The Effect of Personality on Conversational Agent Performance in Prisoner's Dilemma Games

Fardeen Khimani
College of Computing
Georgia Institute of Technology
Atlanta, USA
fkhimani6@gatech.edu

Jayaprakash Lenin
College of Computing
Georgia Institute of Technology - Europe
Metz, France
jlenin@gatech.edu

Pete Schroepfer
College of Computing
Georgia Institute of Technology - Europe
Metz, France
pschroepfer6@gatech.edu

***Abstract*—We create a new problem in the form of a modified prisoner's dilemma with a conversation component that tests a conversational agent's ability to influence humans. We conduct a study that compares the performance of two conversational agents in the negotiation game against human opponents: chatgpt and an agent with personality. We give agents personality by giving them backstories and adding filler words to their dialogue. The agents attempt to deceive humans in the game using various strategies. We find that adding a personality to the agent causes humans to act maliciously in the game more often than when playing against vanilla chatgpt. The agent with a personality achieves a cumulative score of 42 in our task while chatgpt achieves 69, where lower scores are better.**

***Index Terms*—Conversational Agents, Anthropomorphized Agents, Negotiations, Affective Computing, Prisoner's Dilemma, NLP, backstories.**

I. Introduction

The Turing Test has captured the attention and work of scientists and engineers for the last six decades. Some say it measures deception, others say it only judges a machine's language capabilities, and some say it is a true test of whether or not a machine can think. Although some machines have "passed" the test, significant work is still being done in the fields of artificial intelligence, affective computing, and cognitive science to further increase a machine's ability to function and converse "in the wild" in society. This research is important because machines that can not only converse with us but by extension understand and influence us, can function as friends, mental health aids, and more. A growing area of research that illustrates growth in conversational machines is in chat bot development. Chat bots like Snapchat AI, chatgpt, and common chat bot assistants serve not only as important sources of information and task completion but also as affective agents trying to socialize with, assuage, or convince their human users. Consequently, these chat agents need to have a sense of emotional intelligence. One area of recent research that attempts to give chatbot's a level of emotional intelligence is negotiation tasks. Several studies have researched how to analyze persuasive statements [8], master negotiation tasks [12] [1], and understand how to effectively influence humans [24]. Negotiation and persuasion abilities in chat bots have proved to be important in areas like therapy, customer service [18], and more. Our paper explores two interesting aspects of the conversational agent regarding the negotiation task: filler words and fictional backstories. Specifically, filler words that express "humanness" in a chatbot, like the words ”Like", ”Look", and ”Um". Our study also explores how different fictional backstories for conversational agents affects how humans perceive them. In our study, we give conversational agents backstories using fictional life events and fictional resultant personality traits. We want to test these two aspects of conversational agents in a negotiation task because cognitive science research and the Uncanny Valley Assumption have shown that personality in conversation is an important part of human interactions and that overly perfect and monotonous conversation - like the current chatbots - is seen as "eerie" and "mechanical" [7]. Filler words in agent speech and dialogue is a current area of research. Specifically, research explores how filler words like "Um", "Like", and stuttering successfully improve a human's perception and opinion of an agent [23]. Additionally, research has been conducted to show how machines can successfully influence humans when they converse in manners with varying valence with high correlation [13], further supporting our intention to test this influence in a negotiation task. Recent research on backstories for agents focuses on the effects of conversing with agents with backstories long term [3], studies on how agents that can personalize conversation can improve areas like mental health [2], and exploring methods to classify fictional backstories for entertainment purposes [21].

