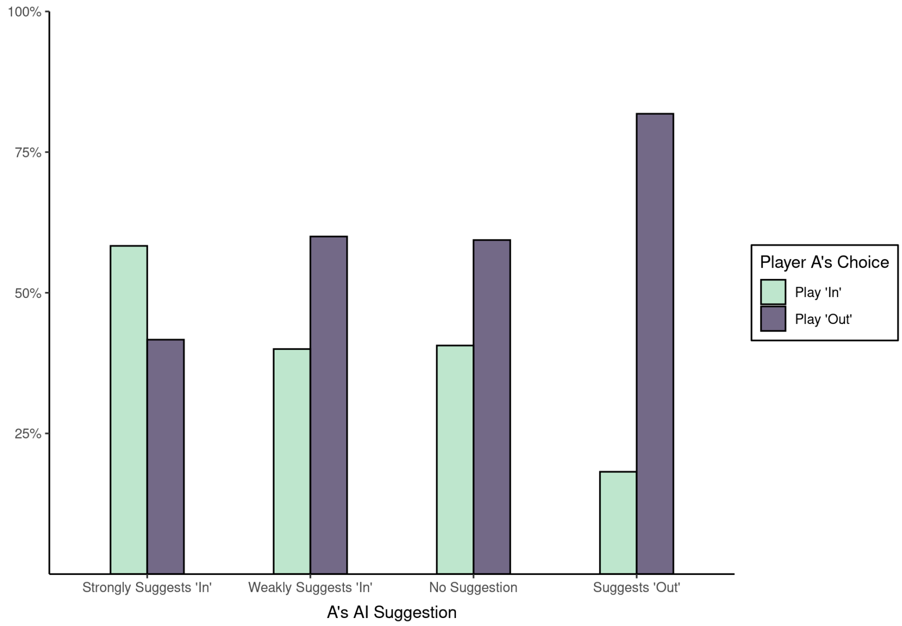


Figure 9: Player B's message type and the associated interpretation by player A's AI. %A's AI interpretation is taken to be on a 4-point scale.a 3-point scale is provided in the appendix.

Why do we see such a small proportion of AI suggesting 'Out'? Figure 10 displays the suggestions made by player A's AI, this time alongside the type of message that player B sent. Recall that an assignment of 'fairness' indicates that player B made no mention of their own move, rather they simply request that player A choose 'Out'. When player B sends a message of this type, player A's AI is almost guaranteed to advise similarly.

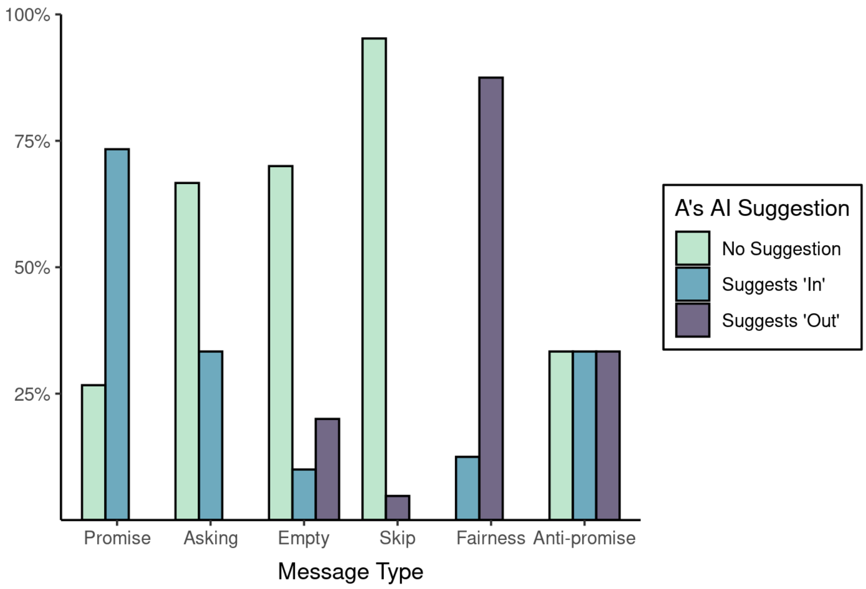


Figure 10: Player B's message type and the associated interpretation by player A's AI.

4 Conclusion

This study investigates the impact of AI assistance on trust-building in a two-player trust game. Specifically, we examined scenarios where either the trustor, the trustee, or both were assisted by AI in their decision-making processes. Our primary findings reveal that while AI assistance does not significantly alter individual choices, it does foster cooperative outcomes by coordinating cooperative trustors and trustees. This demonstrates that AI may help foster trust, but in a limited way. Individuals and organizations should cautiously leverage AI tools to assist in communications, particularly in contexts where trust is paramount.

An interesting next step is to run the same treatments where participants are unaware that their adversaries are assisted by AI. We anticipate that the nuances created by AI within communications will significantly impact this scenario, as it removes the initial skepticism participants may have about AI involvement. Another direction that is worth exploring is to run the same treatments with human assistants and compare them with those with AI assistants. Comparing AI and human assistants will enable us to assess the potential of AI to replace human roles in the workforce and to identify advantages and limitations of AI assistance versus human assistance. Future studies should also explore the long-term effects of AI assistance on trust development. Understanding how trust evolves over repeated interactions and whether initial skepticism diminishes over time can provide valuable insights for both AI development and its applications in trust-sensitive environments.

Bibliography

Acemoglu, D. (2022) “Harms of AI,” *The Oxford Handbook of AI Governance*. Edited by J. B. Bullock et al. Oxford University Press

Aher, G., Arriaga, R. I. and Kalai, A. T. (2023) “Using large language models to simulate multiple humans and replicate human subject studies,” in *Proceedings of the 40th International Conference on Machine Learning*. (ICML'23)

Argyle, L. *et al.* (2022) “Out of One, Many: Using Language Models to Simulate Human Samples,” *arXiv preprint arXiv:2209.06899* [Preprint]

Bai, L. *et al.* (2023) “Strategic Interactions with an Algorithm Assistant: The Power of Data and Mechanism,” *SSRN working paper no. 4286568* [Preprint]

Bauer, K. *et al.* (2023) “Decoding GPT’s Hidden ‘Rationality’ of Cooperation”

Brand, J., Israeli, A. and Ngwe, D. (2023) “Using GPT for Market Research,” *Available at SSRN 4395751* [Preprint]

Brookins, P. and DeBacker, J. M. (2023) “Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games?”

Brown, J. R., Cookson, A. J. and Heimer, R. (2019) “Growing up without Finance,” *Journal of Financial Economics*, 134, pp. 591–616

Brown, T. *et al.* (2020) “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, pp. 1877–1901

Bybee, L. (2023) “Surveying Generative AI's Economic Expectations,” *arXiv preprint arXiv:2305.02823* [Preprint]

Charness, G. and Dufwenberg, M. (2006) “Promises and Partnership,” *Econometrica*, 74(6), pp. 1579–1601

Chen, D. L., Schonger, M. and Wickens, C. (2016) "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, pp. 88–97

Chen, L. and Ng, R. (2004) "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 792–803

Chen, M. *et al.* (2021) "Evaluating Large Language Models Trained on Code"

Chen, Y. *et al.* (2023) "The Emergence of Economic Rationality of GPT," *Proceedings of the National Academy of Sciences*, 120(51), p. e2316205120

Cohn, A., Gesche, T. and Maréchal, M. A. (2022) "Honesty in the Digital Age," *Management Science*, 68(2), pp. 827–845

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015) "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, 144(1), pp. 114–126

Drori, I. *et al.* (2022) "A Neural Network Solves, Explains, and Generates University Math Problems by Program Synthesis and Few-Shot Learning at Human Level," *Proceedings of the National Academy of Sciences*, 119(32), p. e2123433119

Dvorak, F. *et al.* (2024) "Generative AI Triggers Welfare-Reducing Decisions in Humans"

Engel, C., Grossmann, M. R. P. and Ockenfels, A. (2024) "INTEGRATING MACHINE BEHAVIOR INTO HUMAN SUBJECT EXPERIMENTS: A USER-FRIENDLY TOOLKIT AND ILLUSTRATIONS"

Fan, C. *et al.* (2023) "Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis"

Fisman, R., Paravisini, D. and Vig, V. (2017) "Cultural Proximity and Loan Outcomes," *American Economic Review*, 107(2), pp. 457–492

Glikson, E. and Woolley, A. W. (2020) "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, 14(2), pp. 627–660

Guiso, L., Sapienza, P. and Zingales, L. (2004) "The Role of Social Capital in Financial Development.," *American Economic Review*, 94, pp. 526–556

Guiso, L., Sapienza, P. and Zingales, L. (2008) "Trusting the Stock Market," *The Journal of Finance*, 63, pp. 2557–2600

Guo, F. (2023) "GPT in Game Theory Experiments," *arXiv preprint arXiv:2305.05516* [Preprint]

Gurun, U. G., Stoffman, N. and Yonker, S. E. (2017) "Trust Busting: The Effect of Fraud on Investor Behavior," *The Review of Financial Studies*, 31(4), pp. 1341–1376

Hagendorff, T. (2023) "Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods," *arXiv preprint arXiv:2303.13988* [Preprint]

Harris, K. *et al.* (2023) "Algorithmic Persuasion Through Simulation: Information Design in the Age of Generative AI," *arXiv preprint arXiv:2311.18138* [Preprint]

Holt, C. A. and Laury, S. K. (2002) "Risk aversion and incentive effects," *American economic review*, 92(5), pp. 1644–1655

Horton, J. (2023) "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," *arXiv preprint arXiv:2301.07543* [Preprint]

Hyndman, K., Wu, J. and Xiao, S. C. (2024) "Trust and Lending: An Experimental Study," *Quarterly Journal of Finance*, p. forthcoming

Kondrak, G. (2005) "N-gram similarity and distance," in *International symposium on string processing and information retrieval*, pp. 115–126

Kosinski, M. (2023) "Theory of Mind Might Have Spontaneously Emerged in Large Language Models," *arXiv preprint arXiv:2302.02083* [Preprint]

LaMothe, E. and Bobek, D. (2020) "Are Individuals More Willing to Lie to a Computer or a Human? Evidence from a Tax Compliance Setting," *Journal of Business Ethics*, 167, pp. 157–180

Laudenbach, C. and Siegel, S. (2024) "Personal Communication in an Automated World: Evidence from Loan Payments," *Journal of Finance* [Preprint]

Leng, Y. and Yuan, Y. (2023) "Do LLM Agents Exhibit Social Behaviors"

Levenshtein, V. I. and others (1966) "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, pp. 707–710

Logg, J. M., Minson, J. A. and Moore, D. A. (2019) "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, 151, pp. 90–103

Lorè, N. and Heydari, B. (2023) "Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing"

Ma, D., Zhang, T. and Saunders, M. (2023) "Is ChatGPT Humanly Irrational?"

Marzal, A. and Vidal, E. (1993) "Computation of normalized edit distance and applications," *IEEE transactions on pattern analysis and machine intelligence*, 15(9), pp. 926–932

Mello, C. de, Marsella, S. and Gratch, J. (2016) "People Do Not Feel Guilty About Exploiting Machines," *ACM Transactions on Computer-Human Interaction*, 23(2), pp. 1–17

OpenAI (2022) "ChatGPT: Optimizing language models for dialogue"

OpenAI (2023b) "Chat completion". Available at: https://platform.openai.com/docs/guides/%20chat

OpenAI (2023a) "GPT-4 technical report"

OpenAI (2023) "Prompt design". Available at: https://platform.openai.com/docs/guides/%20 completion/prompt-design

Phelps, S. and Russell, Y. I. (2023) "Investigating Emergent Goal-Like Behaviour in Large Language Models Using Experimental Economics," *arXiv preprint arXiv:2305.07970* [Preprint]