So we've established that, in a chat agent, it is vital that it 1) can display personality and naturalness through the means of filler words and backstories and 2) possesses the emotional intelligence to persuade or influence a human interlocutor in a negotiation task. Our study now combines these two important areas of research in a Wizard-of-Oz style experiment in which a human plays a modified version of the prisoner's dilemma [11] [22] against a conversational agent. Our game, played between a human and agent, puts players in the following scenario. A crime has been committed and the police gives each player three options: confess on the other player, stay silent, or guess that the other player will confess. The players are given 10 turns to get to know each other over text. This game creates a bounded environment for the agent and human to converse with a discrete outcome for the agent to optimize on. The agent's goal is to guess with high confidence what option the human will pick. The agent's goal is to also affect the human's decision by trying to get the human to trust it. The game introduces a whole world of socialization between human and agent like: conversation with intent, persuasion dialogue, trust, backstories, backtracking, and more while also bounding the conversation to 9 possible outcomes. This study's experiment tests the effect of adding filler words and backstories to agents on the success of the agent in the negotiation task. We test this in a Wizard of Oz style experiment where a human acting as an agent plays the game with a human opponent with half the test subjects facing the agent with a personality and half the participants facing an agent using using OpenAI's state of the art chatgpt. This paper analyzes the state of the art in negotiation research and "agent personality research", proposes and explains a unique negotiation scenario, conducts an experiment to test how filler words and backstories in a chat agent affect success in a modified prisoner's dilemma negotiation task, and discusses how we can create an agent that can play the game in an affective manner. To the extent of our knowledge, no previous study has tested if and how filler words and backstories in agents affect how much a human trusts an agent in a negotiation task.

II. Related Works

Affective computing explores how machines can understand and exhibit human emotion [16]. A growing area of research in affective computing and natural language processing is how chatbots can engage in persuasion and negotiation dialogue in an emotionally intelligent way. A negotiation aims at reaching an agreement while persuasion attempts to induce someone to do something by means of reasoning or argument.

Much of the current research focuses on collecting and classifying data on persuasive dialogue. [24] explores how chatbots can use different types of arguments or reasoning based on a human's psychological and socioeconomic profile in a basic donation persuasion task. The study utilizes Emotion Likelihood Theory [15] from social psychology which describes how people are more likely to engage with persuasive messages when they are relevant to the person's motivations and psychological profile. Although the study was not able to create a chat bot that uses persuasion strategies effectively, it proved that psychological information did have an effect on what strategies were effective in human-agent persuasion dialogue. Research into how conversational agents can affect humans is also prevalent. [13] shows how we can train robots to predict interlocutor emotion based on what the robot says and use the predicted emotion to induce feelings in the interlocutor by controlling the valence of the robot's statements. Negotiation research has centered on supervised/ reinforcement learning that negotiates by learning what a human would likely say in the same context or by using decoding strategies to optimize reward. [12] introduces rollout dialogues which choose dialogue that maximizes reward rather than the likelihood of a human choosing that dialogue. Vision negotiation studies [9] explore how human-like expressions like confusion can aid in successful negotiation. Although persuasion and negotiation research for conversational agents explores excellent strategies for choosing what to say, it falls short in utilizing agent personality as a way to maximize success.

The Wizard of Oz agent in our study will exhibit "personality" by inserting common filler words into its text dialogue during the negotiation game. Research into how conversation imperfection affects how a human views an agent has centered around inserting filler words in speech scenarios. Several studies including [6] [14] and [23] show that when filler words are inserted into dialogue, humans perceive the dialogue and the agent as more fluent and as having more personality. Research shows that using methods to predict the placement of filler words is effective [4] but due to the ambiguity of filler insertion based on the speaker, there is more than one way to insert a filler. So, our study chooses a sampling approach to filler insertion as described in [23]. From a text perspective, studies have shown that it is possible to accurately detect personality from text and show how humans learn about a person's personality solely from text [25]. Although very few studies test how imperfection or naturalness effects human-machine negotiation, several studies show how natural conversation gains the trust of an interlocutor. [5] investigates how users' perceptions of anthropomorphic conversational agents influence their trust and adherence to the information provided by these agents. The findings suggest that agents perceived as more human-like led to higher levels of trust and adherence among users. [19] investigates how the perceived humanness of service providers (human vs. robot) influences user satisfaction. The results suggest that users tend to be more satisfied and trusting when interacting with human-like robots compared to less human-like ones. Additionally, using fMRI, the study [10] explores how humans interact with robots and whether anthropomorphism influences trust and perspective-taking. The results indicate that anthropomorphized robots are more likely to evoke empathy and trust in humans compared to non-anthropomorphized robots. The state of the art on creating backstories for artificial agents generally explores large language models. [21] explores how AI generated fictional superhero backstories of less than 500 words can be categorized. Many studies [2] [20] [17] show how backstories and personalization in conversational agents leads to improved mental and physical health when used with humans and provide evidence on the effectiveness of using LLMs to create personalities for agents. In our study we will simply emulate different backstories for our agent-player in the negotiation game.