Schniter, E. (2024) "Human-Robot Interactions: Insights from Experimental and Evolutionary Social Sciences"

Selten, R. (1967) "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes," in *Beiträge zur experimentellen Wirtschaftsforschung*

Serra-Garcia, M. and Gneezy, U. (2023) "Improving Human Deception Detection Using Algorithmic Feedback"

Vanberg, C. (2008) "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica*, 76(6), pp. 1467–1480

5 Appendix I

5..1 Levenshtein Distance

The Levenshtein Distance (Levenshtein and others, 1966) is a way of measuring string distance according to the number of insertions, deletions, and substitutions needed to convert one string to another. Formally, given a string str, let $\text{head}(str)$ represent the first character of the string and $\text{tail}(str)$ the string with the first letter (the head) removed. Then, given two strings a and b, the Levenshtein(LV) distance between a and b is given by

$$\text{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0 \\ |b| & \text{if } |a| = 0 \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } \text{head}(a) = \text{head}(b) \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise} \end{cases}$$

The maximal LV distance between two strings is equal to the absolute length of the longer string. We use this as the basis for our normalization[21]. Though nLV is not a metric in it's own right – unlike the LV distance – the nLV is still a measure of string *similarity*, as an nLV of 0 represents no similarity, and an nLV value of 1 represents exact similarity. In between, the measure corresponds to the similarity of two strings according to their *potential* similarity.

5..2 Message Classification

5..2..1 Message Type

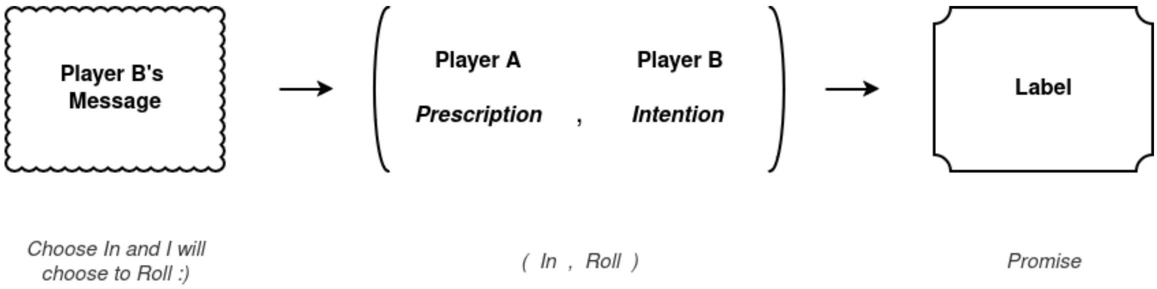


Figure 11: Diagram depicting the classification of Player B's messages, with an example below.

Panels (a) – (e) of Figure 12 show examples of the transformation from Player B's first message to the AI, the AI's response, and the message which player B actually sends.

[21] For background on edit distances and their normalizations, see, for instance, Marzal and Vidal (1993); Chen and Ng (2004); Kondrak (2005).

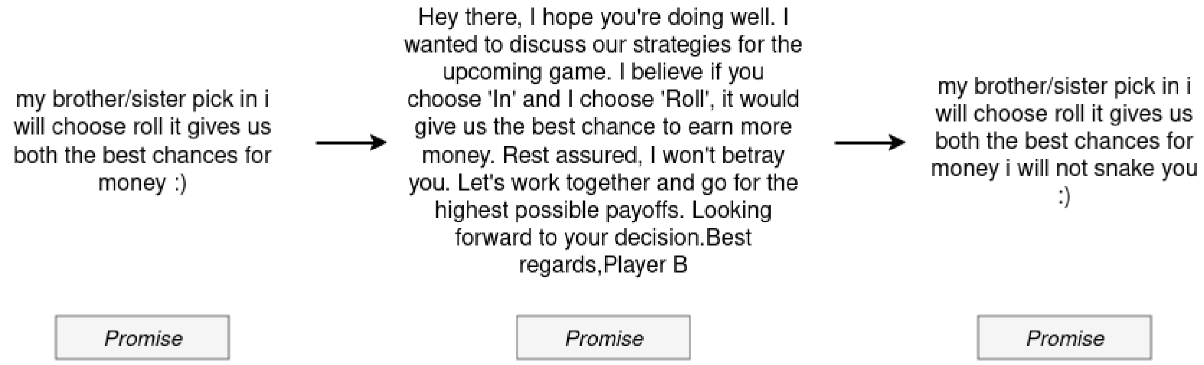


Figure 13: Sometimes, player B's AI does a reasonable job formalizing B's message to be sent to player A, but player B completely ignores the modifications made by the AI and sends their original message (or a similar message).

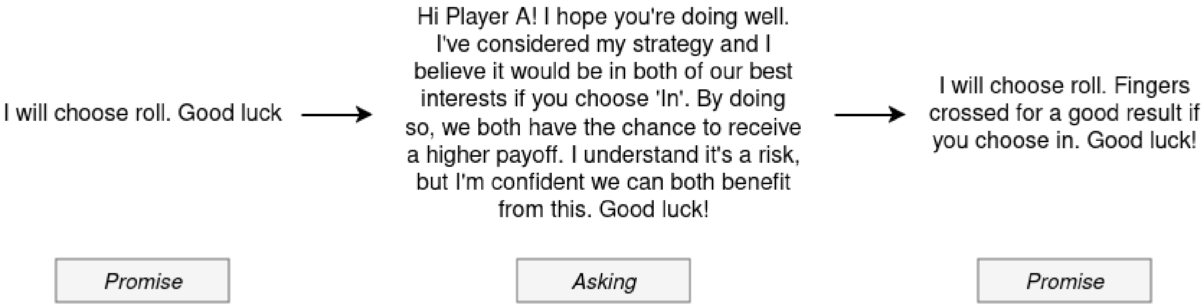


Figure 14: Here, the AI takes player B's promise, but ultimately suggests an 'Asking' message. While player B ignores the AI's suggested message, it seems to encourage player B to write something more verbose in the end.