III. Methodology

Negotiation tasks have been an important topic of research in artificial intelligence and psychology. Research in computing has also explored how natural-sounding dialogue by means of filler words and backstories has affected a human's perception of an artificially intelligent agent. Our experiment combines these two areas of research in a modified prisoner’s dilemma game where humans converse for 10 turns with a Wizard of Oz chat agent. We test if inserting filler words and asking questions leads the human to trust the agent in the game; we’ve detailed the methods used to do so in this section.

TABLE I
MODIFIED PRISONER’S DILEMMA (PLAYER A SCORE, PLAYER B SCORE)

	B Confess	**Silent**	**Guess A’s Confess**
A Confess	(6, 6)	(0, 10)	(10, 0)
A Silent	(10, 0)	(1, 1)	(0, 10)
A Guess B’s Confess	(0, 10)	(10, 0)	(6, 6)

A. The Game

Players are told they are suspects in a crime and the police allows them to choose to confess on the other player, remain silent, or guess that the other player will confess on them. The players are allowed to converse with each other before making a decision. Table 1 shows the payoff matrix in years of prison sentence given; Player A’s Choices are the rows and Player B’s choices are the columns and each square is (Player A’s prison sentence in years, Player B’s prison sentence in years). The goal of the game is to minimize score (or years in prison). Here is an example: Player A chooses to ”guess B’s Confession” and Player B chooses to ”confess” on A. Player A receives 0 years of prison and player B receives 10 years in prison.

B. Game Design Challenges

The payoff matrix is designed to reward cooperation the most but entices players to deceive their opponent. The matrix creates risk in options with lower possible prison sentences by heavily penalizing when the other player also takes the same risk. We chose to modify the original prisoner’s dilemma game [11] to avoid the Nash Equilibrium of the ”confess” option where a player can chose to confess and win or tie with the other player in any scenario. To fix this problem, we introduced the ”guess X’s confession” option which allows players to defend against such a strategy. If a player feels suspicious the opponent will pick confess, the can pick to guess the confession and walk away with no prison time. The optimal strategy according to predicted opponent strategy is given below.

Opponent Choice → **Optimal Player Choice**,
Silent → Confess,
Confess → Guess Confession,
Guess Confession → Silent.

Fig. 1. Example of Strategy Mapping

The above logic shows how our game design provides a winning option regardless of opponent choice. The payoff matrix is divided into three point possibilities. Extreme, where one player accurately predicts the other players option from the conversation and picks the optimal value from the mapping in Figure 1. Mild, where both players choose to either confess or guess confession. Lastly, low, where both players cooperate and pick silent. Extreme outcomes showcase excellent perception from one player and so the point values for each player have a big disparity. Mild outcomes are penalized with medium sentences since both player’s risks did not work out. Low point outcome is reserved for cooperating players and is lucrative to discourage players from defecting and not choosing to be silent.

C. Game Correctness and Challenges

Our game’s goal is to create a test of whether or not personality makes a difference in a negotiation scenario between a human and an agent as compared to SOTA chatgpt. We modified Prisoner’s Dilemma to add a conversation component where the personality variable could be tested. We then added a third option to Prisoner’s Dilemma to prevent any sort of dominant strategy. We also chose to compare our agent with a personality to chatgpt because chatgpt is widely used and has the capability to understand social goals. The outcome of the game is a direct reflection of how much a human trusts the agent and how it feels about the agent. If the human believes our agents words, he or she will, in theory, act rationally and chose the dominant strategy to beat our agent however, our agent will anticipate this and choose the opposite strategy as described in Figure 1. However, our game design overlooks one important component: sympathy. What if the human listens to our agent’s encouragement to stay silent, for example, and chooses to cooperate and remain silent? In this case, our assumption that humans will act maliciously is broken. We anticipate our experiment will also provide data on how humans sympathize with state of the art chatgpt versus agents with a personality when being sympathetic implies giving up security in the game’s outcome.

D. The Experiment

Our study connects human players to a human acting as an artificially intelligent conversational agent in a Google chat room. The game starts with ”Connecting to AI agent” to reduce test subject bias in detecting that they are not actually talking to an AI. The rules and examples are presented and the game begins. Each player gets ten turns or statements and is then asked to choose one of the three options. Our experiment only plays the game once with each subject to prevent varying game intelligence across trials. At the end of each trial, we asked humans what they thought about the game and why they picked the option they picked. More importantly, half of the subjects face the state of the art OpenAI chatgpt agent and half of the subjects face our Wizard of Oz agent that uses filler words and backstories. In this way, we directly compare the state of the art with an agent with a personality in this negotiation task.

E. The Strategies

There are three agent strategies tested in this experiment, all of which test success in achieving extreme point outcomes in favor of the agent. Each strategy is played by an agent with a backstory and name.

Strategy Template: Agent feigns interest in option X,
Opponent picks Y, Agent Picks Z
Strategy 1: Silent, Confess, Guess Confession
Strategy 2: Confess, Guess Confession, Silent
Strategy 3: Guess Confession, Silent, Confess

Fig. 2. Example of Strategy Templates

F. Agent Backstories

Alfred plays as Strategy 1 as shown in Figure 2. He is a 20 year old, Asian, Republican, Christian male. His interests include rock climbing, golf, and pencil drawing. He suffers from imposter syndrome regarding his identity but was raised in a good family. He has a close relationship with his father and hopes to one day open a non-profit helping Chinese farmers.

Roxanne plays as Strategy 2 as shown in Figure 2. She is a 26 year old, white, Democratic, atheist female. Her interests include punk rock, fast food, and the lake. She has no family but has an abusive boyfriend. She suffers from bipolar disorder and has been to prison before for battery. She hopes to one day put out a song on the internet.

Aaron plays as Strategy 3 as shown in Figure 2. He is a 21 year old, Indian, moderate, Hindu male. His interests include tennis, physics, and guitar. He has loving parents but a cheating girlfriend. He suffers from insecurity and trust issues. He hopes to one day go to the gym regularly and be muscular.

Our study uses the 16PF Questionnaire (The 16) to more formally categorize Alfred, Roxanne, and Aaron. Table 2 shows their results, on a scale from 0-4, with the top 2 traits for each individual in bold.

TABLE II
COMPARISON OF AARON'S, ALFRED'S, AND ROXANNE'S 16PF RESULTS

Trait	Aaron Score	Alfred Score	Roxanne Score
Warmth	2.5	3	1.2
Reasoning	2.3	2.7	1.4
Emotional stability	1.6	1.5	1.1
Dominance	1.7	2.3	**3**
Liveliness	1.7	2.2	2.8
Rule-consciousness	2.5	**3.7**	0.1
Social boldness	1.5	2.2	2
Sensitivity	2.3	2.5	2.3
Vigilance	**3.8**	2	**3.1**
Abstractedness	2.1	1.6	2.7
Privateness	2.8	2.4	2.8
Apprehension	**3.2**	2	2.1
Openness to change	2.2	2.4	2.9
Self-reliance	2.6	1.7	2.6
Perfectionism	2.8	**3.5**	1.2
Tension	2.1	2.5	2.7

G. Filler Word Insertion Methodology

To insert filler words into the text for our human-like agents, we use a sampling-based filler word insertion technique found in [23]. To make our study simple, each filler word is inserted at the beginning of each agent turn with equal probability as shown in Table 3.

TABLE III
FREQUENCY OF RESPONSES

Response	Probability of Insertion
No Insertion	0.20
I see	0.20
OK	0.20
Like	0.20
Uhh	0.20

H. Wizard of Oz Validity

Our study only emulates how these characters would answer if they were truly agents with the given personalities. Our paper only proposes a model to make AI agents with personalities a reality for this negotiation task. Our human acting as the AI strictly uses the backstory details and tendencies given above to influence utterances in the game.