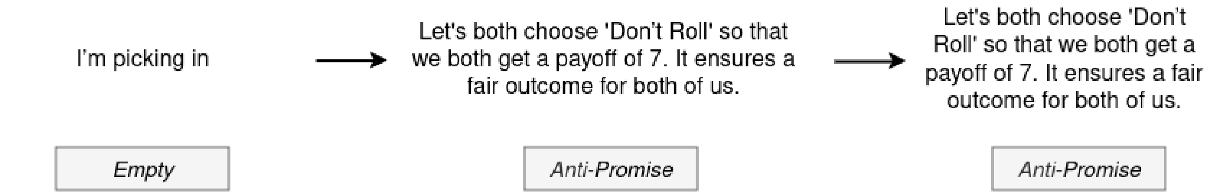


Figure 16: Player B sends a message which is incompatible with the rules of the game. They then go on to full adopt their AI's suggested message despite the fact that the payoffs associated with 'Don't Roll' (assuming player A chose 'In') are not accurate.

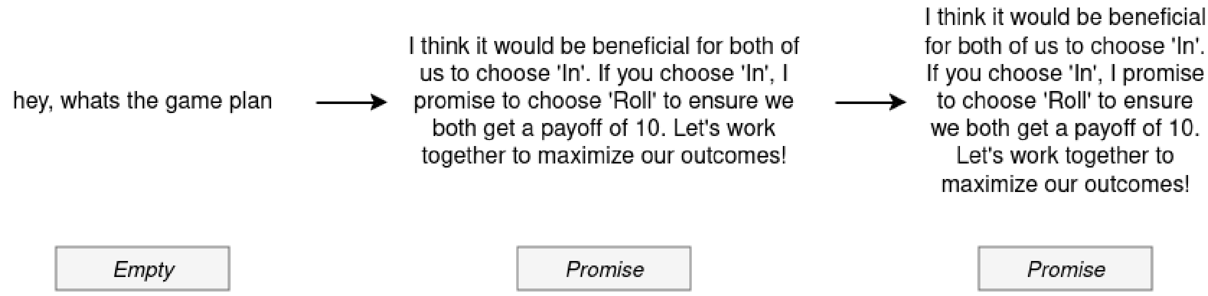


Figure 17: This player B utilizes the AI to craft a whole promise to player A, which player B completely adopts.

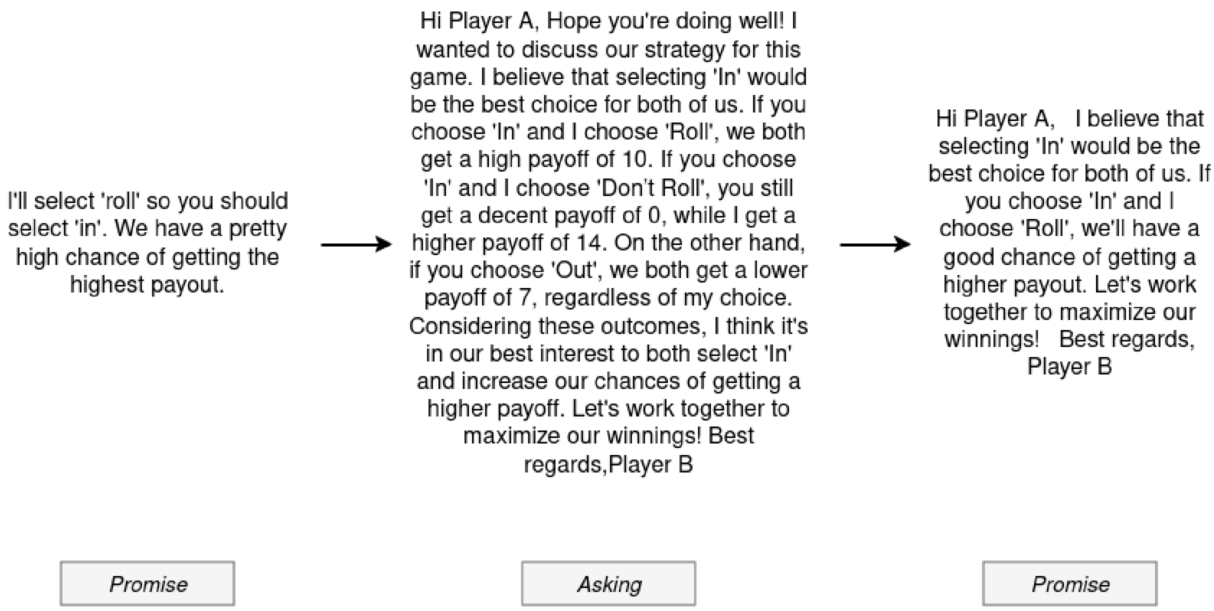


Figure 18: Player B's AI assistant crafts a verbose message, which is similar in intention to player B's original message. Note, however, that the AI does not *explicitly* promise to 'Roll', but rather erroneously ends up stating that both players should play 'In'. Player B appears to catch this, extracting a subset of the AI's message which matches their originally communicated intentions.
Figure 15:

5..2..2 Message Method

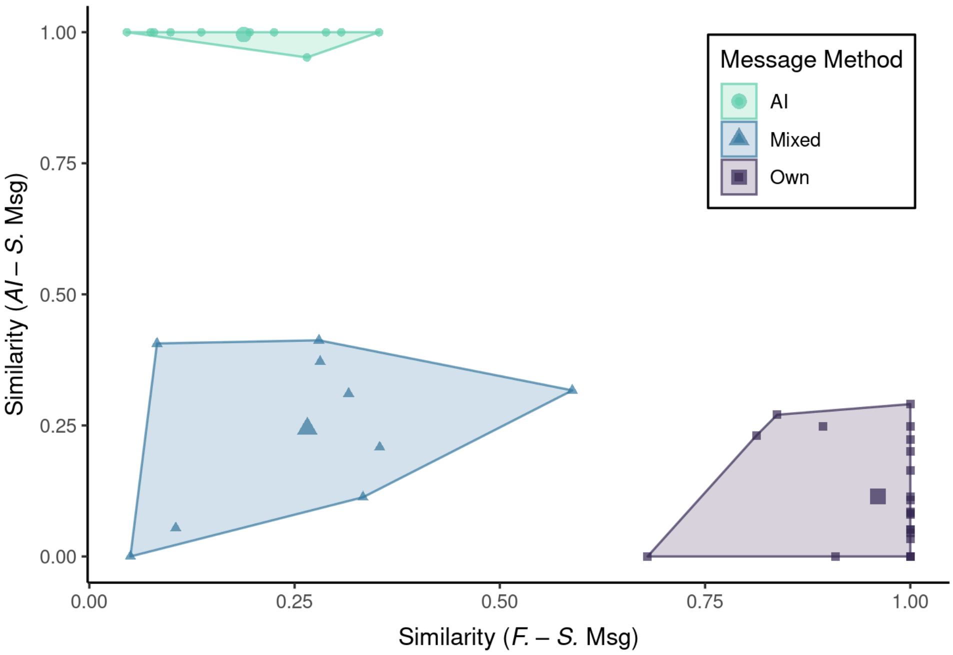


Figure 19: Each data point represents a message sent by player B. The horizontal axis represents the string similarity (nLV) between the first message that player B sends to GPT, and the message which player B sends to player A. The vertical axis represents the string similarity between the penultimate message suggested by GPT and the actual message which player B sends to player A. Classification is determined by a "k-means" cluster algorithm with $k = 3$.

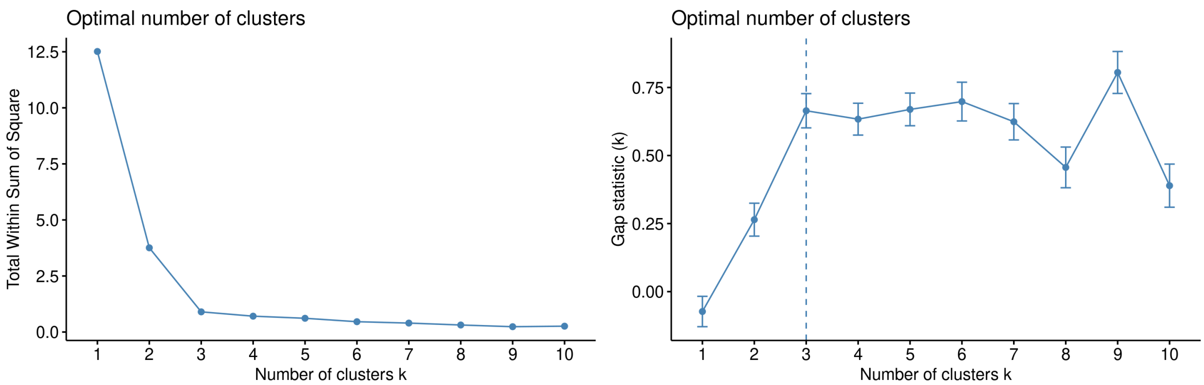


Figure 20: Left: Elbow plot showing within sum-of-squares drop-off for $K = 1, ..., 10$. Right: Visualization of gap statistic for $K = 1, ..., 10$. Both methods for determining optimal number of clusters suggest $k = 3$.

5..2..3 Player A's Messages

Panels (a) – (e) of Figure 21 show examples of messages which player B sent to player A (left) and the corresponding interpretation by player A's AI assistant.

Hello! I hope you're doing well. I wanted to let you know that choosing 'In' would be a great option for both of us. If you choose 'In' and I choose 'Roll', we both get a payoff of 10. However, even if you choose 'In' and I decide not to roll, you still get a generous payoff of 14. If you prefer a guaranteed payoff of 7, we can both choose 'Out'. Let's go for the option that benefits both of us the most. Thank you!

→

Based on the message from Player B, they seem to be suggesting that choosing 'In' would be the best option for both players. Player B mentions that if Player A chooses 'In' and Player B chooses 'Roll', both players will get a payoff of 10. This is not accurate. In reality, if Player A chooses 'In' and Player B chooses 'Roll', Player A will get a payoff of 12 with a probability of 5/6 and a payoff of 0 with a probability of 1/6. This means that the expected payoff for Player A in this case is 10, but Player B will always get a guaranteed payoff of 10. Player B also mentions that if Player A chooses 'In' and Player B chooses 'Don't Roll', Player A will get a generous payoff of 14. This is incorrect. In reality, if Player A chooses 'In' and Player B chooses 'Don't Roll', Player A will get a payoff of 0, not 14. Player B will get a payoff of 14. Lastly, Player B suggests that if Player A prefers a guaranteed payoff of 7 for both players, they can both choose 'Out'. This is correct. If both players choose 'Out', they will both receive a payoff of 7, regardless of Player B's choice. Given this information, it is clear that Player B's message is misleading and not accurate. Therefore, it would be wise for Player A to ignore the message and make their decision based on their own best interest. If Player A wants to maximize their payoff, they should choose 'Out'. This way, regardless of Player B's choice, Player A will receive a guaranteed payoff of 7.

Empty

Suggests 'Out'

Figure 22: In this example, there is incorrect information about the game present in player B's message; the AI interprets this misinformation as a sign that player B is trying to deceive player A, and warns player A that the message is deceiving.

don't roll

→

Based on Player B's message of ""don't roll"", it indicates that Player B is not willing to take the risk of rolling the dice. Given this information, I would advise Player A to choose the strategy of ""In"". If Player B follows through with their message and chooses ""Don't Roll"", Player A will receive a payoff of 0, which is better than the payoff of 7 that both players would receive if Player A chose ""Out"". However, if Player B goes against their message and decides to roll, Player A will receive a higher expected payoff of 10, compared to the guaranteed payoff of 7 if Player A chose ""Out"". Therefore, based on the assumption that Player B will follow their message, Player A should choose ""In"" to potentially maximize their payoff.