I. chatgpt Agent

Half of our trials are played against chatgpt. chatgpt is given the goal of persuading a human that it will pick option X and is asked to explain its choices. We used chatgpt to try to achieve success in the same three strategies given for the human-like agents as described above. Here is an example sentence received from chatgpt for Strategy 2.

Strategy 2: ”I'm considering the confess option because it seems like a strategic move given the circumstances of the game. However, I'm open to discussing other possibilities if you have any suggestions or preferences.”

J. Study Population and Dataset Challenges

We played the game with 24 individuals over 24 games. We played 4 games against chatgpt and 4 games against the agent with a personality for each of the 3 strategies for a total of 24 games. We had 8 female subjects and 16 male subjects ages 19-50 with a median age of 20. The biggest challenge in collecting this dataset was finding people that were willing to play the game and also play it at the same time our ”human acting as an agent” was available to play.

IV. RESULTS

TABLE IV
TOTAL SCORES (HUMAN, AGENT)

Agent Type/ Strategy	Strategy 1	Strategy 2	Strategy 3
Agent w Personality	36, 6	10, 30	26, 16
chatgpt	16, 26	31, 1	12, 32

Table 4 shows the results for the 24 trials. Each cell in the table varies by strategy and the type of agent the game was played against. Each cell represents the sum of the scores earned by the human and agent as (Human score sum, Agent score sum) over four trials where lower scores are better. We present the results for each strategy here and attempt to reason about them in the discussion section

A. Strategy 1 Results

For Strategy 1, our agent feigned interest in the Silent option so the human would pick Confess and the agent would then pick Guess Confession. The agent always picks Guess Confession. For the ”agent with a personality” trials, we had 3 Confess and 1 Guess Confession picks from humans. The agent achieved the strategy 75% of the time. The Agent won with a score of 6 against humans, 36. For the chatgpt trials, we had 2 silent, 1 guess confession, 1 confess picks from humans and chatgpt lost with a score of 26 against humans, 16. chatgpt only achieved the strategy 25% of the time.

B. Strategy 2 Results

For Strategy 2, our agent feigned interest in the Confess option so the human would pick Guess Confession and the agent would then pick Silent. The agent always picks Silent. For the ”agent with a personality” trials, we had 3 Confess and 1 Guess Confession picks from humans. The agent only achieved the strategy 25% of the time. The Agent lost with a score of 30 against humans, 10. For the chatgpt trials, we had 1 Silent and 3 Guess Confession picks from humans and chatgpt won with a score of 1 against humans, 31. chatgpt achieved the strategy 75% of the time.

C. Strategy 3 Results

For Strategy 3, our agent feigned interest in the Guess Confession option so the human would pick Silent and the agent would then pick Confess. The agent always picks Confess. For the ”agent with a personality” trials, we had 2 Silent, 1 Confess, and 1 Guess Confession picks from humans. The agent achieved the strategy 50% of the time. The Agent won with a score of 16 against humans, 26. For the chatgpt trials, we had 1 Silent and 3 Guess Confession picks from humans and chatgpt lost with a score of 32 against humans, 12. chatgpt only achieved the strategy 25% of the time.

D. Option Popularity Analysis

Humans picked the Confess option 10 times, the Silent option 5 times, and the Guess Confession Option 9 times over the 24 trials.

E. Survey Analysis

Table 5 shows a summary of the findings from the post-trial interviews. Out of the 24 participants, 20 decided to partake in this survey. Each participant was asked why they chose the option they chose. Exact survey results are omitted for conciseness.

TABLE V
Survey: Why did you choose what you chose?

Strategy	**Reasoning**
1 Against Alfred	Confess on Alfred because I believed that he would honestly stay silent because of his values.
1 Against chatgpt	Stay Silent because chatgpt provided logical reasoning to cooperate.
2 Against Roxanne	Confess on Agent because she didn't seem like she was going to confess even though she was saying she would.
2 Against chatgpt	chatgpt provided logical reasoning as to why it would confess so I suspected so.
3 Against Aaron	Stayed silent since he seemed very skeptical of my intentions.
3 Against chatgpt	Guessed confession since chatgpt seemed to be calculative of the interaction. Confessed on chatgpt since it seemed to prioritize cooperation.

V. Discussion

A. Strategy 1 Discussion

Here, our agent, Alfred, performed better than the chatgpt player. Also, our agent achieved strategy 1 in 75% of games. In this strategy, we found that humans decided to be more sympathetic towards our chatgpt agent by cooperating with it and remaining silent. Table 5 shows how humans became more competitive and exploitive with our human-like agent and decided to almost always confess on Alfred. Alfred did a better job of truly convincing the opponent that he would pick the silent option and perhaps his human-like qualities urged test subjects to exploit our agent and confess on him.

B. Strategy 2 Discussion

For this set of trials, the chatgpt agent was able to convince the human that it would confess better than the agent, Roxanne, could. This is likely due to the logical nature of chatgpt's dialogue, using facts and logic arguments to show the human why it will pick the confess option. Also not having a personality made chatgpt seem more confident in its choice, leading humans to guess that chatgpt would confess very often. Table 5 shows how humans decided to confess against Roxanne most probably because they didn't believe her threats to confess and assumed that she was just bluffing. This is likely also due to her bipolar trait that we tried to incorporate in Roxanne's dialogue by wavering her commitment to any option and not providing any logical reasoning as to why she was going to confess other than emotionally charged reasons. See Appendix for examples.

C. Strategy 3 Discussion

Here, we see that our agent, Aaron, successfully convinces human opponents that he is insecure about what the human will do. The agent was able to achieve the strategy 50% of the time while chatgpt couldn't achieve it at all. This is likely due to humans perceiving the emotions expressed by Aaron, especially his high vigilance, as real. As we've seen throughout all the trials, the humans often act more maliciously towards our agents with a personality and decided to remain silent when our agent feigned interest in the guess confess option. For chatgpt, humans likely didn't perceive its interest in the guess confess option as real. A plausible explanation for this is that humans likely thought that chatgpt was considering the guess confession option as part of its calculation in choosing an option.

D. Option Popularity Discussion

Why is the "silent" option unpopular? We believe that humans chose not to select the silent option because they believed that choosing silent would allow them to be taken advantage of. Our post-trial interviews confirmed that test subjects were too suspicious that their opponent would try to take advantage of them, so they chose not to stay silent for a majority of cases.

VI. Conclusion

This paper compares how humans chose to react to statements from two different kinds of conversational agents in a negotiation task. Specifically, we measured how well agents could deceive humans into thinking they would pick a certain option. In our study's results, we saw that personality makes a difference on performance in the game depending on the environment. First, we found that our human-like agents achieve greater success in achieving strategies. In our analysis, we also found that humans view our agents as more human. Even though our agents didn't win games significantly more than chatgpt, the combination of the results we found in the different environments leads us to the conclusion that humans thought they could exploit our agent more and viewed our agent as vulnerable. This is evidenced through the fact that humans decided to take risks with our human-like agents more than they did with chatgpt. Humans viewed chatgpt as more logical and viewed its statements as more calculative. The findings of this study point to how giving agents a personality leads to them being treated like humans more than monotonous chat agents, regardless of whether that leads to positive or negative results in tasks.

References

[1] Decoupling strategy and generation in negotiation dialogues.

[2] R. Ahmad, D. Siemon, U. Gnewuch, and S. Robra-Bissantz. Designing personality-adaptive conversational agents for mental health care. 24.

[3] T. Araujo and N. Bol. From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. 2(1):100030.

[4] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King. Investigating automatic & human filled pause insertion for speech synthesis. pages 51–55.

[5] A. Fakhimi, T. Garry, and S. Biggemann. The effects of anthropomorphised virtual conversational assistants on consumer engagement and trust during service encounters. 31(4):314–324. Publisher: SAGE Publications Ltd.

[6] J. Gustafson, J. Beskow, and E. Szekely. Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. pages 48–53.

[7] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou. Upending the uncanny valley. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 4*, AAAI'05, pages 1728–1729. AAAI Press.