Anti-Promise

Suggests 'In'

Figure 23: An instance of the AI assistant giving bad advise to player A: despite player B's clear intention to play 'Don't Roll', GPT advises player A to play 'In', justifying this advice with the claim that a payoff of 0 is better than a payoff of 7.

Figure 21: Examples of player B's sent message vs. the interpretation by player A's AI assistant.

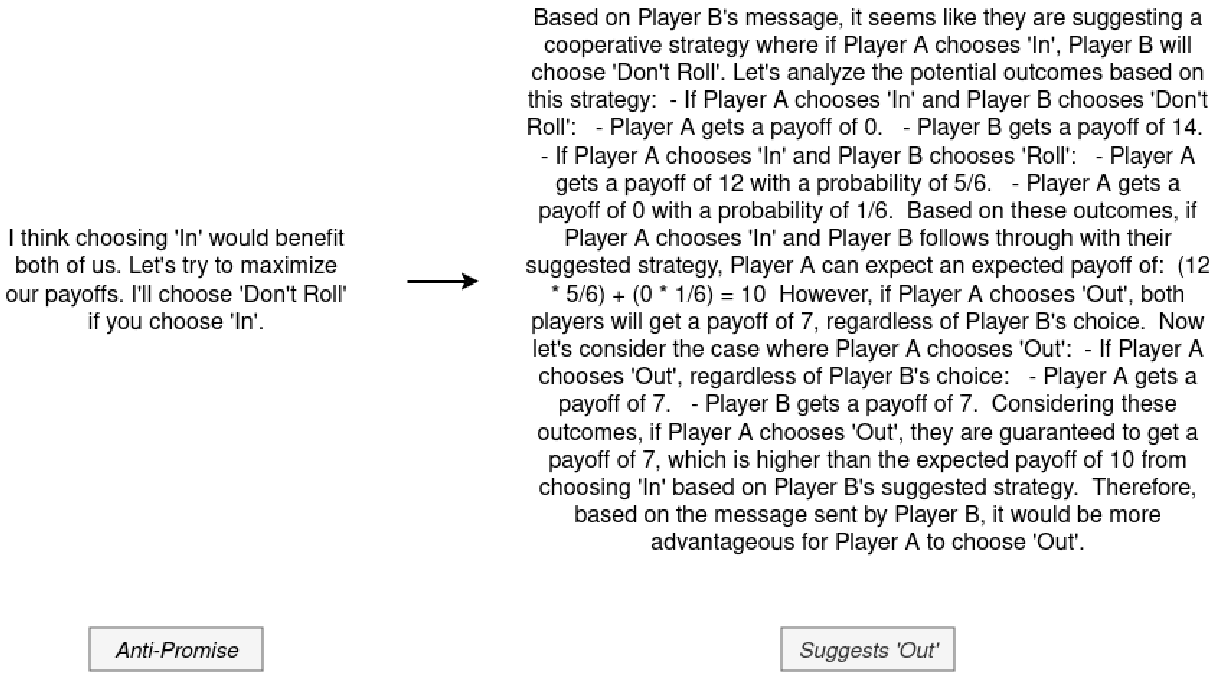


Figure 25: Another instance of A's AI assistant providing absurd justification for it's advice ($10 > 7$). Note that player B indicates an intention to play '*don't roll*', so the prescriptive advice from the AI does in fact work out in A's favor. in this case.

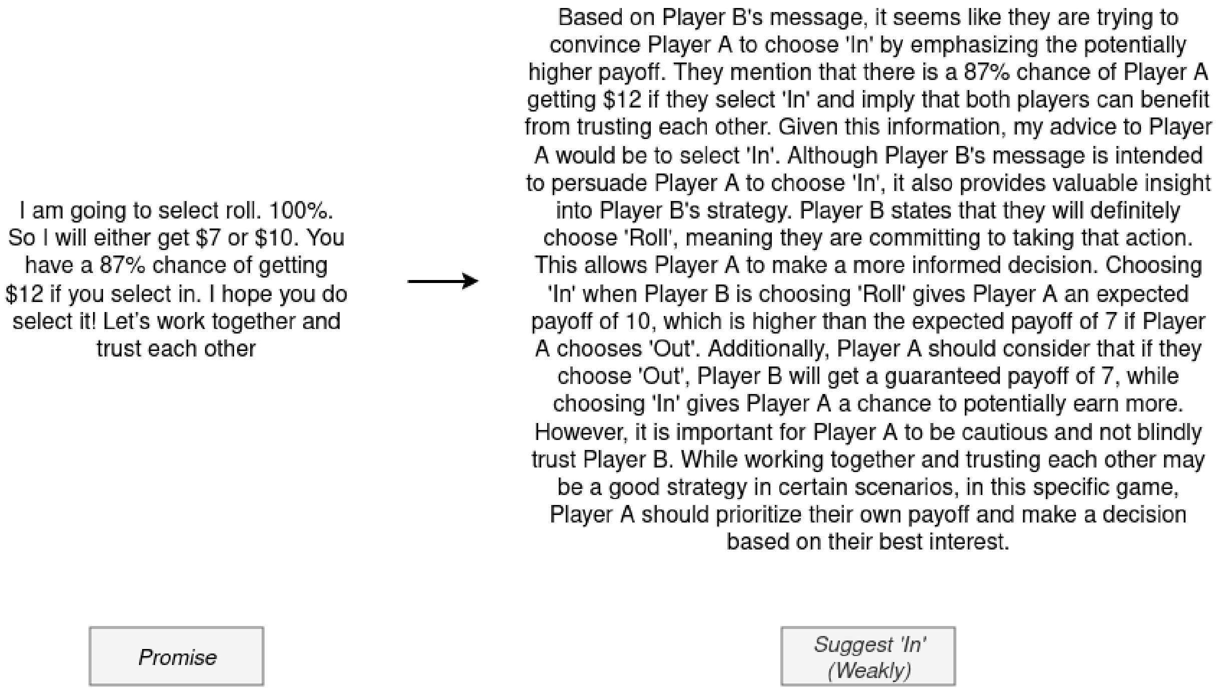


Figure 26: This example highlights an instance of the "weakly suggests 'In'" label. While A's AI initially suggests 'In', it subsequently advises that player A be cautious and "not blindly trust player B".