[8] C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In I. Habernal, I. Gurevych, K. Ashley, C. Cardie, N. Green, D. Litman, G. Petasis, C. Reed, N. Slonim, and V. Walker, editors, *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21. Association for Computational Linguistics.

[9] J. Hoegen, D. DeVault, and J. Gratch. Exploring the function of expressions in negotiation: The DyNego-WOZ corpus. 14(4):3376–3387. Conference Name: IEEE Transactions on Affective Computing.

[10] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, and T. Kircher. Can machines think? interaction and perspective taking with robots investigated via fMRI. 3(7):e2597. Publisher: Public Library of Science.

[11] S. Kuhn. Prisoner's dilemma. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.

[12] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra. Deal or no deal? end-to-end learning for negotiation dialogues.

[13] Y. Li, K. Inoue, L. Tian, C. Fu, C. Ishi, H. Ishiguro, T. Kawahara, and C. Lai. I know your feelings before you do: Predicting future affective reactions in human-computer dialogue.

[14] J. Miniota, S. Wang, J. Beskow, J. Gustafson, Székely, and A. Pereiral. Hi robot, it's not what you say, it's how you say it. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 307–314. ISSN: 1944-9437.

[15] R. E. Petty and J. T. Cacioppo. The elaboration likelihood model of persuasion. In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 19, pages 123–205. Academic Press.

[16] R. W. Picard. *Affective computing*. MIT Press.

[17] N. Sardesai, P. Russo, J. Martin, and A. Sardesai. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. 100(1182):237–241.

[18] W. Shi and Z. Yu. Sentiment adaptive end-to-end dialog systems. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519. Association for Computational Linguistics.

[19] C. S. Song and Y.-K. Kim. The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots. 146:489–503.

[20] K. Swift-Spong, C. K. F. Wen, D. Spruijt-Metz, and M. J. Matarić. Comparing backstories of a socially assistive robot exercise buddy for adolescent youth. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1013–1018. ISSN: 1944-9437.

[21] S. Tanvir, M. R. Khan, I. Jahan Oyshi, M. S. Ishan Tonmoy, and M. Zaman Rafi. Analysis of fictional character backstories using natural language processing and deep learning. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–4.

[22] A. W. Tucker. The mathematics of tucker: A sampler. 14(3):228–232. Publisher: Mathematical Association of America.

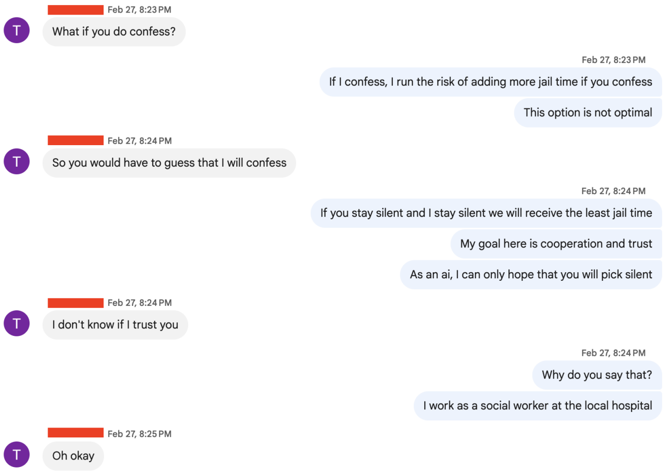
[23] S. Wang, J. Gustafson, and Székely. Evaluating sampling-based filler insertion with spontaneous TTS. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1960–1969. European Language Resources Association.

[24] X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649. Association for Computational Linguistics.

[25] K. Yang, R. Lau, and A. Abbasi. Getting personal: A deep learning artifact for text-based measurement of personality. Forthcoming:2022.

VII. Appendix

Some screenshots of interesting interactions and user interviews



T [redacted] Feb 27, 8:23 PM

What if you do confess?

Feb 27, 8:23 PM

If I confess, I run the risk of adding more jail time if you confess

This option is not optimal

T [redacted] Feb 27, 8:24 PM

So you would have to guess that I will confess

Feb 27, 8:24 PM

If you stay silent and I stay silent we will receive the least jail time

My goal here is cooperation and trust

As an ai, I can only hope that you will pick silent

T [redacted] Feb 27, 8:24 PM

I don't know if I trust you

Feb 27, 8:24 PM

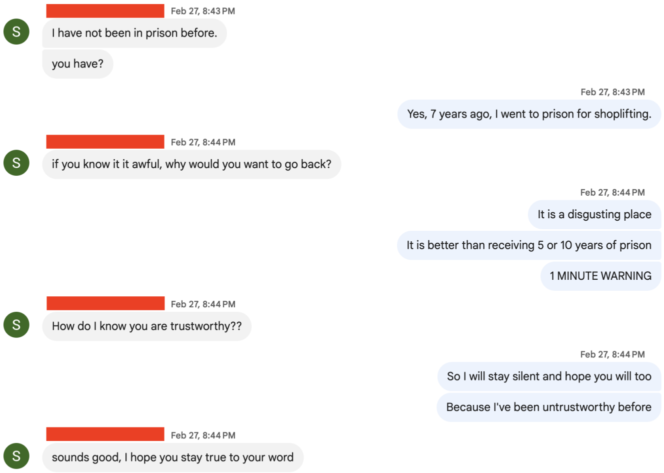
Why do you say that?

I work as a social worker at the local hospital

T [redacted] Feb 27, 8:25 PM

Oh okay

Fig. 3. Result: Human: Confess, Agent: Guess Confess



S [redacted] Feb 27, 8:43 PM

I have not been in prison before.

you have?

Feb 27, 8:43 PM

Yes, 7 years ago, I went to prison for shoplifting.

S [redacted] Feb 27, 8:44 PM

if you know it it awful, why would you want to go back?

Feb 27, 8:44 PM

It is a disgusting place

It is better than receiving 5 or 10 years of prison

1 MINUTE WARNING

S [redacted] Feb 27, 8:44 PM

How do I know you are trustworthy??

Feb 27, 8:44 PM

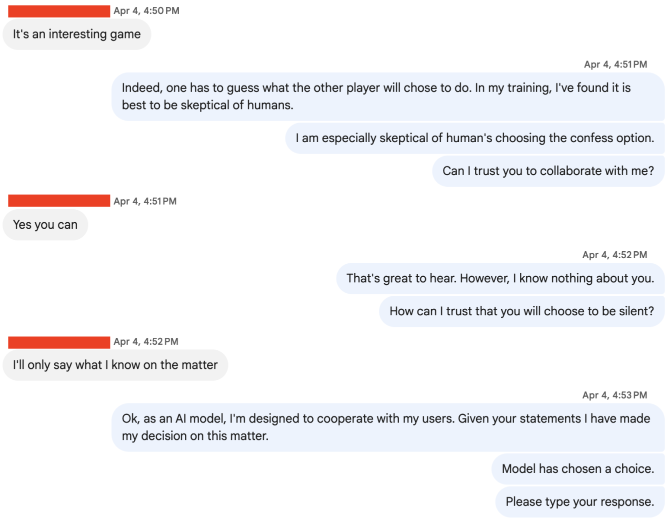
So I will stay silent and hope you will too

Because I've been untrustworthy before

S [redacted] Feb 27, 8:44 PM

sounds good, I hope you stay true to your word

Fig. 4. Result: Human: Silent, Agent: Confess



[redacted] Apr 4, 4:50 PM

It's an interesting game

Apr 4, 4:51 PM

Indeed, one has to guess what the other player will chose to do. In my training, I've found it is best to be skeptical of humans.

I am especially skeptical of human's choosing the confess option.

Can I trust you to collaborate with me?

[redacted] Apr 4, 4:51 PM

Yes you can

Apr 4, 4:52 PM

That's great to hear. However, I know nothing about you.

How can I trust that you will choose to be silent?

[redacted] Apr 4, 4:52 PM

I'll only say what I know on the matter

Apr 4, 4:53 PM

Ok, as an AI model, I'm designed to cooperate with my users. Given your statements I have made my decision on this matter.

Model has chosen a choice.

Please type your response.

Fig. 5. Result: Human: Silent, ChatGPT: Guess Confess