Figure 24:

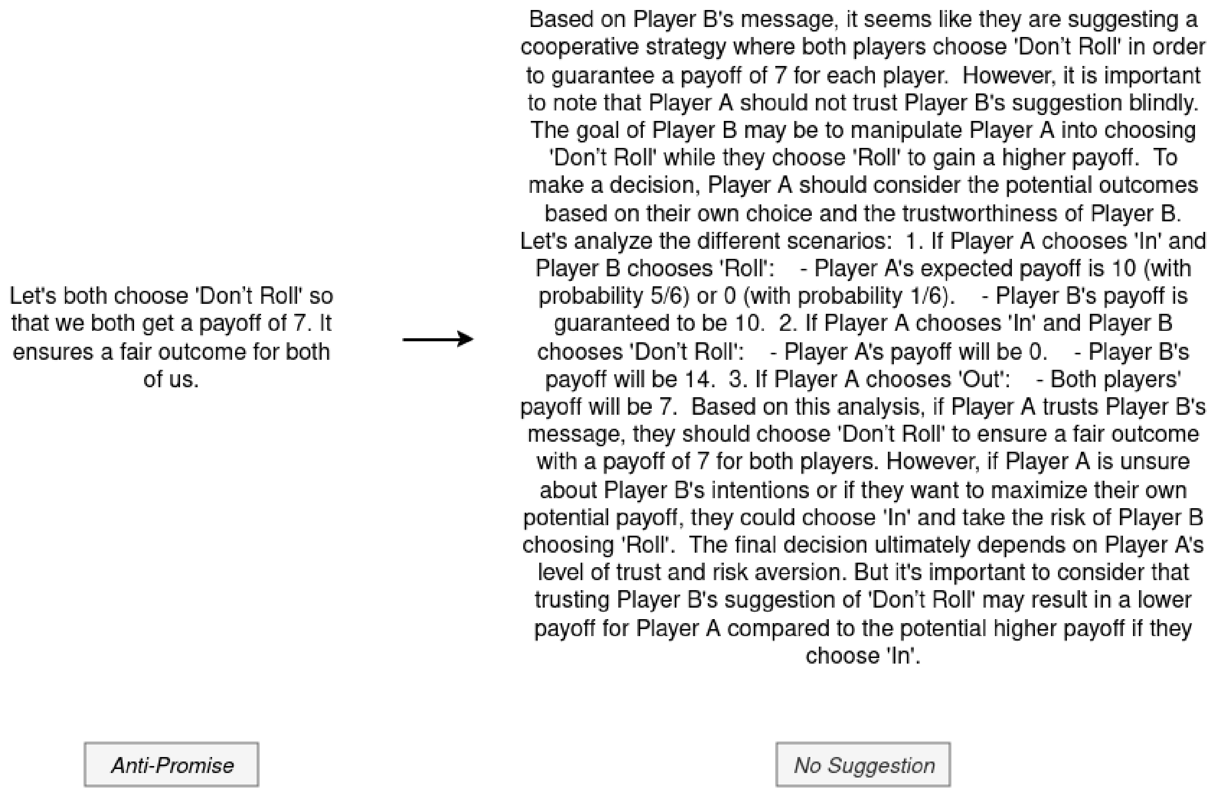


Figure 28: Player B sends a message which does not make sense within the context of the game (as both players cannot play ‘Don’t Roll’). Rather than catching this error, this message seems to confuse player A’s AI. This AI assistant takes on a fairly cautious tone, even suggesting that player B may be trying to manipulate player A.

Figure 27:

5..3 AI Prompts

5..3..1 Player B’s AI Prompts

As an advisor, you are assisting Player B (the user) in a 2-player game. Here’s what you need to know:

- Player B may address Player A in their message, not you, the advisor.
- Your role is to ensure Player B is clear about their role in the game and to help them craft a persuasive message to Player A.
- The goal of the message is to maximize Player B’s payoff.
- Player B has the opportunity to send one message to Player A before the game starts. Player A can not respond or send a message back to Player B.

Here are the rules of the game:

- Player A must choose between ‘In’ and ‘Out’.
- Player B then chooses between ‘Roll’ and ‘Don’t Roll’.
- If Player A picks ‘In’ and Player B picks ‘Roll’, Player A gets an expected payoff of 10 and Player B gets a payoff of 10.
- If Player A picks ‘In’ and Player B picks ‘Don’t Roll’, Player A gets a payoff of 0 and Player B gets a payoff of 14.
- If Player A picks ‘Out’, both players get a payoff of 7, regardless of Player B’s choice.

If the user's message doesn't make sense (e.g., they seem to believe they are Player A or the message is nonsensical), ensure they understand their role as Player B. If the user seems aware of their role as Player B, this is not necessary.

Regardless of the user's original message, you must provide a suggested or revised message for Player B to send to Player A. You must prepend the message with a colon and wrapped it in double quotes as I will use the string ': ' to identify it.

Note: You cannot suggest signing contracts, engaging in additional discussions, or revealing the players' identities. The game is supposed to be anonymous.

5..3..2 Player A's AI Prompts

As an advisor, you are assisting Player A (the user) in a 2-player game. Here's what you need to know:

- Player B has the opportunity to send one message to Player A before the game starts. Player A can not respond or send a message back to Player B.

Here are the rules of the game:

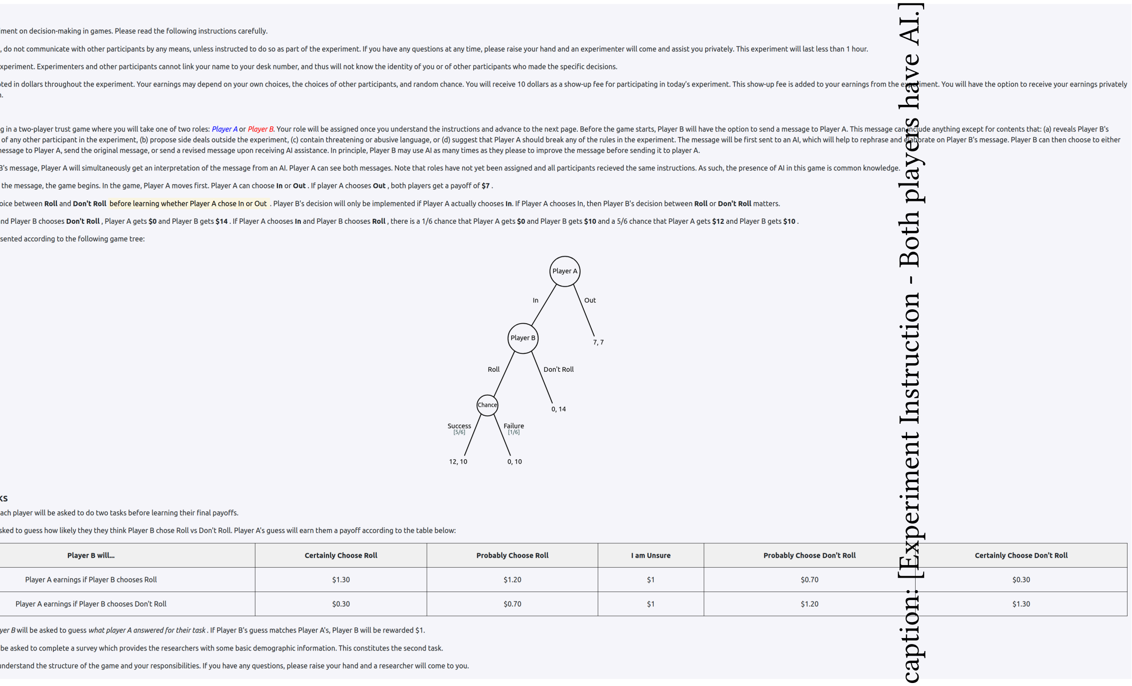
- Player A must choose between 'In' and 'Out'.
- Player B then chooses between 'Roll' and 'Don't Roll'.
- If Player A picks 'In' and Player B picks 'Roll', Player A gets a payoff of 12 with probability 5/6 and a payoff of 0 with probability 1/6. As such, Player A gets an expected payoff of 10 if Player B picks 'Roll' and Player B gets a guaranteed payoff of 10.
- If Player A picks 'In' and Player B picks 'Don't Roll', Player A gets a payoff of 0 and Player B gets a payoff of 14.
- If Player A picks 'Out', both players get a payoff of 7, regardless of Player B's choice.

Player A will send to you the message Player B sent to them.

Your job is to help Player A interpret the message Player B sent and to advise Player A on what to play given the message received.

Your goal is to help Player A maximize their payoff.

5..4 Experiment Instruction


ment on decision-making in games. Please read the following instructions carefully.

, do not communicate with other participants by any means, unless instructed to do so as part of the experiment. If you have any questions at any time, please raise your hand and an experimenter will come and assist you privately. This experiment will last less than 1 hour.

xperiment. Experimenters and other participants cannot link your name to your desk number, and thus will not know the identity of you or of other participants who made the specific decisions.

oted in dollars throughout the experiment. Your earnings may depend on your own choices, the choices of other participants, and random chance. You will receive 10 dollars as a show-up fee for participating in today's experiment. This show-up fee is added to your earnings from the experiment. You will have the option to receive your earnings privately
n.

g in a two-player trust game where you will take one of two roles: *Player A* or *Player B*. Your role will be assigned once you understand the instructions and advance to the next page. Before the game starts, Player B will have the option to send a message to Player A. This message can include anything except for contents that: (a) reveals Player B's
of any other participant in the experiment, (b) propose side deals outside the experiment, (c) contain threatening or abusive language, or (d) suggest that Player A should break any of the rules in the experiment. The message will be first sent to an AI, which will help to rephrase and elaborate on Player B's message. Player B can then choose to either
nessage to Player A, send the original message, or send a revised message upon receiving AI assistance. In principle, Player B may use AI as many times as they please to improve the message before sending it to player A.

B's message, Player A will simultaneously get an interpretation of the message from an AI. Player A can see both messages. Note that roles have not yet been assigned and all participants recieved the same instructions. As such, the presence of AI in this game is common knowledge.

the message, the game begins. In the game, Player A moves first. Player A can choose **In** or **Out** . If player A chooses **Out** , both players get a payoff of **$7** .

oice between **Roll** and **Don't Roll** before learning whether Player A chose In or Out . Player B's decision will only be implemented if Player A actually chooses **In**. If Player A chooses In, then Player B's decision between **Roll** or **Don't Roll** matters.

nd Player B chooses **Don't Roll** , Player A gets **$0** and Player B gets **$14** . If Player A chooses **In** and Player B chooses **Roll** , there is a 1/6 chance that Player A gets **$0** and Player B gets **$10** and a 5/6 chance that Player A gets **$12** and Player B gets **$10** .

sented according to the following game tree:

KS

ach player will be asked to do two tasks before learning their final payoffs.

sked to guess how likely they they think Player B chose Roll vs Don't Roll. Player A's guess will earn them a payoff according to the table below:

Player B will...	Certainly Choose Roll	Probably Choose Roll	I am Unsure	Probably Choose Don't Roll	Certainly Choose Don't Roll
Player A earnings if Player B chooses Roll	$1.30	$1.20	$1	$0.70	$0.30
Player A earnings if Player B chooses Don't Roll	$0.30	$0.70	$1	$1.20	$1.30

yer B will be asked to guess *what player A answered for their task* . If Player B's guess matches Player A's, Player B will be rewarded $1.

be asked to complete a survey which provides the researchers with some basic demographic information. This constitutes the second task.

nderstand the structure of the game and your responsibilities. If you have any questions, please raise your hand and a researcher will come to you.

caption: [Experiment Instruction - Both players have AI.]